Federated Learning for Privacy-aware Cognitive Workload Estimation

Dario Fenoglio
Università della Svizzera Italiana
Lugano, Switzerland
dario.fenoglio@usi.ch

Daniel Josifovski
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
daniel.yosifovski@outlook.com

Alessandro Gobbetti
Università della Svizzera Italiana
Lugano, Switzerland
alessandro.gobbetti@usi.ch

Mattias Formo
Università della Svizzera Italiana
Lugano, Switzerland
mattias.formo@usi.ch

Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
hristijang@feit.ukim.edu.mk

Martin Gjoreski
Università della Svizzera Italiana
Lugano, Switzerland
martin.gjoreski@usi.ch

Marc Langheinrich
Università della Svizzera Italiana
Lugano, Switzerland
marc.langheinrich@usi.ch

ABSTRACT

Human physiological monitoring has become easily accessible by integrating wearable devices into our lives, providing valuable real-time data. Methods for Cognitive Workload (CW) estimation utilize such physiological data to quantify CW during task execution. These methods are crucial for various domains, including mobile healthcare, forecasting human errors, and human-machine interaction. However, accurately estimating CW continues to pose a challenge due to the absence of objective ground truth data, context dependency, and the privacy sensitivity of the data. This study tackled the complex task of estimating CW based on privacy-sensitive data (e.g., eye movement, pupil diameter, blink information, and other physiological signals) using Federated Learning (FL) methods to improve user privacy. We compared the outcomes of the FL models with the more conventional centralized approach on two publicly-available datasets COLET and ADABase, which include data from 75 participants overall. The results highlight the efficacy of FL in collaboratively training a global (person-independent) model. The FL models achieved performances on par with centralized state-of-the-art models while preserving data privacy. Recognizing the importance of privacy in user sensing, FL presents a promising approach that enables wearable sensing applications in privacy-sensitive domains.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Security and privacy → Distributed systems security**; • **Applied computing** → *Health care information systems*; • **Computer systems organization** → *Client-server architectures*; • **Information systems → Decision support systems**.

KEYWORDS

Cognitive Workload Estimation, Federated Learning, Data Privacy, Machine Learning.

1 INTRODUCTION

The widespread use of sensors embedded in wearable devices has facilitated human physiological monitoring, enabling real-time biofeedback. Wearable devices are extensively employed to enhance the performance of athletes, students, soldiers, and pilots [21]. Cognitive Workload (CW) aims to quantify the cognitive effort exerted by individuals during task execution, directly referencing the cognitive resources allocated by that task [17]. Accurate CW estimation could enable valuable use cases of wearable devices, e.g., for preventing burnout, reducing medical errors during complex diagnosis processes, and avoiding sub-optimal care outcome [5]. Future human-machine interfaces [29] could also be augmented by accurate CW estimation. For example, automatic CW estimation could enhance driving safety by identifying high CW situations due to secondary tasks that may compromise a driver's focus [28].

However, accurate estimation of CW remains an ongoing challenge. First, the main limitation is the lack of objective ground truth data, as CW has a substantial subjective component, and it often relies on self-report measures or subjective ratings [31]. To achieve a more precise evaluation of CW, a multidimensional measure based on a set of self-validated questions, such as the NASA-Task Load Index (TLX) [13], can be employed.

Furthermore, CW is highly context-dependent, depending on task characteristics and environmental factors. Thus, methods for CW estimation must consider and account for these external influences. Common approaches for estimating CW involve physiological signals such as breathing patterns, brain activity, eye movement, skin temperature, electrodermal activity (EDA), Photoplethysmography signals (PPG), electrocardiography (ECG) signals, electromyography (EMG), Electroencephalography (EEG) signals, and similar [11, 18, 28].

Privacy protection, however, remains a significant concern that has received limited attention in prior research. The state-of-the-art sensors currently employed for CW estimation, including cameras, eye trackers, and biomedical sensors [18, 28]. The use of these sensor data poses the risk of unintentional exposure of users' physiological data, potentially leading to privacy attacks that can reveal users' identity, emotional states, or health conditions. This information could further be exploited for purposes unrelated to the intended application, such as targeted advertising or profiling. Thus, it is crucial to comply with data protection regulations (e.g., GDPR) that relate to critical ethical aspects such as data collection, storage, and sharing practices [20]. Privacy-awareness is necessary for practical deployment of CW estimation systems in various contexts, including education, healthcare, and human-computer interaction.

Addressing these privacy challenges, one might consider training individual models on local user datasets, thereby avoiding data transfer across users. However, this approach sacrifices the shared information across individuals, mitigating user heterogeneity and potentially compromising the model's generalizability and performance. A possible solution is to employ privacy-aware modeling techniques, such as Federated Learning (FL). This approach addresses privacy concerns by allowing only non-sensitive modeling parameters to be shared for training a global model. The strength of FL lies in its ability to emulate the comprehensive insights of a centralized global dataset as if all users' data were combined without compromising individual privacy. Despite the promise, FL is a recent technique with many open challenges, rendering its applicability for CW estimation questionable. Example challenges include federated model selection, evaluation, and hyperparameter optimization [15]. The domain of CW estimation magnifies these challenges, adding to the mix of challenges: noisy data, limited data, and domain shifts in both labels and sensor data (due to subjectivity).

Our study explores the application of FL to overcome data privacy concerns in CW estimation. This contribution was realized by:

- The creation of preprocessing and feature-extraction pipelines for CW estimation from privacy-sensitive data — such as eye movement, pupil diameter, and blink information — to enable a direct comparison with the state-of-the-art methods in the field.
- The development of both feature-based and end-to-end machine learning models for CW estimation.
- Experimental comparisons between the FL models with the more conventional centralized models on two publicly-available datasets COLET and ADABase, which include data from 75 participants overall.

To the best of our knowledge, this is the first study that explores FL for the development of privacy-aware CW estimation pipelines.

2 RELATED WORK

2.1 Cognitive Workload Estimation

CW evaluation is typically examined by inducing single or dual-task workloads to manipulate the intensity of mental effort. During these tasks, various physiological measurements, such as EMG, ECG, PPG, EDA, skin temperature, and respiration rate, are recorded. It's important to acknowledge that all of these data sources are vulnerable to external sources of noise, such as physical movements, which can distort the sensor information. Additionally, variations in individuals' baseline physiological activity can introduce uncertainty, making it challenging to differentiate CW-related changes.

Additionally, behavioral measurements derived from actions captured by cameras provide further insights for the CW estimation. However, changes in lighting conditions can impact the accuracy of eye-tracking measurements, requiring continuous adjustments to accommodate shifting environmental factors, which may affect the precision of CW assessments [4].

To counteract the inherent subjectivity of CW assessments, the actual mental state of individuals is often assessed using performance metrics on the tasks or through subjective questionnaires [28], aiming to provide a more objective and precise evaluation.

Previous studies have delved into the evaluation of mental states using EEG signals [2, 26], ECG signals [12, 16], and heart rate variability (HRV) [3], mainly employing machine learning techniques. Notably, eye-tracking metrics, such as pupil diameter and blink information, have been demonstrated to be closely associated with CW [7]. Elevated CW levels result in prolonged fixation latencies and saccade durations, increasing average peak saccadic velocity [18]. For a detailed overview of approaches for measuring CW, we refer the readers to the recent survey presented by Kosch et al. [17].

