A Limitations

- The selected class-conditional Gaussian distribution may not work well for all neural network architectures. For example, if the output features are the result of an activation such as ReLU, a truncated Gaussian distribution may be a better model. Future work can look to exploit knowledge of the neural network architecture to improve the accuracy of the feature distribution estimate.
- In this work, we leverage the insights from a fusion of global and local feature space. As in many applications there is often an underlying cluster structure between clients datasets, future works may explore the identification and efficient estimation of feature distributions of client clusters, in order to reduce the degree of bias introduced in client collaboration.

B Broader Impacts

Federated learning has become the main trend for distributed learning in recent years and has deployed in many popular consumer devices such as Apple's Siri, Google's GBoard, and Amazon's Alexa. Our paper addresses the practical limitations of personalization methods in adapting to clients with covariate shifts and/or limited local data, which is a central issue in cross-device FL applications. We are unaware of any potential negative social impacts of our work.

C Details of Experimental Setup

All experiments are implemented in PyTorch 2.1 [36] and were each trained with a single NVIDIA A100 GPU. Compute time per experiment ranges from approximately 2 hours for CIFAR10/100 and 20 hours for TinyImageNet. Code for re-implementing our method is provided at the following GitHub URL: `https://github.com/cj-mclaughlin/pFedFDA`.

C.1 Dataset Description

The EMNIST [4] dataset contains over 730,000 28×28 grayscale images of 62 classes of handwritten characters. The CIFAR10/CIFAR100 [22] datasets contain 60,000 32×32 color images in 10 and 100 different classes of natural images, respectively. TinyImageNet [23] contains 120,000 64×64 color images of natural images.

C.1.1 CIFAR-S Generation.

We implement the following 10 common image corruptions at 5 levels of severity as described in [14]: Gaussian noise, shot (Poisson) noise, impulse noise, defocus blur, motion blur, fog, brightness, contrast, frost, JPEG compression. We apply a unique corruption-severity pair to all samples of the first 50 clients.

C.1.2 Non-i.i.d. Partitioning.

On CIFAR and TinyImageNet datasets, we distribute the proportion of samples of class C across M clients according to a Dirichlet distribution: $q_c, m \sim \mathtt{Dir}_M(\alpha)$, where we consider $\alpha \in (0.1, 0.5)$ as in [30].

We provide a visualization of Dirichlet partitioning strategies on CIFAR10 below. The size of each point represents the number of allocated samples. Notably, as α increases, $\mathrm{Dir}(\alpha)$ becomes less heterogeneous.

C.2 Training Settings

All methods are trained using mini-batch SGD for 200 global rounds with 5 local epochs of training. We use a fixed learning rate of 0.01, momentum of 0.5, and weight decay of 5e-4. The batch size is set to 50 for all experiments, except for EMNIST, where we use a batch size of 16. We sample the set of active clients uniformly with probability q=0.3 for CIFAR and TinyImageNet and q=0.03 for EMNIST. The last global round of training employs full client participation. We split the data of each client 80-20% between training and testing.

15

(a) CIFAR10 Dirichlet(0.1) (b) CIFAR10 Dirichlet(0.5)
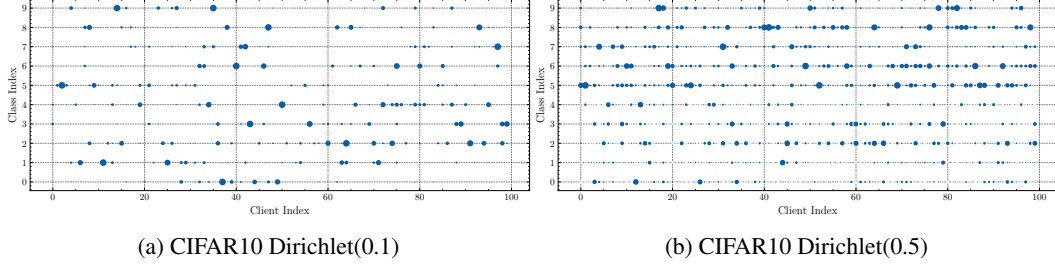
Figure 3: Comparison of Dirichlet Partitions on CIFAR10.

