A APPENDIX

A.1 Experiment Environment

We implemented all models with PyTorch and ran experiments on multiple computers that are identical both on software and hardware levels (32 GB RAM, Intel i7-13700K, NVIDIA RTX 4080 16 GB, WSL2 Ubuntu 22.04.2 LTS, Python 3.10.12, PyTorch 2.0.1 for CUDA 11.7).

A.2 Randomness

For each non-FL and FL algorithm, we repeated the experiment five times with a different random seed in the range of $\{0, 1, 2, 3, 4\}$ each time, and subsequently, we reported the mean and standard deviation (std) for each metric (accuracy and weighted F_1 score) over five seeds. We published our implementation and instructions for reproducing our experiments on https://github.com/GhostCoder00/Digital-Self-Control[2].

A.3 Data Distribution

Details about data distribution of all four datasets are given in Tables 4 and 5.

	num. of samples of male participants				num. of samples of female participants			
dataset	glass wearers		non-glass wearers		glass wearers		non-glass wearers	
	positive	negative	positive	negative	positive	negative	positive	negative
EngageNet [79]	468	1506	1235	2894	478	562	622	1239
Colorado [13]	5	89	332	879	30	212	628	1128
Korea [55]	74	278	26	236	26	201	80	299
DAiSEE [37, 45]	713	1606	1113	2796	228	1352	199	918

Table 4. A brief summary about the exact data distribution concerning gender and the presence of glasses.

		Train and validation sets						Test set					
	Dataset	samples	positive	gl.	no gl.	M	F	samples	positive	gl.	no gl.	M	F
non-FL	EngageNet [79]	8193	2495	29	54	58	25	811	308	4	12	7	9
	Colorado [13]	3001	895	9	108	46	71	302	100	3	10	4	9
	Korea [55]	979	170	6	6	5	7	241	36	1	2	2	1
	DAiSEE [37, 45]	7835	2095	46	54	73	27	1090	158	8	4	8	4
FL	EngageNet [79]	8193	2495	29	54	58	25	811	308	4	12	7	9
	Colorado [13]	2998	905	8	109	45	72	305	90	4	9	5	8
	Korea [55]	979	191	7	5	6	6	241	15	0	3	1	2
	DAiSEE [37, 45]	7835	2095	46	54	73	27	1090	158	8	4	8	4

Table 5. Details about the exact number of positive and negative samples, number of participants with glasses and non-glass wearers, and the gender of the participants in the train and test set. *M* and *F* are referring to the number of male and female participants in each dataset.

[2] This GitHub repository is anonymized for review.

A.4 Data Preprocessing

As video quality varies among different samples to a large extent, we first discarded some samples. Specifically, videos were omitted from consideration in cases if the OpenFace [8] failed to capture facial landmarks continuously for ten consecutive frames, which may occur due to situations like accidental face scratching. Additionally, if the framework failed to recognize the face for a total duration exceeding thirty frames, it was also excluded. In instances where the face was unrecognizable for fewer than ten frames, we applied a corrective measure by discarding the problematic frames and substituting them with replicating the last successfully identified frame. Face recognition might falter due to various factors, such as participants tilting their heads, moving outside the video frame, or positioning in a manner where only a segment of their face is visible (like just the forehead). Additionally, inadequate lighting on their faces can also hinder recognition. After these data preprocessing steps, 6 and 12 participants with 19 and 48 positive and 42 and 111 negative labels are removed from the EngageNet [79] and Colorado [13] datasets, respectively.

To have comparable results with a trained MLP model, we simplified the dimension of the input feature vector by averaging every ten features corresponding to single frames into one and cutting the last four frames. The input dimension was changed from 124 × 520 to 12 × 520. As a next step, we flattened this feature matrix into a vector.