The related work has demonstrated that physiological signals offer valuable insights into CW. However, the sensors' susceptibility to noise, external factors, limited datasets, absence of ground truth data, intrinsic difficulty of the task, and high-subject-dependency can make precise CW estimation a challenging task. To address some of these challenges and achieve fine-grained CW estimation, advanced preprocessing techniques are required, including normalization methods and supervised or unsupervised domain adaptation [14, 30]. However, these challenges become even more pronounced in FL, where the entire dataset, and consequently, global dataset statistics, cannot be collected centrally. In our pursuit to advance the field, we are the first to explore CW estimation based on FL while aiming to maintain comparable performance with centralized approaches and safeguard data privacy. We propose a comprehensive pipeline for locally preprocessing the client data, ensuring compatibility with federated approaches.

2.2 Federated Learning

FL is a machine learning approach offering higher privacy levels than centralized machine learning. In a typical FL environment, each model is trained locally on user devices without transmitting raw data [25] in a common (centralized) location. Specifically, in horizontal FL, each user computes training gradients locally, and only an encrypted version of these gradients is sent to the central server for secure aggregation. Similarly, the aggregated results are sent back to the user, where they can be decrypted to allow for model updates. To prevent indirect leakage of personal information, the most used privacy techniques in the FL framework are

DOI: 10.1145/3626705.3627783
Published December 2023

Secure Multi-party Computation, Differential Privacy, and Homomorphic Encryption [34]. To mitigate some of the FL challenges (e.g., non-IID data), various adaptations of the traditional Federated Averaging [25] algorithm have been introduced, which involve the simple average aggregation of weights from different models. These adaptations include FedProx [22], q-FedAvg [23], and Personalized-FedAvg [8]. Horizontal FL finds application in diverse wearable computing domains, such as human behavior classification based on EEG signals [9] and human mobility modeling using location data [6].

FL is also gaining attention in the field of Affective Computing, which aims to detect an individual's mental state. Affective Computing spans a wide spectrum, encompassing stress-level recognition, CW estimation, and emotion recognition, among others. Liu et al. [1] highlighted the advantages of FL in stress classification tasks using physiological and motion data. Their approach showed improved results compared to models trained solely on local datasets, a scenario where data sharing is impossible due to privacy constraints. However, their study did not use a user-independent test set for model validation. Using a user-independent test set is crucial as it helps ensure the model's generalization capability across different users, providing a more robust assessment of its real-world applicability. Similarly, Yekta Said Can and Cem Ersoy [24] employed FL to analyze heart activity data captured by smart bands for stress monitoring. Another notable study is Fed-ReMECS [27], which introduced a real-time emotion classification system using multimodal physiological data sources, such as EDA and RB (respiratory belt).

Building on these studies, our work expands the analysis to the field of CW estimation. We aim to show that FL can be an effective approach for creating high-performance models that respect privacy while still benefiting from shared user data. We propose a CW-specific preprocessing on client devices that only relies on client dataset, and incorporates signal alignment, common resampling and data scaling. In this way, we avoid traditional analysis and preprocessing that need the collection of the entire dataset. Our results demonstrated that our framework, although subject to additional challenges compared to centralized CW estimation, can maintain high-performance. Such an approach holds significant promise in the healthcare sector, where wearable technology provides a wealth of sensitive information that can be leveraged to enhance decision-making, intervention, and prevention processes. To the best of our knowledge, this is the first study to apply FL methods in developing a privacy-aware model for CW estimation, with experiments spanning two distinct datasets.

3 DATASETS

This section describes the two datasets for CW estimation that we utilized in our study, COLET [18] and ADABase [28].

3.1 COLET Dataset

The COLET dataset contains eye-tracking data, including eye movement, pupil diameter, and blink metrics, collected from 47 test subjects participating in four distinct activities (A1/A2/A3/A4). The dataset is publicly available online [18, 19]. Data was recorded using the "Pupil Core" eye-tracker from Pupil Labs. The four activities were intended to evoke different levels of CW. Precisely, CW is elicited by two tasks: main task where the participants need to point out an object in an image divided in nine squares, and a secondary task in which the participants counted aloud backwards from 1000 deducting 4. Whereas, the four activities are obtained from a two-by-two factorial design with the factors time constraint and tasking. Thus, A1 is no time constraint and single task, A2 – time constraint and single task, A3 – no time constraint and multi-tasking, and A4 – time constraint and multi-tasking. The effect of the CW on one participant (ID32) can be observed on the plot (Figure 1) which depicts the pupil diameter for each activity, where the x-axis is the time, and the y-axis is the pupil diameter. In the boxes the mean value is annotated, and it can be verified that indeed the pupil diameter increases proportional to the CW. Upon completing each activity, participants were asked to complete a simplified version of the NASA Task-Load Index questionnaire, NASA-RTLX. This questionnaire is a widely used assessment tool for measuring the subjective workload experienced during task performance. Figure 2 displays a boxplot depicting the participants' responses to the questionnaire. The plot shows an increase in perceived CW across the four activities. However, there is considerable variability among the subjects. Our study focused on the classification task of separating two levels of CW, high (activity A1 and A2) and low (activity A3 and A4).

3.2 ADABase Dataset

The ADABase dataset utilized multimodal sensor data to estimate CW [28]. Its primary objective is to assess how CW varies across different driving challenges. The dataset includes physiological metrics such as ECG, EDA, EMG, PPG, respiration, skin temperature, and eye tracker data recorded during tasks. Behavioral measurements, depicted by action units from facial videos, and performance metrics like reaction times are also incorporated. After each task, the participants provided subjective feedback using NASA-TLX and NASA-RTLX questionnaires. Originally involving 51 participants, but due to privacy considerations, data from only 30 subjects is publicly available. Prospective researchers can access the dataset upon signing an EULA document. CW in the study is induced in two distinct ways: via the k-drive test conducted in a semi-autonomous driving simulation experiment and the n-back test. Participants observed an autonomously driving vehicle under varying levels of CW.

For the n-back task, the participants were presented sequentially with stimuli, such as letters, and asked to compare the current stimulus to one presented n items back in the sequence. The n-back test consists of six variations: single n-back and dual n-back, where $n \in \{1, 2, 3\}$. During the test, positive and negative hits, as well as reaction times for positive hits (from stimuli start until button press), are recorded. The primary task is visual, showing a blue square on the screen. The secondary task is auditory, involving prerecorded German consonants. Positive and negative hits are recorded, along with reaction times for positive hits (from stimuli start until button press).