Hyper-parameters. For APFL, we tune α over [0.25, 0.5, 0.75, 1.0], and set $\alpha = 0.25$. For pFedMe, we tune λ over [1.0, 5.0, 10.0, 15.0] and set $\lambda = 5.0$. For Ditto, we use five local epochs for personalization and tune μ over [0.05, 0.1, 0.5, 1.0, 2.0] and set $\mu = 1.0$. For FedRep and FedBABU, we use five local epochs for training the head parameters. For FedPAC, we tune λ over [0.1, 0.5, 1.0, 5.0, 10.0], and set $\lambda = 1.0$. FedPAC uses one local epoch for training head parameters with a higher learning rate of 0.1, following the original implementation.

C.3 Evaluation on New Clients

Our fine-tuning procedure on new clients largely follows the methodology above. For FedAvgFT, we fine-tune the global model for five local epochs. For FedBABU and FedPAC, we personalize the model in 2 different ways and report the best result: (1) fine-tuning only the head for 5 local epochs, and (2) fine-tuning both the body and head for 5 local epochs. For pFedFDA, each new client estimates their local interpolated statistics (i.e., lines 8-11 of Algorithm 1) to obtain a personalized generative classifier.

For our covariate shift evaluation, we apply a medium severity corruption (level 3) to all samples.

D Additional Results

D.1 Multi-Domain FL

In Table 7, we present results on the DIGIT-5 domain generalization benchmark [53]. This presents an alternate form of covariate shift, as the data from each client is drawn from one of 5 datasets (SVHN, USPS, SynthDigits, MNIST-M, and MNIST). In particular, we use 20 clients trained with full participation, and assign 4 clients to each domain. Within each domain, we use the Dirichlet(0.5) partitioning strategy to assign data to each client. We observe that pFedFDA is effective in all settings, but has the most significant benefits over prior work in the low-data regime.

Table 7: Results on multi-domain DIGIT-5 benchmark for varying data volumes.

DIGIT-5 % Samples	25	50	75	100	Avg. Improvement
Local	76.84	83.11	86.97	88.51	-
FedAvg	81.75 (+4.91)	85.09 (+1.98)	87.41 (+0.44)	88.19 (+0.32)	1.91
FedAvgFT	85.61 (+8.77)	88.72 (+5.61)	90.75 (+3.78)	**91.73 (+3.22)**	5.34
Ditto	83.85 (+7.01)	85.53 (+2.42)	87.43 (+0.46)	88.80 (+0.29)	2.54
FedPAC	82.78 (+5.94)	87.94 (+4.83)	**91.12 (+4.15)**	91.04 (+2.53)	4.36
pFedFDA	**86.54 (+9.70)**	**90.05 (+6.94)**	90.75 (+3.78)	91.56 (+3.05)	**5.86**

D.2 Effect of Local Epochs

In many FL settings, we would like clients to perform more local training between rounds to reduce communication costs. However, too much local training can cause the model to diverge. In Fig. 4, we compare the effect of the local amount of epochs for CIFAR100 and CIFAR100-S-25% sample datasets. We observe that (1) pFedFDA outperforms FedAvgFT at all equivalent budgets of E, (2)

16

both methods follow exhibit a general plateau in accuracy after $E = 5$, and (3) pFedFDA learns much faster than FedAvgFT, with significantly higher accuracy for $E = 1$.
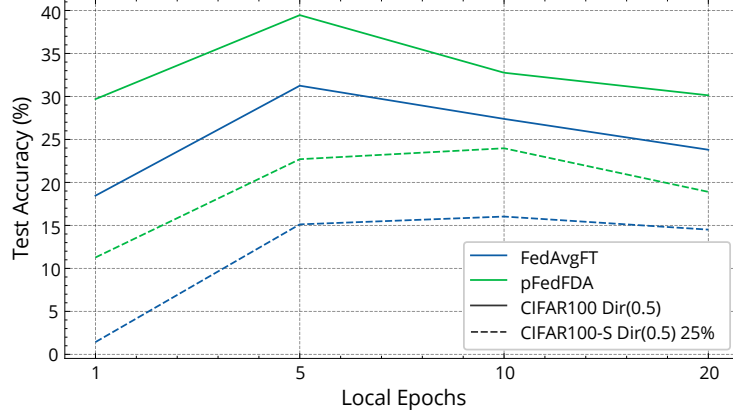
Figure 4: Comparison of average test accuracy with varying local epochs on CIFAR100.

D.3 Communication Load Examples

In Table 8, we compare the number of distinct parameters in our Gaussian estimates to that of a typical linear classifier for the models and datasets used in this paper, along with some additional examples. We display the resulting overhead relative to the base parameter count of the shared representation backbone.

Table 8: Comparison of communication load (parameters/iter.) between our Gaussian distribution parameters (μ, Σ) and standard linear classifiers.