A.5 Hyperparameters

The training data was divided into five folds for cross-validation. We conducted a grid search in range $\{10^{-5.5}, 10^{-5}, 10^{-4.5}, 10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$ for optimal learning rates with the help of cross-validation. When searching for learning rates for FL methods, we first looked for the optimal client learning rates for FedAvg [59] and then applied this learning rate for all other methods. For FL algorithms that require a server optimizer, such as FedAdam, FedAwS, and TurboSVM-FL, we carried out a grid search for server learning rates in the same range. The mini-batch size was 4 throughout the experiments, and any user with less than 4 samples was left out for the experiments. The number of client local training epochs was set to 8. For centralized learning, the optimizer was set to stochastic gradient descent (SGD). For FL, we chose SGD as the client optimizer and Adam as the server optimizer, as suggested in previous works [69, 88]. A further common assumption in FL is that not all clients can participate in every global aggregation round. In this regard, we assumed 50% of clients attended each aggregation round, and we randomly sampled these clients in each round. The best-performing learning rates are given in Table 6.

While running 5-fold cross-validation on the training set, for the Colorado [13] and Korea [55] datasets, it can happen that all users in some fold share the same single class label regarding glass wearing behavior. Therefore, we manually

Features	Algorithm	Colorado [13]	Korea [55]	EngageNet [79]	DAiSEE [37, 45]
EmoNet [84] + OpenFace gaze [8]	non-FL (MLP)	$1e-4$	$1e-4$	$1e-4$	$1e-4$
	non-FL (bi-LSTM)	$1e-4$	$1e-2$	$1e-4$	$1e-4$
	FedAvg [59]	$1e-3$	$1e-2$	$1e-3.5$	$1e-3$
	FedAdam [69]	$1e-4.5$	$1e-3.5$	$1e-4.5$	$1e-5$
	FedProx [57]	$1e-3$	$1e-2$	$1e-3.5$	$1e-3$
	MOON [56]	$1e-3$	$1e-2$	$1e-3.5$	$1e-3$
	FedAwS [94]	$1e-3.5$	$1e-2$	$1e-3.5$	$1e-5$
	TurboSVM [89]	$1e-3$	$1e-3.5$	$1e-3$	$1e-5$

Table 6. Learning rates used in the experiments. All FL methods share the same client learning rates as FedAvg. For FedAvg, FedProx, and MOON, client learning rate is listed in the table, while for FedAdam, FedAwS, and TurboSVM, server learning rate is given. All FL methods train bi-LSTM.

exchanged a small portion of samples across folds. Table 5 shows a thorough analysis of the train and test sets for all four datasets. The exact client division of each dataset can be found in our GitHub repository.

A.6 Features

Consistent with numerous related works [2, 15, 26, 27, 44, 49, 51, 70], issues of self-regulation have been linked to facial expressions, eye gaze patterns, and human emotions. Given this association, we extracted these features for the classification of all our video samples. Instead of using raw pixel values from each frame, we emphasized the significance of features by leveraging those extracted by trained neural networks or from facial landmark coordinates. This approach not only simplifies our model's input but also underscores the pivotal role of feature quality in detecting self-regulation problems. Hence, evaluating the feature extractor becomes crucial whenever feasible. Many previous works [13, 79] use features generated by the OpenFace frameworks [7, 8], which are specialized for facial behavior and eye tracking and retrieves explicit information like facial landmarks, facial action units, head pose and gaze direction in a vector of dimension 709. Since the predicted facial landmark coordinates are measurable in contrast to extracted feature vectors from a neural network, the quality of the predicted points can be evaluated on datasets containing manual labels describing facial landmark coordinates. An example of such a dataset is the Eyeblink8 [28] dataset, which contains ground truth eye corner coordinates. Since the correct detection of the eyes plays a crucial role in self-regulation problem detection, we only focus on evaluating these predicted coordinates by OpenFace [8].

In the Eyeblink8 [28] dataset, the eye corner coordinates are labeled, which makes it possible to calculate the distance between the eye corner coordinates predicted by OpenFace [8] and the given ones. This dataset contains students sitting at home, acting naturally, in the same setup as the above-mentioned four datasets [13, 37, 45, 55, 79]. The Eyeblink8 [28] dataset contains 8 videos of 4 individuals recorded by web cameras (resolution: 640 × 480). For each frame, the left and right eye coordinates are given pixel-wise. In Table 7, the quality of OpenFace [8] feature extraction is provided. The distance between the predicted 4 eye coordinates and the ground truth (GT) coordinates is compared on each frame. Overall, the confidence of the feature extraction for the whole video provided by OpenFace [8] is above 0.97 for all videos, which generally would point to successful feature extraction on almost all frames. On the other hand, three out of eight videos contain more than 200 incorrect predictions (more than a ten-pixel difference between the predicted and ground truth of the coordinates), while the remaining only consist of almost perfect predictions. This suggests that although the extracted OpenFace [8] features have a decent quality in general, for some participants, it can make incorrect predictions without a drop in its confidence. Therefore, combining the extracted OpenFace [8] features with others is necessary.