The k-drive test, a novel introduction from the study, mirrors the n-back test. The primary task involves participants detecting three car actions. The secondary task requires participants to add

DOI: 10.1145/3626705.3627783
Published December 2023

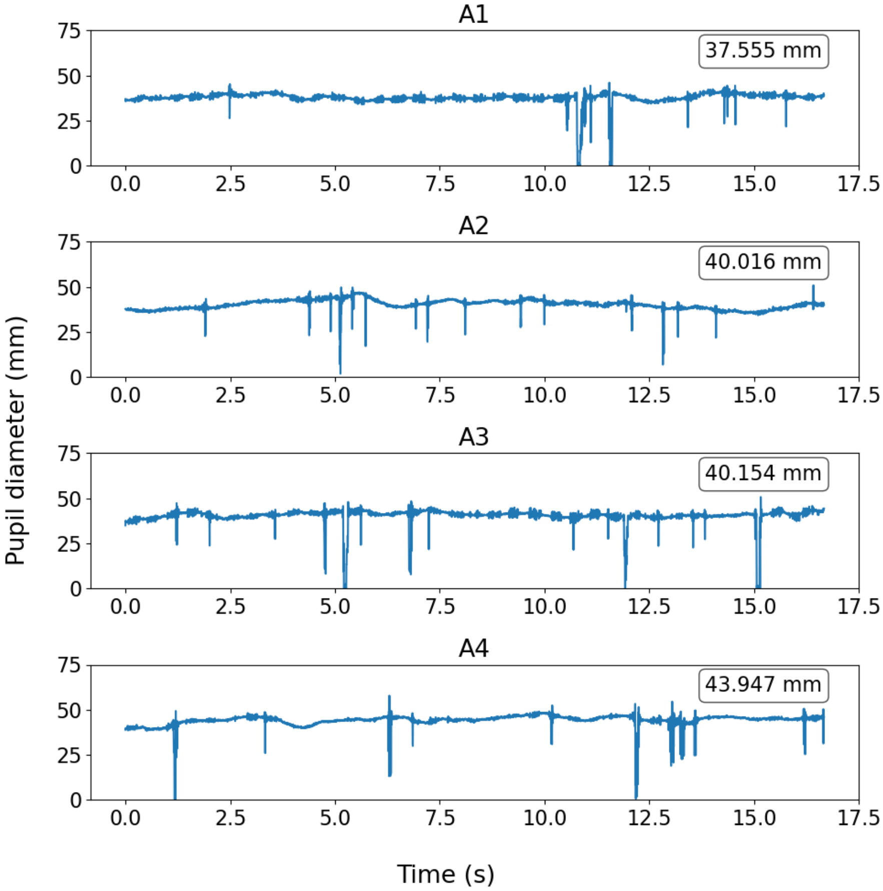


Figure 1: Pupil diameter in each activity. The boxes represent the average pupil diameter value.

a song to a playlist on a mobile app. The test uses k to denote the number of necessary user actions, where $k \in \{1, 2, 3\}$. Specifically, k=1 indicates a car overtaking, k=2 signifies the car being overtaken, and k=3 represents the car accelerating/decelerating. Levels 2 and 3 incorporate the secondary task.

Since ADABase has multiple different tests conducted, the CW can be encoded using various settings, as suggested in the paper [28]. Thus, we opted for encoding the CW in two classes where:

- low CW is n-back baseline $\{1, 2\}$, visual, audio-visual: $n \in \{1\}$ k-drive baseline $\{1, 2, 3\}$;
- high CW is visual audio-visual: $n \in \{2, 3\}$ and dual $k \in \{1, 2, 3\}$.

4 METHODS

This section presents the machine learning pipelines developed in our study. We first describe the common preprocessing of the data, then the development of centralized models replicating the state-of-the-art studies to serve as baselines for developing the FL models. The centralized and the FL approach use the same steps: preprocessing (segmentation, filtering, outlier removal), feature extraction, and model learning.

4.1 Preprocessing and Segmentation

To mitigate the inter-variability inherent in the sensor data, signals from both datasets underwent a subject-wise normalization to position them within a uniform, comparable range [32]. As highlighted in Figure 3(b) relating to the COLET dataset, user-specific scaling can be executed locally on the user's device, making it particularly compatible with the FL methodology. In Figure 4, (a) displays the average absolute pupil diameter, while (b) presents the user-scaled average pupil diameter across the four activities in the COLET dataset. These figures showcase the impact of user-specific scaling on pupil diameter, enhancing the discrimination between activities with varying CWs. Similar outcomes were observed for signals from the ADABase dataset. Each user's data was scaled using the StandardScaler from the scikit-learn library. Scaling offers a significant additional advantage as it accounts for the varying scales across the measured data types. Since these variables do not equally contribute to model fitting, scaling the data aids in normalizing all variables to comparable ranges of the values [18].

In the context of the COLET dataset, after the user-specific scaling, the data underwent a series of preprocessing steps. As depicted

DOI: 10.1145/3626705.3627783
Published December 2023

in Figure 3(b), the initial phase involved aligning the pupil measurements, originally captured at a frequency of 480 Hz. These were downsampled to 240 Hz to ensure consistency with other measurements. Notably, subsequent experimentation validated the viability of further downsampling to 120 Hz without compromising performance. Additionally, the encoding of participant blinking as a binary signal enhances interpretability, with values of 1 indicating blink occurrence and 0 representing non-blink at respective timestamps. The dataset is then segmented into windows of 5, 10, 20, and 30 seconds with 50% of overlap.

In the context of the ADABase dataset, after user-specific scaling, the initial stage involves harmonizing signals with diverse sampling frequencies, ranging from 100 Hz to 1000 Hz, through uniform downsampling to 50 Hz. This process aligns the signals across the dataset, preserving their performance integrity. Furthermore, participant blink events are encoded in a binary signal in the same manner as the COLET dataset. Subsequently, the dataset is transformed into a 3D array using sliding-window segmentation, employing 20-second windows with 5-second overlap, so it can be fed into the 1D Convolutional Neural Network. Notably, owing to computational constraints, an end-to-end approach is exclusively employed in experimentation, showcasing the dataset's readiness for further investigation.

Another important note to make is that from both datasets, data for one participant was removed because of suspicious data quality, but we still managed to preserve the performance of the models.

Table 1 presents summary information about the preprocessed datasets, such as the downsampled frequency, number of clients, recording length per client, class balance, and data volume. We can observe that the number of samples (timestamps) of the data acquisition of each participant in both datasets is comparable. However, after delving deeper, the first discrepancy that comes out is the class imbalance in the COLET dataset, whereas this has been taken care of in the ADABase dataset, as it is also explained in the paper [28].

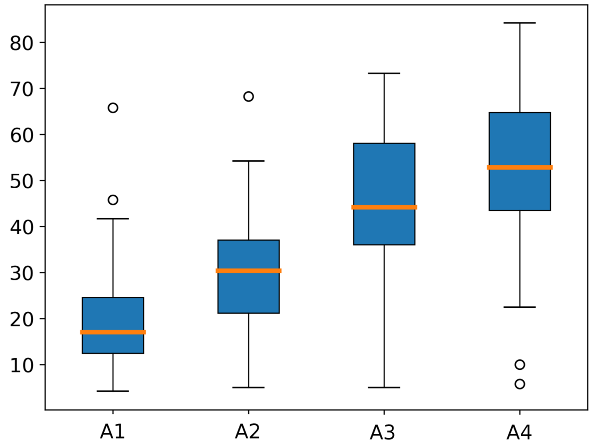


Figure 2: Boxplot of the NASA-TLX score. A higher NASA-TLX score indicates a higher CW.

This finding is also reflected in the average number of class labels per participant.

4.2 Feature Extraction

To replicate the state-of-the-art baselines reported in the original COLET study [18], we performed dataset-specific feature extraction. This enabled us to perform a detailed comparison of the proposed FL approach with the feature-based, centralized, state-of-the-art approaches. It is worth noting that feature-based models continue to hold significance in wearable computing, achieving state-of-the-art results in machine learning competitions [33]. This is probably because wearable computing datasets have limited data volume and intrinsic noise. This noise, primarily attributed to the relative movement between the sensor and the wearer, frequently results in faulty sensor readings. These challenges, compounded by the intricacies introduced by FL, might hinder the effective application of end-to-end learning for wearables.

