Functional Imaging data classification with crowdsourced noisy labels

Shuoyang WANG[1], Grace Y. YI[2], and Guanqun CAO[3]

[1] *University of Louisville, Louisville, KY 40208, USA*

[2] *University of Western Ontario, London, ON N6A 3K7, Canada*

[3] *Michigan State University, East Lansing, MI 48824, USA*

Supplementary Material

In this supplementary material, we provide technical details, including additional assumptions, lemmas, the proofs for Theorems 1 and 2, and Proposition 1, and additional numerical results.

S1 Technical proofs

In this section, we show the proof of Theorems 1 and 2, and additional simulation results. Throughout the section, for any vectors $\boldsymbol{a} = (a_1, \ldots, a_K)^\intercal$ and $\boldsymbol{b} = (b_1, \ldots, b_K)^\intercal$ with length K, denote

$$\boldsymbol{a}^\intercal \log\left(\frac{\boldsymbol{a}}{\boldsymbol{b}}\right) = \sum_{k=1}^{K} a_k \log\left(\frac{a_k}{b_k}\right).$$

For notation simplicity and readability, we denote $\widehat{\boldsymbol{f}}^{\texttt{afDNN}} = \widehat{\boldsymbol{f}}^*$ as an estimator of $\boldsymbol{f}^*$.

S1.1 Additional assumptions

For function class $\mathcal{H}$, denote

$$(\widehat{k}, \widehat{u}) = \arg \max_{\substack{u=0,\ldots,q^{(k)} \\ (k,u),k=1\ldots,K}} \frac{t_u^{(k)}}{\tilde{\beta}_u^{(k)}}$$

as the roughest compositional function layer index pair, where

$$\tilde{\beta}_u^{(k)} = \beta_u^{(k)} \prod_{\ell=u+1}^{q^{(k)}} \beta_\ell^{(k)} \wedge 1 \text{ and } (\widehat{l}, \widehat{v}) = \arg \max_{\substack{(k,u),k=1\ldots,K, \\ u=0,\ldots,q^{(k)}}} \frac{t_u^{(k)}}{\tilde{\beta}_u^{(k)}(1+\alpha_k)}.$$

Let

$$\theta = \frac{(1+\alpha_{\widehat{l}})\tilde{\beta}_{\widehat{v}}^{(\widehat{l})}}{(1+\alpha_{\widehat{l}})\tilde{\beta}_{\widehat{v}}^{(\widehat{l})} + t_{\widehat{v}}^{(\widehat{l})}}, \ \nu = \frac{\theta t_{\widehat{u}}^{(\widehat{k})}}{\tilde{\beta}_{\widehat{u}}^{(\widehat{k})}\,(1+\tilde{\alpha})} \text{ , and } \quad \widetilde{\beta} = \max_{k=1\ldots,K} \left(\widetilde{\beta}_0^{(k)} \wedge 1 \right),$$

where $\widetilde{\alpha} = \min_k(\alpha_k \wedge 1)$. The following two assumptions characterize the DNN function class for densely observed functional data and discretely observed functional data, respectively, both of which are necessary to ensure fast convergence.

Assumption S1. *The DNN class $\mathcal{D}$ satisfies*

(a) $L \asymp \log n$;

(b) $n^{\theta/\rho} \lesssim J \lesssim n^{\nu}$;

(c) $\max_{1 \le \ell \le L} \gamma_\ell \asymp n^{\nu}$;

(d) $s \asymp n^{\nu} \log n$.

Assumption S2. *Let $\underline{N} = \min_{i=1,\ldots,n} N_i$ denote the least sampling frequency for the functional observations. For some phase transition point $N^* \in \mathbb{N}^+$, the bias of extracted scores satisfies $\sup_{j=1,\ldots,J} E|\widehat{\xi}_{ij} - \xi_{ij}| \lesssim N_i^{-\tau}$, and the DNN class $\mathcal{D}$ satisfies*

(a) $L \asymp (\log n)\mathbb{I}(\underline{N} \geq N^*) + (\log \underline{N})\mathbb{I}(\underline{N} < N^*)$;

(b) $n^{\theta/\rho}\mathbb{I}(\underline{N} \geq N^*) + \underline{N}^{\theta'/\rho}\mathbb{I}(\underline{N} < N^*) \lesssim J \lesssim n^{\nu}\mathbb{I}(\underline{N} \geq N^*) + \underline{N}^{\nu'}\mathbb{I}(\underline{N} < N^*)$;