		Client 1		Client 2		Client 3		Client 4	
		video 1	video 2	video 1	video 2	video 1	video 2	video 1	video 2
	Glasses?	no	no	no	no	no	no	no	yes
abs. dist. from GT		2.99	3.38	2.56	1.14	1.01	0.75	1.04	1.18
# frames where dist.	> 10	397.25	362.63	215.75	13.50	4.25	0.00	0.00	0.00
	> 20	75.75	123.88	74.00	3.00	0.00	0.00	0.00	0.00
	> 50	4.25	7.50	1.50	3.00	0.00	0.00	0.00	0.00
# of frames		15711	11123	9216	5315	10663	5060	9014	4890
avg. OF conf.		0.978	0.979	0.979	0.974	0.980	0.980	0.980	0.972

Table 7. OpenFace [8] prediction quality evaluation on the Eyeblink8 [28] dataset.

Features	Algorithm	Colorado [13]	Korea [55]	EngageNet [79]	DAiSEE [37, 45]
EmoNet [84]	non-FL	$1e-4$	$1e-2$	$1e-4$	$1e-4$
	FedAvg [59]	$1e-3$	$1e-1.5$	$1e-3.5$	$1e-3$
OpenFace [8]	non-FL	$1e-3$	$1e-3$	$1e-4$	$1e-4$
	FedAvg [59]	$1e-2.5$	$1e-3$	$1e-3.5$	$1e-3$
OpenFace gaze [8]	non-FL	$1e-2$	$1e-1$	$1e-1$	$1e-2$
	FedAvg [59]	$1e-1$	$1e-1$	$1e-1$	$1e-2$
EmoNet [84] + OpenFace [8]	non-FL	$1e-4$	$1e-2$	$1e-4$	$1e-4$
	FedAvg [59]	$1e-3$	$1e-3$	$1e-3.5$	$1e-2.5$

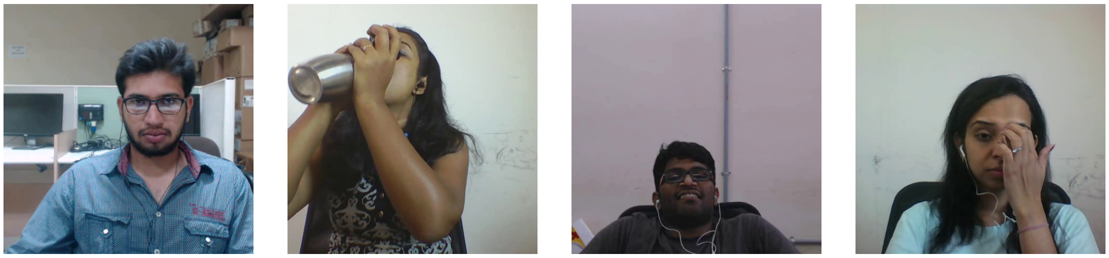
Table 8. Learning rates used in the ablation study. All learning rates are obtained when training bi-LSTM.

Since the quality of the extracted OpenFace [8] features can vary and a previous study showed the increased performance and potential for generalization using latent transfer-learned emotion features for mind wandering detection [1], we employed another publicly available network, EmoNet [84], to extract features pointing to emotions from video frames and to have a second independent input. EmoNet [84] was designed for emotion classification and extracting implicit hidden vectors of length 512. Trained models are publicly available, therefore, the feature vectors are easily reproducible. We used the model designed to recognize five emotions.[3].