As delineated in Figure 3(b), the preliminary stages of data preprocessing for feature-based models align with those of the end-to-end approach. Post this phase, we extracted statistical attributes of the signals and domain-specific metrics from each data window. The statistical features extracted include mean, standard deviation, range, skewness, kurtosis, differential mean, second-order differential mean, lower and upper quartiles, inter-quartile range, and the inverse coefficient of variation of the signals. For domain-specific metrics, we integrated descriptors for saccades (velocity, peak velocity, frequency, and duration), fixation metrics (frequency and duration), pupil diameter, gaze positioning, and its differential. In total, this approach yielded 71 distinctive features per data window. After the feature extraction process, any statistical outliers were automatically identified and eliminated. The features subsequently served as inputs for both the centralized and FL feature-based models, which we detail in the following subsections.

4.3 Machine Learning Models

4.3.1 Feature-based models. To replicate the state-of-the-art results reported in the original COLET study [18], we initially used feature-based ML models, such as Gaussian Naive Bayes (GNB) and Random Forest (RF) on the extracted features. We also included a Feed-Forward Neural Network (FFNN) suitable for training in FL scenarios. We used simple FFNN architecture with two hidden layers and one dropout layer. Each fully connected layer had 256 neurons, and the dropout rate was 25%. The output layer employs a sigmoid activation function, enabling binary classification to discern high and low CW categories. The FFNN was optimized with the Adam optimizer, fine-tuned with a learning rate of 0.0001, using the binary cross-entropy loss function. Demonstrating efficient resource management, a strategic callback mechanism halts training when performance plateaus for ten epochs.

4.3.2 End-to-end models. For the end-to-end approach regarding both datasets, we designed a neural network architecture combining a single 1D convolutional layer with the previous FFNN (ConvFFNN). These additional layers are designed to intrinsically extract features from the signals. Specifically, our architecture begins with a 1D convolutional layer equipped with 64 filters, each with a kernel size of 3. We used a ReLU activation function and

DOI: 10.1145/3626705.3627783
Published December 2023

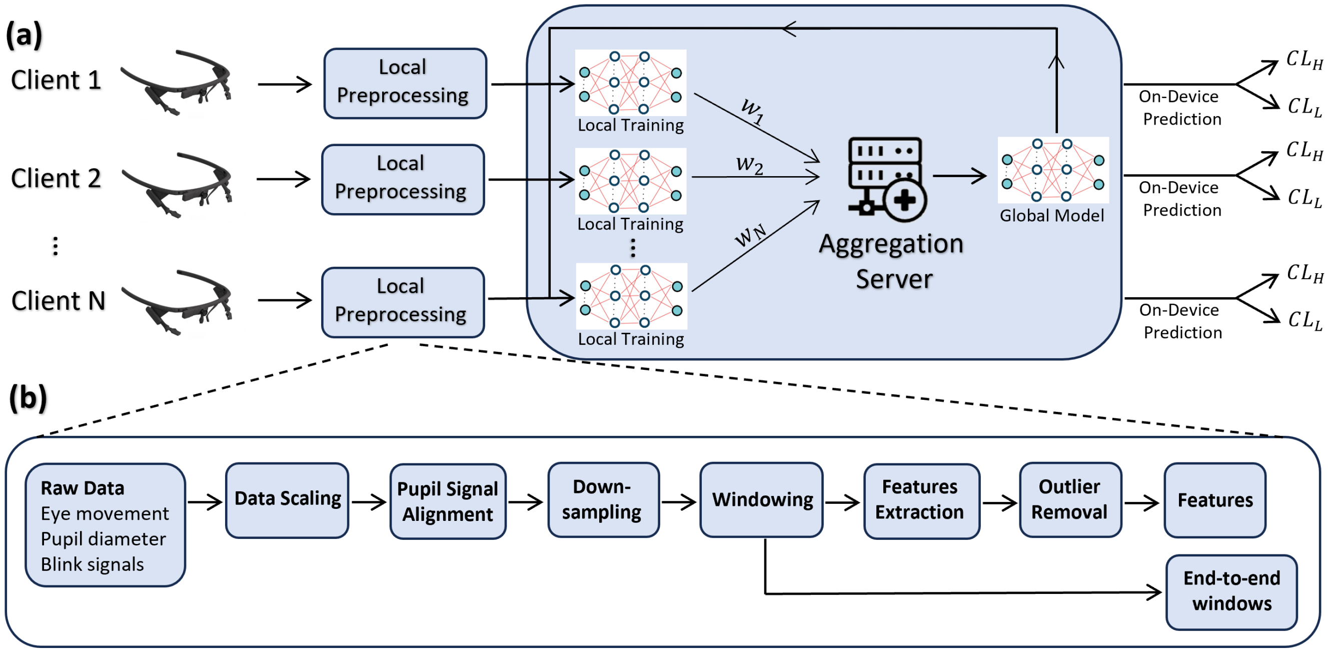
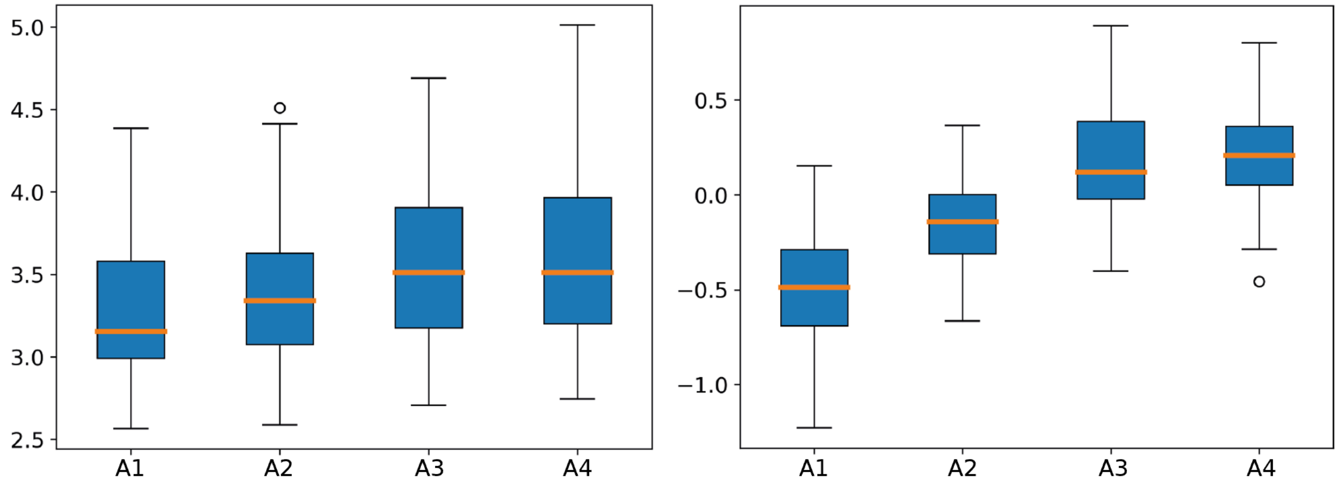


Figure 3: Federated Learning Pipeline for COLET. (a) Training Pipeline: Data from the smartglasses of each user is locally processed and trained. Subsequently, model weights from individual users are aggregated on the server to produce a global model, which is then used for on-device predictions. (b) User-Preprocessing Pipeline: Raw data undergoes several processing steps, including scaling, pupil alignment, downsampling, windowing, feature extraction, and outlier removal to either yield features or provide end-to-end windows.