(c) $\max_{1\leq \ell \leq L} \gamma_\ell \asymp n^{\nu}\mathbb{I}(\underline{N} \geq N^*) + \underline{N}^{\nu'}\mathbb{I}(\underline{N} < N^*)$;

(d) $s \asymp (n^{\nu}\log n)\mathbb{I}(\underline{N} \geq N^*) + (\underline{N}^{\nu'}\log \underline{N})\mathbb{I}(\underline{N} < N^*)$.

where $\theta' = \tau\widetilde{\beta}$, *and* $\nu' = \frac{\theta' t_{\widetilde{u}}^{(\widehat{k})}}{\widetilde{\beta}_{\widetilde{u}}^{(\widehat{k})}(1+\widetilde{\alpha})}$.

S1.2 Technical lemmas

For any fixed J, define the training loss difference between $\widehat{\boldsymbol{f}}^*$ and the global minimum over $\mathcal{D}$ with respect to the true labels Y_i as

$$\delta_n(J) = \mathrm{E}\left\{-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}\left(Y_i = k\right)\log \widehat{f}_k^*\left(\boldsymbol{\xi}_{iJ}\right) - \min_{\boldsymbol{f}\in\mathcal{D}} -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}\left(Y_i = k\right)\log f_k\left(\boldsymbol{\xi}_{iJ}\right)\right\}.$$

Lemma S1. *If a network* $\widehat{\boldsymbol{f}}^* \in \mathcal{D}(L, J, \boldsymbol{p}, s)$ *and* $\mathcal{D}(L, J, \boldsymbol{p}, s)$ *satisfies*

- $L \asymp \log n$,
- $\|\boldsymbol{p}\|_{\max} \asymp \max\{J, K, C_2 n^{\nu}\}$,
- $s \asymp C_3 n^{\nu}\log n$,

then

$$\mathrm{E}_{\boldsymbol{f}^*}\left\{\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J)\left[\left(C_0\epsilon^{-1}\right) \wedge \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)}\right)\right]\right\} \lesssim n^{-\theta}\log^3 n$$

for $\widehat{\boldsymbol{f}}^*$ *taking values on* $[a, b]$, $a < b \in \mathbb{R}$, *and* ϵ *is defined in Assumption 2(a).*

Lemma S2. *(Bos and Schmidt-Hieber (2022) Theorem 3.5) Given a fixed* $J \in \mathbb{N}^+$, *for*

any classifier $\widehat{\boldsymbol{f}}^*$*, we have*

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{f}^*}\left\{\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J)\left[C_0 \wedge \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)}\right)\right]\right\} \\
\leq\ & C^* n^{-\theta}\log^3 n + \delta_n(J) \\
&+4n^{-1}\left[68C_0(s+1)\log\left(2^{2L+6}n(L+1)K^3J^2s^L\right) + 272C_0 + 2K\left(\log n + C_0\right)\right],
\end{aligned}$$

where C^* *is some universal constant.*

In the sequel, define the empirical risk with respect to the proposed loss function (3.4):

$$\begin{aligned}
\widetilde{\delta}_n(J) \ =\ & \mathrm{E}\Bigg\{-\frac{1}{nR}\sum_{i=1}^{n}\sum_{r=1}^{R}\sum_{k=1}^{K}\mathbb{I}\left(\widetilde{Y}_i = k\right)\log\left[\widehat{\boldsymbol{A}}^{*(r)}\widehat{\boldsymbol{f}}^*\left(\boldsymbol{\xi}_{iJ}\right)\right]_k + \lambda\,\mathrm{tr}(\widehat{\boldsymbol{A}}^{*(r)}) \\
&- \min_{\boldsymbol{f}\in\mathcal{D},\boldsymbol{A}^{(r)}\in\mathcal{A}} -\frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(\widetilde{Y}_i^{(r)} = k)\log[\boldsymbol{A}^{(r)}\boldsymbol{f}(\boldsymbol{\xi}_{iJ})]_k + \lambda\,\mathrm{tr}(\boldsymbol{A}^{(r)})\Bigg\},
\end{aligned}$$

where J is considered as a fixed positive integer.

The following lemma shows an information upper bound of the KL divergence with respect to the classifier derived from the proposed loss function (3.4).