Parameters	Backbone	Linear Classifier $(C \times (d + 1))$	Gaussian (μ, Σ) $(C \times d + \frac{1}{2}(d^2 + d))$	Δ Overhead
EMNIST-CNN (EMNIST, $C = 62$)	115776	7998	16192	6.620%
CIFAR-CNN (CIFAR100, $C = 100$)	106400	12900	21056	6.837%
ResNet18 (TinyImageNet, $C = 200$)	11167212	102600	233728	1.164%
MobileNetV3-Small (ImageNet, $C = 1000$)	927008	1615848	1548800	-2.637%

D.4 Runtime of Method Components

In Table 9, we evaluate the proportion of each local iteration of pFedFDA associated with each line of our algorithm. **Network Passes** refers to the time taken to train the base network parameters ϕ (Line 7 of Alg. 1). **Mean/Covariance Est.** refers to the time taken to estimate the local mean and covariance from features extracted during model training (Line 8 of Alg. 1). **Interpolation Optimization** refers to the time taken to optimize the local coefficient β (Line 9 of Alg. 1). Overall, we find that the majority of the overhead of our method comes from estimating the interpolation parameter β.

Table 9: Percentage (average (std)) of the local training time associated each component of our algorithm.

	Network Passes	Mean/Covariance Est.	Interpolation Optimization
CIFAR10	84.88 (6.79)%	0.765 (0.281)%	14.36 (6.62)%
CIFAR100	77.70 (5.75)%	2.861 (0.899)%	19.43 (5.70)%
TinyImageNet	87.41 (1.50)%	2.701 (0.659)%	9.886 (1.14)%

17

E   On the Bias-Variance Tradeoff

This section justifies the bias-variance tradeoff under some simplified technical assumptions. For simplicity, we assume that at any given round, the extracted feature vectors for a class are independent. We illustrate the bias-variance tradeoff in estimating the mean feature of a given class c at round t.

Proof of Theorem 1. For ease of exposition, we drop the time index and class index.

Let i be an arbitrary client with local dataset size n_i of class c. Let N be the total data volume of class c over the entire FL system. Assuming that the distribution of client i's features z follow a multivariate Gaussian distribution $\mathcal{N}(\theta_i, \Sigma_i)$, and the global feature distribution follows $\mathcal{N}(\theta, \Sigma)$ where $\theta_g := \sum_{i=1}^{M} n_i \theta_i / (\sum_{i \in [M]} n_i)$, $\Sigma_g := \sum_{i=1}^{M} n_i^2 \Sigma_i / (\sum_{i \in [M]} n_i)^2$. Note θ_i, θ_g are deterministic parameters.

We denote the local and global mean estimates as:

$$\mu_i := \frac{1}{n_i} \sum_{j=1}^{n_i} z_i^j, \text{ and } \mu_g := \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{n_i} z_i^j.$$

Let $\widehat{\mu}_i$ be the local estimate that interpolates between local and global knowledge, defined as

$$\widehat{\mu}_i := \beta \mu_i + (1 - \beta) \mu_g. \tag{10}$$

We will focus on bounding the high probability local estimation error $\|\widehat{\mu}_i - \mathbb{E}[\mu_i]\|_2$.

Note that Eq.(10) can be further expanded as

$$\widehat{\mu}_i = \beta \mu_i + (1 - \beta) \left(\frac{n_i}{N} \mu_i + \sum_{i' \neq i} \frac{n_{i'}}{N} \mu_{i'} \right)$$

$$= (\beta + (1 - \beta) \frac{n_i}{N}) \mu_i + (1 - \beta) \sum_{i' \neq i} \mu_{i'}$$

$$= \gamma \mu_i + (1 - \beta) \bar{\mu},$$

where $\gamma := \beta + (1 - \beta) \frac{n_i}{N}$, and $\bar{\mu} := \frac{1}{N} \sum_{i' \neq i} \sum_{j=1}^{n_{i'}} \mu_{i'}^j$.