(a) Pupil diameter without scaling

(b) Pupil diameter with scaling

Figure 4: The impact of user-specific scaling on the average pupil diameter across the four activities with increased CW in COLET: (a) Original unscaled data, and (b) Data after user-specific scaling.

adopted a padding strategy to ensure that the output has the same width as the original input. Following the convolutional layer, a max-pooling layer with a pool size of 2 is applied, effectively halving the size of the feature maps. After the pooling operation, the feature maps are flattened to form a 1D vector, which is then fed into the FFNN. The optimization process was the same as with the FFNN model, i.e., we optimized the ConvFFNN model using the Adam optimizer, with a learning rate of 0.0001, using the binary

Table 1: Summary statistics for the two datasets, COLET and ADABase

Statistic	COLET	ADABase
Downsampled frequency	120 Hz	50 Hz
Total Number of Participants	46	29
Number of raw signals	45	56
Number of timestamps per window	2400	1000
Average Participant Split: Train-Val-Test	34-6-6	21-4-4
Total Low CW Labels	35.10%	46.99%
Total High CW Labels	64.90%	53.01%
Total Number of 20 sec-Windows	1682	2926
Average Number of 20 sec-Windows/Participant	36.57 ± 11.74	100.90 ± 4.74

cross-entropy loss function. The same callback mechanism with patience of ten epochs was also used.

4.4 Federated Machine Learning

Figure 3(a) shows the main concept of the FL where only the weights of the neural network models are transmitted to the server and are aggregated equally to a single global model. This FL concept is applicable both for the feature-based model (FFNN), and for the end-to-end learning model (ConvFFNN).

We employed the FL approach using the weighted Federated Averaging algorithm. In this implementation, the updates from each client are weighted relative to the number of samples in each client's data before the aggregation on the server. This ensures that each client contributes equally during the federated training. We chose this method because the clients in our datasets exhibit an inhomogeneous distribution of data, leading to varying numbers of samples per client.

Experiments were conducted with varying numbers of participating clients in each communication round to investigate the impact of client participation on the FL process. Specifically, the experiments were performed with 5, 10, 20, and 30 clients in each round of the training. For each participation value, 100 rounds of communication were executed. A binary cross-entropy loss function was employed during model training. The clients used the Adam optimizer, with learning rates of 0.001. The final model selected was the one that exhibited the best accuracy on the validation set during training. The same training settings were used for both datasets.

5 RESULTS

We performed experiments on two separate datasets, COLET and ADABase. We used the COLET dataset to perform a detailed comparison of the FL models with machine learning models based on the original COLET study [18]. These experiments involve several feature-based models, several segmentation window sizes, and several participation rates (for the FL scenario). The results of these experiments are presented in the subsections 5.2 and 5.3. Once we confirmed the suitability of the FL approach and the suitable hyperparameters on the COLET dataset, we performed an additional comparison of the best-performing end-to-end FL model with the corresponding centralized version. The results of these experiments are presented in the subsection 5.4.

5.1 Experimental Setup

Our study focused on the classification task of separating two levels of CW, high and low, for both COLET and ADABase datasets. Given the size of our dataset and the hardware at our disposal, we opted for 4-fold cross-validation rather than Leave-One-Out validation to limit the computational time required. With 4-fold cross-validation, each experiment was run four times, reducing the computational burden compared to the 29 and 46 times required for Leave-One-Out validation (where 29 and 46 represent the number of users in ADABase and COLET, respectively). Each iteration of our experiments lasted approximately 12 hours. This allowed us to assess the models' ability to generalize to completely unseen users (person-independent models). Using 4-fold cross-validation, the dataset was split into four exclusive groups of users, ensuring that the data coming from one user was left out of the training only once. For each fold, the clients not used for training were equally split into validation and test sets. The model with the best validation accuracy was tested on the unseen test clients.

The performance of the models was evaluated based on the mean and standard deviation of accuracy, as well as the F1-score, Precision, and Recall. The inclusion of these metrics is especially valuable for the COLET dataset, where a higher class imbalance is present. All experiments were performed on a workstation featuring a GPU–NVIDIA-RTX-A5000 (with 24 GB VRAM).

5.2 Centralized Machine Learning — COLET Dataset

Table 2 presents the experimental results for the centralized models. Our optimal combination of model and window size surpassed the results presented in the original paper [18]. However, these results might not be directly comparable because the original study does not specify if the authors utilized window-based feature extraction or if they calculated features across the entire duration of each subject's recordings for a given task. This variation might account for the observed differences in results.

A positive correlation exists between window size and accuracy across all models. This observation is consistent with the insights from [32], suggesting that larger window sizes typically yield enhanced accuracy in CW detection tasks. One possible explanation is that features derived from more extensive windows are potentially less susceptible to noise compared to those from narrower windows.

DOI: 10.1145/3626705.3627783
Published December 2023

Table 2: Experimental results comparison between [18]'s best score (k-NN COLET), our centralized models, and the FFNN model. All models are person-independent.

Classifier	Window Size (s)	Accuracy	σ
k-NN (COLET)	–	90.00%	–
RF	**30**	**97.54%**	**1.16%**
RF	20	95.38%	2.51%
RF	10	92.76%	2.23%
GNB	**30**	**94.02%**	**1.91%**
GNB	20	93.05%	1.65%
GNB	10	85.57%	3.88%
ConvFFNN	**30**	**97.21%**	**1.50%**
ConvFFNN	20	94.70%	2.04%
ConvFFNN	10	91.34%	2.82%
FFNN	30	94.81%	4.92%
FFNN	**20**	**95.07%**	**1.34%**
FFNN	10	91.46%	2.09%

In our experiments, the Random Forest (RF) model with a 30-second window achieved the topmost performance, with an accuracy of 97.54%. Similar outcomes were observed for the ConvFFNN, which achieved its best accuracy of 97.21% with a 30-second window. For the FFNN, the results between the 20-second and 30-second windows did not show significant statistical variance. Interestingly, the peak accuracy for the FFNN was attained with the 20-second window, marked at 95.07%, in contrast to the 94.81% with the 30-second window. However, as indicated in Table 2, the 30-second window for the FFNN demonstrated a greater variance than the 20-second window.

In the following experiments, we focused only on the end-to-end learning model (ConvFFNN) given that it does not require feature-extraction steps, and yet it performs on par with the best-performing feature-based models.

5.3 Federated Learning — COLET Dataset

In this section, we build upon the best combination of window size and neural network model, identified during centralized learning (subsection 5.2). Here, we explore the influence of the client participation rate (C) within the federated environment, examining the end-to-end (ConvFFNN) approach. Table 3 shows that FL models benefit from an increased participation rate in each round. Indeed, the highest performances were achieved with 30 out of 30 participants per round with an accuracy of 95.40% ± 1.42% and an F1-score of 94.68% ± 1.77%. Furthermore, since all participation rates were tested with equal communication rounds, it can be inferred that training with a higher participation rate resulted in faster convergence. This observation is further solidified by Figure 5, which depicts the average accuracy across training phases for all participation rates over communication rounds. A participation rate of 5 manifested a slightly oscillatory curve and lower accuracy, whereas higher rates achieved more steady convergence.