Lemma S3. *Assume Assumption 1. For any* $J \in \mathbb{N}^+$*, and for any classifier* $\widehat{\boldsymbol{f}}^*$ *obtained by the loss function in Equation (3.4), when* $\boldsymbol{\xi}_J$ *is defined on some* $[-M, M]^J \subset \mathbb{R}$ *for* $M > 0$*, we have*

$$\begin{aligned}
&\mathrm{E}\left\{\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J)\left[\left(C_0\epsilon^{-1}\right) \wedge \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)}\right)\right]\right\} \\
\lesssim\ & n^{-\theta}\log^3 n + \widetilde{\delta}_n(J) + R^{-1}\sum_{r=1}^{R}\mathrm{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right) \\
&+n^{-1}\left[s\log\left(2^{2L+6}n(L+1)K^3J^2s^L\right) + K\log n\right].
\end{aligned}$$

Proof. First, for any $\boldsymbol{f} = (f_k)_{k=1,\ldots,K}$ and $\boldsymbol{A}^{(r)}, r = 1, \ldots, R$, it satisfies that

$$\begin{aligned} \mathrm{E}\left(\sum_{k=1}^{K} \mathbb{I}\left(Y_i = k\right) \log f_k\left(\boldsymbol{\xi}_{iJ}\right)\right) &= \mathrm{E}\left(\sum_{k=1}^{K} \mathrm{E}\left(\mathbb{I}\left(Y_i = k\right) | \boldsymbol{\xi}_{iJ}\right) \log f_k\left(\boldsymbol{\xi}_{iJ}\right)\right) \\ &= \mathrm{E}\left(\sum_{k=1}^{K} f_k^*\left(\boldsymbol{\xi}_J\right) \log f_k\left(\boldsymbol{\xi}_{iJ}\right)\right), \end{aligned}$$

and

$$\begin{aligned} & \mathrm{E}\left(\sum_{k=1}^{K} \mathbb{I}\left(\widetilde{Y}_i^{(r)} = k\right) \log\left[\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right]_k\right) \\ = & \mathrm{E}\left(\sum_{k=1}^{K} \mathrm{E}\left(\mathbb{I}\left(\widetilde{Y}_i^{(r)} = k\right) | \boldsymbol{\xi}_{iJ}\right) \log\left[\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right]_k\right) \\ = & \mathrm{E}\left(\sum_{k=1}^{K} \left[\boldsymbol{A}^{*(r)} \boldsymbol{f}^*\left(\boldsymbol{\xi}_{iJ}\right)\right]_k \log\left[\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right]_k\right). \end{aligned}$$

Next,

$$\begin{aligned} & \left(\boldsymbol{A}^{*(r)} \boldsymbol{f}^*\left(\boldsymbol{\xi}_{iJ}\right)\right)^{\mathsf{T}} \log\left(\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) - \left(\boldsymbol{f}^*\left(\boldsymbol{\xi}_{iJ}\right)\right)^{\mathsf{T}} \log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) \\ = & \left(\boldsymbol{f}^*\right)^{\mathsf{T}} \left[\left(\boldsymbol{A}^{*(r)}\right)^{\mathsf{T}} \log\left(\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) - \boldsymbol{I}_k \log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right)\right] \\ = & \left(\boldsymbol{f}^*\right)^{\mathsf{T}} \left[\left(\boldsymbol{A}^{*(r)}\right)^{\mathsf{T}} \left\{\log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) + \frac{\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right) - \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)}{\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)} + o\left(\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right) - \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right)\right\}\right. \\ & \left. - \boldsymbol{I}_k \log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right)\right] \\ = & \left(\boldsymbol{f}^*\right)^{\mathsf{T}} \left[\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_k\right)^{\mathsf{T}} \log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) + \left(\boldsymbol{A}^{*(r)}\right)^{\mathsf{T}} \frac{\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right) - \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)}{\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)}\right. \\ & \left. + o\left(\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right) - \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right)\right], \end{aligned}$$

where the second equality is derived by the Taylor series expansion. Therefore, under Assumption 1 and Proposition 1, when $\boldsymbol{A}^{(r)} = \widehat{\boldsymbol{A}}^{*(r)}$ and $\boldsymbol{f} = \widehat{\boldsymbol{f}}^*$, since $\|\boldsymbol{f}^*\|_{\max} \leq 1$ and $\|\boldsymbol{f}\|_{\max} \leq 1$, it follows that

$$\left(\boldsymbol{A}^{*(r)} \boldsymbol{f}^*\left(\boldsymbol{\xi}_{iJ}\right)\right)^{\mathsf{T}} \log\left(\boldsymbol{A}^{(r)} \boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) - \left(\boldsymbol{f}^*\left(\boldsymbol{\xi}_{iJ}\right)\right)^{\mathsf{T}} \log\left(\boldsymbol{f}\left(\boldsymbol{\xi}_{iJ}\right)\right) = O\left(\mathrm{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right)\right).$$