Thus $\mu_i \sim \mathcal{N}(\theta_i, \frac{1}{n_i} \Sigma_i)$ and $\bar{\mu} \sim \mathcal{N}(\frac{N\theta_g - n_i \theta_i}{N}, \frac{N\Sigma_g - n_i \Sigma_i}{N^2})$. Since μ_i and $\bar{\mu}$ are independent, we have

$$\widehat{\mu}_i - \theta_i \sim \mathcal{N}\left((1 - \beta)(\theta_g - \theta_i), \ \gamma^2 \frac{1}{n_i} \Sigma_i + (1 - \beta)^2 \frac{N\Sigma_g - n_i \Sigma_i}{N^2} \right).$$

Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We have

$$\widehat{\mu}_i - \theta_i = (1 - \beta)(\theta_g - \theta_i) + \bar{\Sigma}^{1/2} \boldsymbol{g},$$

where $\widehat{\Sigma}^{1/2}$ is the square root matrix of $\widehat{\Sigma} := \gamma^2 \frac{1}{n_i} \Sigma_i + (1 - \beta)^2 \frac{N\Sigma_g - n_i \Sigma_i}{N^2}$. It holds that

$$\|\widehat{\mu}_i - \theta_i\|_2^2 = (1 - \beta)^2 \|\theta_g - \theta_i\|_2^2 + 2(1 - \beta) \left\langle \theta_g - \theta_i, \widehat{\Sigma}^{1/2} \boldsymbol{g} \right\rangle$$

$$+ \left\langle \widehat{\Sigma}^{1/2} \boldsymbol{g}, \widehat{\Sigma}^{1/2} \boldsymbol{g} \right\rangle.$$

Taking the expectation with respect to the randomness in the Gaussian random variable $\boldsymbol{g}$ and by the law of total expectation, we have

$$\mathbb{E}\left[2(1 - \beta) \left\langle \theta_g - \theta_i, \widehat{\Sigma}^{1/2} \boldsymbol{g} \right\rangle \right] = 2(1 - \beta) \left\langle \theta_g - \theta_i, \widehat{\Sigma}^{1/2} \mathbb{E}[\boldsymbol{g}] \right\rangle = 0,$$

18

and

$$\mathbb{E}\left[\left\langle \widehat{\Sigma}^{1/2}\boldsymbol{g}, \widehat{\Sigma}^{1/2}\boldsymbol{g}\right\rangle\right] \overset{(a)}{=} \mathbb{E}\left[\boldsymbol{g}^\top \widehat{\Sigma}\boldsymbol{g}\right]$$

$$= \mathbb{E}\left[\boldsymbol{g}^\top \gamma^2 \frac{1}{n_i}\Sigma_i \boldsymbol{g}\right] + \mathbb{E}\left[\boldsymbol{g}^\top (1-\beta)^2 \frac{N\Sigma_g - n_i\Sigma_i}{N^2}\boldsymbol{g}\right]$$

$$= \gamma^2 \frac{1}{n_i}\operatorname{Tr}(\Sigma_i) + (1-\beta)^2 \frac{N\operatorname{Tr}(\Sigma_g) - n_i\operatorname{Tr}(\Sigma_i)}{N^2}.$$

where equality (a) holds because $(\widehat{\Sigma}^{1/2})^\top (\widehat{\Sigma}^{1/2}) = \widehat{\Sigma}$ as $\widehat{\Sigma}^{1/2}$ is symmetric.

By Hanson-Wright inequality [44, Theorem 6.2], we conclude that with probability at least $1-\delta$ (for any given $\delta \in (0,1)$),

$$\|\widehat{\mu}_i - \theta_i\|_2^2 \le (1-\beta)^2 \|\theta_g - \theta_i\|_2^2$$
$$+ \left(\frac{\beta^2}{n_i} + \frac{2\beta(1-\beta)}{N}\right)\operatorname{Tr}(\Sigma_i) + (1-\beta)^2 \frac{1}{N}\operatorname{Tr}(\Sigma_g)$$
$$+ 4\left\|\left(\frac{\beta^2}{n_i} + \frac{2\beta(1-\beta)}{N}\right)\operatorname{Tr}(\Sigma_i) + (1-\beta)^2 \frac{1}{N}\operatorname{Tr}(\Sigma_g)\right\|_{\mathrm{F}} \max\left\{\sqrt{\frac{\log 1/\delta}{c}}, \frac{\log 1/\delta}{c}\right\}$$

$$\overset{(b)}{\le} (1-\beta)^2 \|\theta_g - \theta_i\|_2^2$$
$$+ \frac{\beta^2 + 2\beta(1-\beta)}{n_i}\operatorname{Tr}(\Sigma_i) + \frac{(1-\beta)^2}{N}\operatorname{Tr}(\Sigma_g)$$
$$+ 4\left\|\left(\frac{\beta^2}{n_i} + \frac{1}{2}\right)\operatorname{Tr}(\Sigma_i) + (1-\beta)^2 \frac{1}{N}\operatorname{Tr}(\Sigma_g)\right\|_{\mathrm{F}} \max\left\{\sqrt{\frac{\log 1/\delta}{c}}, \frac{\log 1/\delta}{c}\right\}$$