5.4 Federated Learning — ADABase Dataset

In the context of ADABase, our primary comparison focused on the end-to-end approach, examining both our centralized and federated models. Based on our findings from the COLET dataset, where the end-to-end approach demonstrated comparable performance to the feature-based approach, we decided to exclusively experiment with the end-to-end approach in the ADABase dataset. Our objective was to compare these results with those presented in the original paper [28]. It's worth acknowledging that various features were derived from the raw signals, including but not limited to mean heart rate, standard deviation of successive NN intervals, and Root Mean Square of Successive Differences (RMSSD) from ECG, RMS, maximum amplitude and number of onsets per minute from EMG. From the EMG data, they derived features such as RMS, maximum amplitude, and the number of onsets per minute. Additionally, some of the features derived from EDA included mean amplitude values of SCR peaks and changes in skin conductance level within a window. The eye-tracking signals provided statistical features for fixations, saccades, blinks, and pupil measurements. Furthermore, they incorporated the number of active action units within a window. For a more comprehensive understanding of each of these derived features, we encourage the reader to refer to the original paper [28].

Notably, as displayed in Table 4, our centralized model (ConvFFNN) outperformed the results achieved with specific feature extraction and an XGBoost classifier from the original study [28], with a consistent improvement of 3.32% in F1-score. It's worth mentioning that while their classifier was evaluated using nested 10-fold cross-validation splitting the data subject-wise, our approach utilized 4-fold cross-validation with person-independent splitting across subjects. Although the evaluation methods are not directly comparable, the improvement in performance is noteworthy. This result confirms the advantages of using a convolution-based method

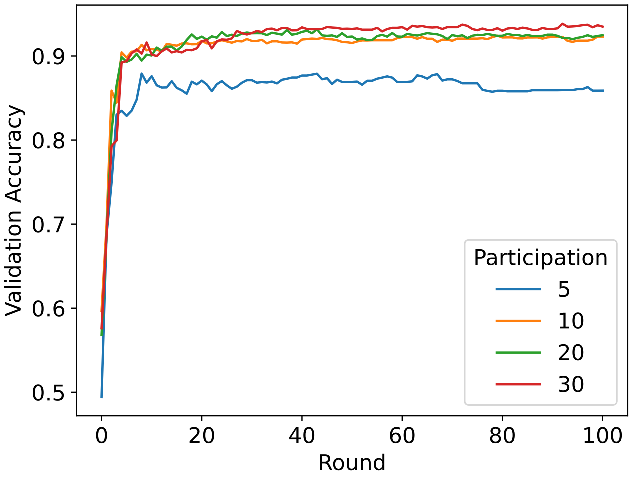


Figure 5: Plots of mean accuracy from different participation rates for COLET.

DOI: 10.1145/3626705.3627783
Published December 2023

Table 3: Cross-validation results with our end-to-end federated neural network (COLET).

Classifier	Part. rate	Window (s)	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
ConvFedNN	5	30	90.13 ± 3.91	87.62 ± 4.62	90.48 ± 3.36	89.00 ± 3.87
ConvFedNN	10	30	94.15 ± 3.00	93.18 ± 3.19	94.17 ± 3.17	94.36 ± 2.79
ConvFedNN	20	30	92.22 ± 1.20	90.53 ± 1.41	92.27 ± 1.55	91.63 ± 1.43
ConvFedNN	**30**	**30**	**95.40 ± 1.42**	**94.68 ± 1.77**	**95.04 ± 1.40**	**95.54 ± 1.83**

to extract features. As regards the end-to-end comparison, our ConvFFNN achieved an accuracy of 87.38%, while its federated counterpart (ConvFedNN) registered an accuracy of 85.58%. The obtained results for ADABase with higher standard deviation across the tested folds compared to COLET may suggest the presence of more diverse clients and measurements in the ADABase dataset.

This is also confirmed by Figure 6, which depicts the variation in accuracy and loss during the training of the 5 ConvFedNNs. Specifically, the plot illustrates the mean values along with the standard deviation (± σ) observed during the 4-fold cross-validation.

6 DISCUSSION

The findings from our study provide strong evidence that the FL approach can rival the performance of centralized, privacy-intrusive methods. In the end-to-end approach on the ADABase dataset, the disparity in accuracy between the two methods is 1.80 percentage points. On the COLET dataset, the difference was 1.81 percentage points. Nevertheless, the slight decrease in performance seems acceptable, given the increased privacy awareness of the FL models.

As also confirmed in Figure 7 , the confusion matrix for COLET shows that the federated end-to-end model is perfectly balanced between classes, demonstrating that the model is resilient to the class imbalance present in the original dataset. Contrarily, a small bias toward low CW is present in ADABase, indicating possibilities for improvement.

One explanation for the performance decrease might be "out-of-distribution clients", i.e., when training in a centralized fashion, those outlier clients are shuffled in the overall training data and considered noise without a big influence on the models. On the other hand, in the FL scenario, during local training with outlier clients, the local model trains only on those outlier samples, resulting in a bigger influence on the global model. In such scenarios, the simple weighted Federated Averaging might not be the optimal choice for the models' aggregation on the server. These findings accentuate the need for deeper exploration into federated configurations. In this context, an interesting topic for future research is exploring other FL algorithms, such as FedProx [22] and Personalized-FedAvg [8], in more realistic simulations that consider variations in the number of samples or users' importance. These algorithms seek to address issues associated with the original FedAvg algorithm, including fairness across devices and device heterogeneity.

Furthermore, CW estimation studies often face limitations in labeled data and dataset size, where feature-based models typically excel. In our experiments, the end-to-end models achieved high evaluation scores – on par with the state-of-the-art feature-based models. However, it should also be noted that for more fine-grained CW estimation, the quantity of the labelled data may have a bigger influence. Moreover, ADABase contains almost double the labeled data compared to COLET and data from multiple sensors, such as ECG, EDA, EMG, PPG, respiration rate and skin temperature in addition to the eye-tracker alone. Surprisingly, despite this substantial difference in data volume, the evaluation scores of federated models on ADABase are slightly lower, with an F1-score of 83.18% compared to COLET's 94.68%. This discrepancy suggests that factors such as data quality and the specific type of CW task undertaken by users may have a more pronounced effect on evaluation scores than the data volume. It's worth noting that this relationship might change if the size of the labeled data significantly decreases, such as in the case of only having a few hundred labeled samples.

Considering the importance of accurate CW estimation, future research should also focus on evaluating federated approaches for more detailed CW estimation. This could involve identifying more CW states or providing continuous CW estimates instead of discrete ones. However, it's essential to recognize that achieving higher discrimination among CW levels significantly increases the task's complexity. Therefore, to maintain high performance in these more complex tasks, researchers can explore the use of complementary techniques. Self-supervised or unsupervised methods, for instance, could enhance model performance by leveraging unlabeled data for training [14, 30]. For example, a recent work of Google, SimPer, offers an intuitive and flexible approach for learning robust feature representations from periodic signals, which could be instrumental in enhancing the labeled dataset and optimizing FL model performance [35]. Another avenue we intend to explore is the augmentation of convolutional layers within our network. While the primary objective of our study focused on contrasting centralized and FL, the parallel results obtained from both feature-based and end-to-end methods prompt us to consider whether a deeper automatic feature extraction could yield improved performance. A promising deep learning architecture might be the Spectro-Temporal ResNet, which learns both from the temporal and the spectral (frequency) representations of the input signals, and is specialized for multimodal sensor data [10].