The same bound can also be derived by letting $\boldsymbol{A}^{(r)} = \widehat{\boldsymbol{A}}^{*(r)}$ and

$$\boldsymbol{f} = \arg\min_{\boldsymbol{f}' \in \mathcal{D}} -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}\left(Y_i = k\right) \log f_k'\left(\boldsymbol{\xi}_{iJ}\right).$$

As a result, we obtain that

$$\widetilde{\delta}(J) - \delta(J) = O\left(R^{-1} \sum_{r=1}^{R} \operatorname{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right)\right),$$

where the result holds under Assumption 1. Substituting the $\delta(J)$ by $\widetilde{\delta}(J) - \widetilde{\delta}(J) + \delta(J)$ in Lemma S2 and adjusting the truncation constant as $C_0\epsilon^{-1}$, we complete the proof, where the constant ϵ is defined in Assumption 2. □

S1.3 Proof of Proposition 1

Proof. According to Theorem 1 in Tanno et al. (2019), it suffices to show that for any positive λ,

$$\left(\boldsymbol{A}^{*(1)}, \ldots, \boldsymbol{A}^{*(R)}\right) = \underset{\left\{\widehat{\boldsymbol{A}}^{*(r)} \in \mathcal{A}\right\}_{r=1}^{R}}{\arg\min} \lambda \operatorname{tr}\left(\sum_{r=1}^{R} \widehat{\boldsymbol{A}}^{*(r)}\right)$$

and this solution is unique. Since $\left(\boldsymbol{A}^{*(1)}, \ldots, \boldsymbol{A}^{*(R)}\right)$ is the maximizer of the likelihood function, $\left(\boldsymbol{A}^{*(1)}, \ldots, \boldsymbol{A}^{*(R)}\right)$ is the minimizer of the objective function. The uniqueness is straightforward, as if the solution to

$$\underset{\left\{\widehat{\boldsymbol{A}}^{*(r)} \in \mathcal{A}\right\}_{r=1}^{R}}{\arg\min} -\frac{1}{nR} \sum_{r=1}^{R} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(\widetilde{Y}_i^{(r)} = k) \log[\widehat{\boldsymbol{A}}^{*(r)} \boldsymbol{f}^*(\boldsymbol{\xi}_J^{(i)})]_k$$

is not unique, the uniqueness of $\arg\min_{\left(\widehat{\boldsymbol{A}}^{*(1)}, \ldots, \widehat{\boldsymbol{A}}^{*(R)}\right)} \operatorname{tr}\left(\sum_{r=1}^{R} \widehat{\boldsymbol{A}}^{*(r)}\right)$ still ensures the solutions are $\left(\boldsymbol{A}^{*(1)}, \ldots, \boldsymbol{A}^{*(R)}\right)$, thus the proof is complete.

□

S1.4 Proof of Theorem 1

Let the set of the effective inputs of the kth be grouped by

$$\mathcal{P}_k = \left\{ j : \xi_j \text{ is effective for } g_{0v}^{(k)} \text{ for all } v = 1, \ldots, d_1^{(k)} \right\}.$$

By definition of $\mathcal{G}$, $|\mathcal{A}_k| \leq t_0^{(k)} d_1^{(k)} < \infty$. Let $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{P}_k$ denote the index set of effective inputs among all K groups, where $|\mathcal{P}| \leq \sum_{k=1}^{K} t_0^{(k)} d_1^{(k)} < \infty$. Given any positive integer m, define $\mathcal{Q}_M = \{|\xi_j| \leq M \text{ for all } j \in \mathcal{P}\}$ to be the concentration set for all effective inputs.