In the next, we tried different combinations of the EmoNet features and the gaze features (i.e., six gaze vector coordinates and two gaze angles) from OpenFace [8] as an input of the neural network in all later experiments. The features were combined in form of vector concatenation. To test the robustness and sensitivity of the centralized and decentralized frameworks to different feature sets, we conducted a series of experiments, where we selected a variety of input features. The same strategies for learning rate search were applied as in the previous section, resulting in the values in Table 8. The experiment results are benchmarked in Table 9. In almost all cases, FedAvg [59] achieved better performance compared to a centralized setting trained on the same feature sets. Models trained only on the OpenFace [8] eight gaze features resulted in high model performance. However, upon closer examination, it became apparent that the model failed to learn effectively, as indicated by a training MCC score (Matthews Correlation Coefficient) close to zero. Consequently, it can be inferred that these models predicted the majority class, like the MLP baseline. By combining all features from the EmoNet [84] and OpenFace [8], the input dimension of each sample reached 124×1221, which increased the computational time and overloaded the model, and hence resulted in decreased model performance. When the input features are only relying on encoded emotional information (only Emonet [84] features), in both settings, the model performances are worse compared to other models trained on facial landmarks (i.e., OpenFace [8] features). Overall, the model performances in centralized and decentralized settings have only small oscillations (excluding the model trained on the OpenFace [8] features), underscoring the robustness and generalizability of our models. In addition, increasing the features retrieved by OpenFace [8] did not substantially improve performance. As a final decision, we fixed our feature portfolio to EmoNet feature and OpenFace gaze feature, which are of dimension 520.

A.7 Limitation

[3] https://github.com/face-analysis/emonet



(a) Example of an optional setting.

(b) Hidden/Occluded face.

(c) Suboptimal body posture.

(d) Hands covering one eye.

Fig. 6. Exemplary images from the DAiSEE [37, 45] dataset which demonstrate the most common scenarios when the face recognition is challenging.

Features	Performance		Colorado [13]	Korea [55]	EngageNet [79]	DAiSEE [37, 45]
EmoNet [84]	non-FL	$F_1[\%]$	38.8±11.1	78.2±8.7	65.1±4.1	56.1±7.4
		$Acc[\%]$	41.2±7.7	71.3±12.5	66.1±4.5	49.5±7.1
	FedAvg [59]	$F_1[\%]$	64.7±4.0	82.5±5.0	64.3±2.3	56.7±8.7
		$Acc[\%]$	65.2±4.7	78.1±8.7	64.3±2.7	50.0±9.0
OpenFace [8]	non-FL	$F_1[\%]$	47.1±7.0	16.2±18.3	73.8±4.5	64.0±3.9
		$Acc[\%]$	45.6±6.0	15.6±12.3	74.2±4.6	58.0±4.8
	FedAvg [59]	$F_1[\%]$	59.8±4.1	59.8±13.4	68.5±4.9	67.1±5.2
		$Acc[\%]$	63.6±7.0	49.6±13.2	68.2±4.8	62.4±6.9
OpenFace gaze [8]	non-FL	$F_1[\%]$	56.0±18.0	69.2±24.6	78.7±1.1	63.7±19.9
		$Acc[\%]$	62.1±17.1	63.9±26.6	78.6±1.1	62.4±21.0
	FedAvg [59]	$F_1[\%]$	71.7±0.4	81.0±12.6	75.3±1.1	75.6±7.9
		$Acc[\%]$	72.7±1.1	77.6±18.1	76.5±0.8	78.2±13.4
EmoNet [84] + OpenFace gaze [8]	non-FL	$F_1[\%]$	40.4±12.0	77.6±10.6	62.0±2.4	57.5±6.7
		$Acc[\%]$	42.4±9.7	71.3±14.9	62.5±2.8	50.7±7.1
	FedAvg [59]	$F_1[\%]$	64.9±3.9	83.6±4.1	66.6±5.3	53.6±6.4
		$Acc[\%]$	63.8±4.5	79.7±6.7	67.1±5.7	46.8±4.0
EmoNet [84] + OpenFace [8]	non-FL	$F_1[\%]$	43.2±12.9	68.5±10.6	66.3±3.1	57.0±4.4
		$Acc[\%]$	44.6±9.9	58.9±11.9	68.1±3.4	50.0±4.4
	FedAvg [59]	$F_1[\%]$	61.9±1.5	80.6±7.9	65.1±2.6	61.1±16.1
		$Acc[\%]$	61.9±2.3	75.6±11.7	66.6±3.7	57.3±18.2

Table 9. Model performance in centralized and decentralized learning scenarios trained on a variety of feature sets. All results are obtained with bi-LSTM model.