Finally, we intend to integrate both datasets to establish a unified federated training system. This fusion is anticipated to foster a global model with an enhanced capacity for generalization across diverse user profiles, adeptly addressing data and sensor heterogeneity. As part of this effort, we envisage simulating cross-silo FL, treating the two datasets as distinct institutions, and cross-client scenarios by incorporating clients from varied datasets, each contributing a different spectrum of sensor inputs.

DOI: 10.1145/3626705.3627783
Published December 2023

Table 4: Comparison of cross-validation results on ADABase between the best, centralized model (XGB) from [28] and our models (ConvFFNN and federated ConvFedNN). The participation rate for FL training was 21 out of 21, with a window size of 20 seconds.

Classifier	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
XGB (ADABase)	—	82.00 ± 6.00	—	—
ConvFFNN	87.38 ± 6.07	85.32 ± 8.15	85.64 ± 7.31	85.02 ± 7.20
ConvFedNN	85.58 ± 7.25	83.18 ± 8.71	84.32 ± 9.04	85.65 ± 7.21

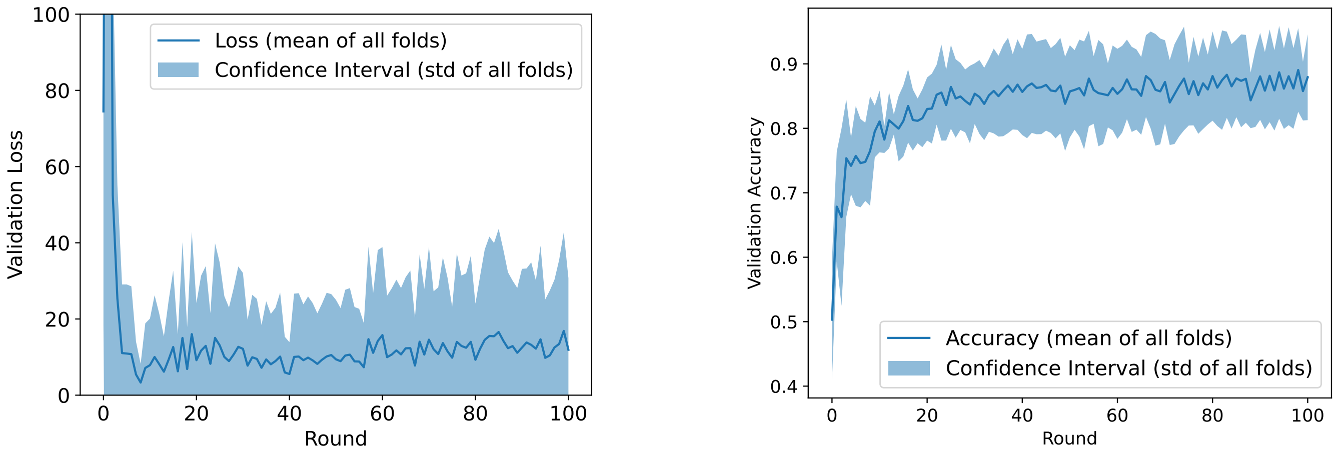


Figure 6: Learning curves during the federated training (ADABase).

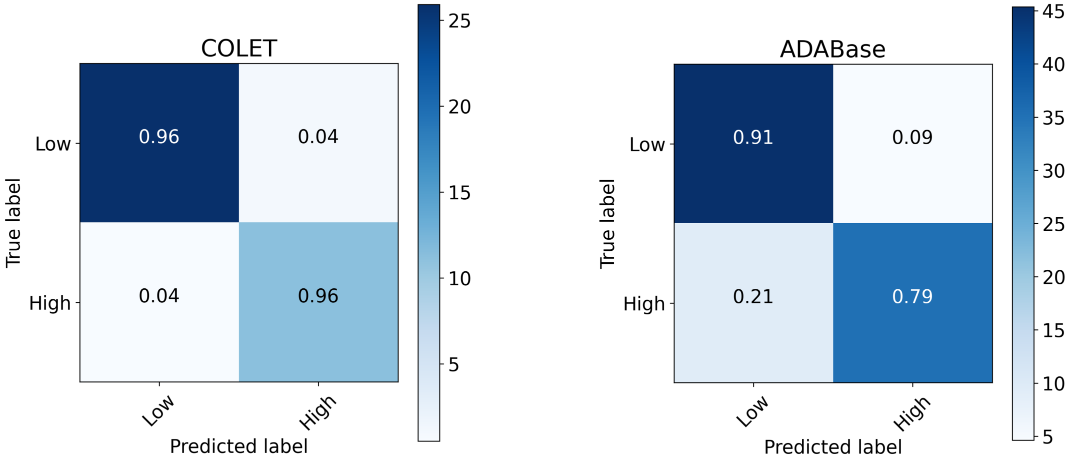


Figure 7: Confusion matrices for the ConvFedNNs in COLET and ADABase.

7 CONCLUSION

This study tackled the complex task of estimating CW using FL methods that prioritize user privacy. We compared the outcomes of the FL models with the more conventional centralized approach. Our experimental results highlight the efficacy of FL in collaboratively training a global model, yielding performances on par with centralized state-of-the-art machine learning models. Recognizing the importance of user privacy in user sensing, FL presents a promising approach that enables wearable sensing applications in privacy-sensitive domains.

More specifically, on the COLET dataset, the person-independent, end-to-end FL model achieved an accuracy of 95.40% for recognizing two levels of CW (high vs. low). For comparison, the centralized version of the same model achieved an accuracy of 97.21%. To further test the proposed FL approach, we tested the end-to-end models on a second dataset (ADABase), where a similar trend was observed.

DOI: 10.1145/3626705.3627783
Published December 2023

The end-to-end learning model obtained an accuracy of 87.38% and 85.58% using centralized and federated training, respectively. These findings emphasize the potential of FL systems in safeguarding user data without substantially compromising performance.

ACKNOWLEDGMENTS

This study was partially supported by the Swiss National Science Foundation through "TRUST-ME: TRUstworthy enhancement of job SaTisfaction and productivity using Micro-sensing in work Environments" under project 205121L_214991; and by the SmartCHANGE project, funded within the Horizon Europe R&I programme under the GA No. 101080965.