We first focus on the convergence rate by considering $\mathcal{Q}_M$ and $\mathcal{Q}_M^c$ separately. Let M be relatively large, and write identity vector $\mathbf{1}_K = (1, \ldots, 1)^\intercal$ with length K. For any J, we have

$$\begin{aligned}
& \mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^\intercal \left\{ C_0 \mathbf{1}_K \wedge \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \right\} \right] \\
= \; & \mathrm{E}\left[(\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^\intercal C_0) \wedge \left\{ \boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^\intercal \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \right\} \right] \\
\leq \; & \mathrm{E}\left\{ C_0 \wedge \left[(\boldsymbol{f}^{\circledast}(\boldsymbol{\xi}) - \boldsymbol{f}^*(\boldsymbol{\xi}_J) + \boldsymbol{f}^*(\boldsymbol{\xi}_J))^\intercal \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\boldsymbol{f}^*(\boldsymbol{\xi}_J)} \cdot \frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \right] \right\} \\
\leq \; & \mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^\intercal \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\boldsymbol{f}^*(\boldsymbol{\xi}_J)} \right) \right] + \mathrm{E}\left[C_0 \wedge \left| (\boldsymbol{f}^{\circledast}(\boldsymbol{\xi}) - \boldsymbol{f}^*(\boldsymbol{\xi}_J))^\intercal \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \right| \right] \\
& + \mathrm{E}\left\{ C_0 \wedge \left[\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J) \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \right] \right\} \\
\leq \; & \sum_{k=1}^{K} \mathrm{E}\left[\frac{(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J))^2}{f_k^*(\boldsymbol{\xi}_J)} \right] \\
& + \sum_{k=1}^{K} \mathrm{E}\left[C_0 \wedge \left| (f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)) \log\left(1 + \frac{f_k^*(\boldsymbol{\xi}_J) - \widehat{f}^*(\boldsymbol{\xi}_J)}{\widehat{f}^*(\boldsymbol{\xi}_J)} \right) \right| \mathbb{I}(\mathcal{Q}_M) \right] \\
& + \mathrm{E}\left\{ C_0 \wedge \left[\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J) \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)} \right) \mathbb{I}(\mathcal{Q}_M) \right] \right\} + 2C_0 K \mathbb{P}(\mathcal{Q}_M^c), \qquad \text{(S1)}
\end{aligned}$$

where the first inequality holds because of $f_k^{\circledast} \leq 1$, the second inequality holds since for any $a, b, c \geq 0$, we have that $a \wedge (b+c) \leq (a+c) \wedge (b+c) = (a \wedge b) + c$. The last inequality holds for the first term since KL-divergence is upper bounded by χ^2-divergence.

For the first term in (S1), when J is relatively large, for defined in Assumption 2, we have

$$
\begin{aligned}
& \sum_{k=1}^{K} \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right)^2}{f_k^*(\boldsymbol{\xi}_J)}\right] \\
= & \sum_{k=1}^{K} \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right)^2}{f_k^*(\boldsymbol{\xi}_J)} \mathbb{I}\left(f_k^*(\boldsymbol{\xi}_J) > \epsilon\right)\right] \\
& + \sum_{k=1}^{K} \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right)^2}{f_k^*(\boldsymbol{\xi}_J)} \mathbb{I}\left(f_k^*(\boldsymbol{\xi}_J) \leq \epsilon\right)\right] \\
< & \sum_{k=1}^{K} \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right)^2}{\epsilon}\right] + \sum_{k=1}^{K} \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right)^2}{f_k^*(\boldsymbol{\xi}_J)}\right] \mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J) \leq \epsilon\right) \\
\lesssim & \ \zeta(J),
\end{aligned}
$$

where the second asymptotic inequality is derived by using Lemmas 7.7 and 7.8 in Wang and Cao (2024) and Theorem 3.2 in Bos and Schmidt-Hieber (2022).

For any $a, b \in \mathbb{R}$ and fixed C_0, there exists C_0', such that

$$
|a \log(1+b)| \wedge C_0 \leq |a \log(1+b)| \mathbb{I}(|b| \leq C_0').
$$

Therefore, by the fact that given M, within the concentrated set $\mathcal{Q}_M$, there exists an M^* such that $\left[\widehat{f}^{(k)}(\boldsymbol{\xi}_J)\right]^{-1} \mathbb{I}\left(\mathcal{Q}_M\right) \leq M^*$ for any $\boldsymbol{\xi}_J$, the second term in (S1) can be