$$\overset{(c)}{\le} (1-\beta)^2 \|\theta_g - \theta_i\|_2^2$$
$$+ \frac{2\beta - \beta^2}{n_i}\operatorname{Tr}(\Sigma_i) + \frac{(1-\beta)^2}{N}\operatorname{Tr}(\Sigma_g)$$
$$+ 4\left(\sqrt{\frac{\log 1/\delta}{c}} + \frac{\log 1/\delta}{c}\right)\left(\frac{2\beta - \beta^2}{n_i}\sqrt{\operatorname{Tr}(\Sigma_i^2)} + \frac{(1-\beta)^2}{N}\sqrt{\operatorname{Tr}(\Sigma_g^2)}\right)$$

$$\overset{(d)}{\le} (1-\beta)^2 \|\theta_g - \theta_i\|_2^2$$
$$+ \left[1 + 4\left(\sqrt{\frac{\log 1/\delta}{c}} + \frac{\log 1/\delta}{c}\right)\right]\left(\frac{2\beta}{n_i}\operatorname{Tr}(\Sigma_i) + \frac{(1-\beta)^2}{N}\operatorname{Tr}(\Sigma_g)\right),$$

where $c > 0$ is some absolute constant, inequality (b) holds as $a \cdot b \le \left(\frac{a+b}{2}\right)^2$, and $N \ge n_i$, inequality (c) holds because of triangular inequality $\|A + B\|_{\mathrm{F}} \le \|A\|_{\mathrm{F}} + \|B\|_{\mathrm{F}}$, that $\|A\|_{\mathrm{F}} = \sqrt{\|A\|_{\mathrm{F}}^2} = \sqrt{\operatorname{Tr}(A^\top A)} = \sqrt{\operatorname{Tr}(A^2)}$ if matrix A is symmetric, and that $\max\{a,b\} \le a + b$, inequality (d) holds because $\operatorname{Tr}(A^2) \le (\operatorname{Tr}(A))^2$ for positive semidefinite matrix A and that $\operatorname{Tr}(\Sigma_i), \beta^2 \ge 0$, and $\operatorname{Tr}(\Sigma_g)$ are by definition non-negative.

$$\square$$

The first term $(1-\beta)^2 \|\theta_g - \theta_i\|_2^2$ is the bias introduced when client i uses global knowledge; the smaller the β, the more bias introduced. The last term reveals the interaction of β and the tradeoff between local and global variance. When β approaches 0, we have the global feature variance $\operatorname{Tr}(\Sigma)$ reduced by the average of N global samples. When β approaches 1, we have local feature variance $\operatorname{Tr}(\Sigma_i)$ reduced by the average of only n_i local data. Thus the bias-variance tradeoff on client i crucially depends on the degree of local-global distribution shift, $\|\theta_g - \theta_i\|_2^2$, the local data volume n_i and its quality (i.e., Σ_i), and the volume and quality of the data across clients N, Σ_g.

19

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The scope of the paper is on an important topic of client model personalization in federated learning. We faithfully state our contributions in both the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and considerations of our Gaussian modelling of the feature space both in the main text and with additional notes in the appendix. While the focus of the paper is in improving client personalization in the challenging setting of data scarcity and client distribution shift, we additionally benchmark our method in more general settings to demonstrate the widespread applicability of our work. Finally, we provide an assessment of the communication and computation overhead of our method compared to state-of-art approaches.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the key assumptions in the main text. The missing proof is deferred to Appendix E.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental setups and review the hyperparameters in Section 5.4 and Appendix C.2. We additionally provide code and instructions to train our method.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our evaluations are based on open-accessed datasets that are publically available. An official implementation code is provided in the supplementary materials.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed experimental setups and review the hyperparameters in Section 5.4 and Appendix C.2. We additionally provide code and instructions to train our method.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

20

Answer: [Yes]

Justification: For our main experiments we include the standard deviation of client accuracies, and include std error bars in our ablation visualization of the method components.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please find the software/hardware specifications in Appendix C.

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The NeurIPS code of ethics is strictly enforced throughout our research.

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have discussed the broader impacts of our work in Appendix B. Please find details therein.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The existing assets used in this paper has been adequately cited or credited to.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our attached code is well documented and comes with a README file indicating how reviewers may set up our experiments and train the proposed method.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

21

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

22