REFERENCES

[1] Yekta Said Can and Cem Ersoy. 2021. Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring. *ACM Transactions on Internet Technology (TOIT)* 21, 1 (2021), 1–17. https://doi.org/10.1145/3428152

[2] Maher Chaouachi and Claude Frasson. 2012. Mental workload, engagement and emotions: an exploratory study for intelligent tutoring systems. In *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11*. Springer, 65–71. https://doi.org/10.1007/978-3-642-30950-2_9

[3] Stéphane Delliaux, Alexis Delaforge, Jean-Claude Deharo, and Guillaume Chaumet. 2019. Mental workload alters heart rate variability, lowering non-linear dynamics. *Frontiers in physiology* 10 (2019), 565. https://doi.org/10.3389/fphys.2019.00565

[4] Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13. https://doi.org/10.1145/3173574.3173856

[5] Daniel E. Ehrmann, Sara N. Gallant, Sujay Nagaraj, Sebastian D. Goodfellow, Danny Eytan, Anna Goldenberg, and Mjaye L. Mazwi. 2022. Evaluating and reducing cognitive load should be a priority for machine learning in Healthcare. *Nature Medicine* 28, 7 (2022), 1331–1333. https://doi.org/10.1038/s41591-022-01833-z

[6] Castro Elizondo Jose Ezequiel, Martin Gjoreski, and Marc Langheinrich. 2022. Federated learning for privacy-aware human mobility modeling. *Frontiers in Artificial Intelligence* 5 (2022), 867046. https://doi.org/10.3389/frai.2022.867046

[7] Marzieh Salehi Fadardi, Javad Salehi Fadardi, Monireh Mahjoob, and Hassan Doosti. 2022. Post-saccadic Eye Movement Indices Under Cognitive Load: A Path Analysis to Determine Visual Performance. *Journal of Ophthalmic & Vision Research* 17, 3 (2022), 397. https://doi.org/10.18502/jovr.v17i3.11578

[8] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems* 33 (2020), 3557–3568.

[9] Dashan Gao, Ce Ju, Xiguang Wei, Yang Liu, Tianjian Chen, and Qiang Yang. 2019. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *arXiv preprint arXiv:1909.05784* (2019). https://doi.org/10.48550/arXiv.1909.05784

[10] Martin Gjoreski, Vito Janko, Gašper Slapničar, Miha Mlakar, Nina Reščič, Jani Bizjak, Vid Drobnič, Matej Marinko, Nejc Mlakar, Mitja Luštrek, et al. 2020. Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors. *Information Fusion* 62 (2020), 47–62. https://doi.org/10.1016/j.inffus.2020.04.004

[11] Martin Gjoreski, Tine Kolenik, Timotej Knez, Mitja Luštrek, Matjaž Gams, Hristijan Gjoreski, and Veljko Pejović. 2020. Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences* 10, 11 (2020), 3843. https://doi.org/10.3390/app10113843

[12] Martin Gjoreski, Bhargavi Mahesh, Tine Kolenik, Jens Uwe-Garbas, Dominik Seuss, Hristijan Gjoreski, Mitja Luštrek, Matjaž Gams, and Veljko Pejović. 2021. Cognitive load monitoring with wearables–lessons learned from a machine learning challenge. *IEEE Access* 9 (2021), 103325–103336. https://doi.org/10.1109/ACCESS.2021.3093216

[13] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. https://doi.org/10.1177/154193120605000909

[14] Liangze Jiang and Tao Lin. 2023. Test-Time Robust Personalization for Federated Learning. *arXiv* (2023). https://doi.org/10.48550/arXiv.2205.10920 arXiv:2205.10920

[15] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210. https://doi.org/10.48550/arXiv.1912.04977

[16] Apostolos Kalatzis, Ashish Teotia, Vishnunarayan Girishan Prabhu, and Laura Stanley. 2021. A database for cognitive workload classification using electrocardiogram and respiration signal. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Neuroergonomics and Cognitive Engineering, Industrial Cognitive Ergonomics and Engineering Psychology, and Cognitive Computing and Internet of Things, July 25-29, 2021, USA*. Springer, 509–516. https://doi.org/10.1007/978-3-030-80285-1_58

[17] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. 2023. A survey on measuring cognitive workload in human-computer interaction. *Comput. Surveys* (2023). https://doi.org/10.1145/3582272

[18] Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S. Tachos, Evanthia Tripoliti, Dimitrios I. Fotiadis, and Manolis Tsiknakis. 2022. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine* 224 (2022), 106989. https://doi.org/10.1016/j.cmpb.2022.106989

[19] Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S. Tachos, Evanthia Tripoliti, Dimitrios I. Fotiadis, and Manolis Tsiknakis. 2022. COLET: A Dataset for Cognitive workLoad estimation based on Eye-Tracking (Version 2). (2022). https://doi.org/10.5281/zenodo.6801166

[20] Marc Langheinrich. 2001. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on ubiquitous computing*. Springer, 273–291. https://doi.org/10.1007/3-540-45427-6_23

[21] Ryan T. Li, Scott R. Kling, Michael J. Salata, Sean A. Cupp, Joseph Sheehan, and James E. Voos. 2016. Wearable Performance Devices in Sports Medicine. *Sports health* 8, 1 (2016), 74–78. https://doi.org/10.1177/1941738115616917

[22] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2 (2020), 429–450. https://doi.org/1812.06127/arXiv:1812.06127

[23] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497* (2019). https://doi.org/10.48550/arXiv.1905.10497

[24] Jessica Chia Liu, Jack Goetz, Srijan Sen, and Ambuj Tewari. 2021. Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. *JMIR mHealth and uHealth* 9, 3 (2021), e23728. https://doi.org/10.2196/23728

[25] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial intelligence and statistics*. PMLR, 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

[26] Caitlin Mills, Igor Fridman, Walid Soussou, Disha Waghray, Andrew M. Olney, and Sidney K. D'Mello. 2017. Put your thinking cap on: detecting cognitive load using EEG during learning. In *Proceedings of the seventh international learning analytics & knowledge conference*. 80–89. https://doi.org/10.1145/3027385.3027431

[27] Arijit Nandi and Fatos Xhafa. 2022. A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods* 204 (2022), 340–347. https://doi.org/10.1016/j.ymeth.2022.03.005

[28] Maximilian P. Oppelt, Andreas Foltyn, Jessica Deuschel, Nadine R. Lang, Nina Holzer, Bjoern M. Eskofier, and Seung Hee Yang. 2022. ADABase: A Multimodal Dataset for Cognitive Load Estimation. *Sensors* 23, 1 (2022), 340. https://doi.org/10.3390/s23010340

[29] P. Ramakrishnan, B. Balasingam, and F. Biondi. 2021. Cognitive load estimation for adaptive human-machine system automation. In *Learning control*. Elsevier, 35–58. https://doi.org/10.1016/b978-0-12-822314-7.00007-9

[30] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30. https://doi.org/10.1145/3328932

[31] John Sweller, Jeroen JG van Merriënboer, and Fred Paas. 2019. Cognitive architecture and instructional design: 20 years later. *Educational psychology review* 31 (2019), 261–292. https://doi.org/10.1007/s10648-019-09465-5

[32] Jaakko Tervonen, Kati Pettersson, and Jani Mäntyjärvi. 2021. Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. *Electronics* 10, 5 (2021), 613. https://doi.org/10.3390/electronics10050613

[33] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2021. Three-year review of the 2018–2020 SHL challenge on transportation and locomotion mode recognition from mobile sensors. *Frontiers in Computer Science* 3 (2021), 713719. https://doi.org/10.3389/fcomp.2021.713719

[34] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19. https://doi.org/10.48550/arXiv.1902.04885

DOI: 10.1145/3626705.3627783
Published December 2023

[35] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. 2022. Simper: Simple self-supervised learning of periodic targets. *arXiv preprint arXiv:2210.03115* (2022). https://doi.org/10.48550/arXiv.2210.03115

DOI: 10.1145/3626705.3627783
Published December 2023