upper bounded as

$$
\begin{aligned}
& \sum_{k=1}^{K} \mathrm{E}\left[C_0 \wedge \left|\left(f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right) \log\left(1 + \frac{f_k^*(\boldsymbol{\xi}_J) - \widehat{f}_k(\boldsymbol{\xi}_J)}{\widehat{f}_k(\boldsymbol{\xi}_J)}\right)\right| \mathbb{I}\left(\mathcal{Q}_M\right)\right] \\
\leq & \sum_{k=1}^{K} \mathrm{E}\left[\left|f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right| \log\left(1 + \frac{\left|f_k^*(\boldsymbol{\xi}_J) - \widehat{f}_k(\boldsymbol{\xi}_J)\right| \wedge C_0'}{\widehat{f}_k(\boldsymbol{\xi}_J)}\right) \mathbb{I}\left(\mathcal{Q}_M\right)\right] \\
\leq & \sum_{k=1}^{K} \mathrm{E}\left[\left|f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right| \frac{\left|f_k^*(\boldsymbol{\xi}_J) - \widehat{f}_{(k)}(\boldsymbol{\xi}_J)\right| \wedge C_0'}{\widehat{f}_{(k)}(\boldsymbol{\xi}_J)} \mathbb{I}\left(\mathcal{Q}_M\right)\right] \\
\lesssim & \sum_{k=1}^{K} \mathrm{E}\left(\left|f_k^{\circledast}(\boldsymbol{\xi}) - f_k^*(\boldsymbol{\xi}_J)\right| M^*\right) \\
\lesssim & \ \zeta(J),
\end{aligned}
$$

where the second last inequality comes from Lemmas 7.5 and 7.6 in Wang and Cao (2024).

For any $\delta > 0$, there exists an $M = m > 0$ such that

$$
\mathbb{P}(\mathcal{B}_m) = \mathbb{P}(|\xi_j| \leq m \text{ for all } j \in \mathcal{P}) \geq 1 - \delta, \quad \text{for all } j \in \mathcal{P}.
$$

Therefore, for any $J \geq J_0$, where J_0 is defined in Assumption 3, when δ is relatively small, there exists a corresponding m, such that

$$
\begin{aligned}
& \mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^{\intercal}\left(C_0 \wedge \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)}\right)\right)\right] \\
\lesssim & \ \zeta(J) + \mathrm{E}\left\{\boldsymbol{f}^{*\intercal}(\boldsymbol{\xi}_J)\left[\left(C_0\epsilon^{-1}\right) \wedge \log\left(\frac{\boldsymbol{f}^*(\boldsymbol{\xi}_J)}{\widehat{\boldsymbol{f}}^*(\boldsymbol{\xi}_J)}\right)\right] \mathbb{I}\left(\mathcal{Q}_m\right)\right\}.
\end{aligned}
$$

By Lemma S3, we obtain that

$$\begin{aligned}
&\mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^{\intercal}\left(C_0 \wedge \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^{*}(\boldsymbol{\xi}_J)}\right)\right)\right] \\
\lesssim\ & n^{-\theta}\log^3 n + \zeta(J) + \widetilde{\delta}_n(J) + R^{-1}\sum_{r=1}^{R}\mathrm{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right) \\
& + n^{-1}\left[s\log\left(2^{2L+6}n(L+1)K^3J^2s^L\right) + K\log n\right].
\end{aligned}$$

When Assumption S1 holds, the inequality can be simplified as

$$\begin{aligned}
&\mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^{\intercal}\left(C_0 \wedge \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^{*}(\boldsymbol{\xi}_J)}\right)\right)\right] \\
\lesssim\ & n^{-\theta}\log^3 n + \widetilde{\delta}_n(J) + R^{-1}\sum_{r=1}^{R}\mathrm{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right),
\end{aligned}$$

where the $\widetilde{\delta}_n(J)$ can be sufficiently small when the loss function (3.4) is minimized well. Thus, the proof is completed.

S1.5 Proof of Theorem 2

We first identify an the upper bound of $\mathbb{P}\left(f_k^*\left(\widehat{\boldsymbol{\xi}}_J\right) \le x\right)$. By the definition of f_k^*, we have

$$\left|f_k^*(\boldsymbol{\xi}_{iJ}) - f_k^*(\widehat{\boldsymbol{\xi}}_{iJ})\right| \lesssim \|\boldsymbol{\xi}_{iJ} - \widehat{\boldsymbol{\xi}}_{iJ}\|_{\max}^{\widetilde{\beta}_0^{(k)}\wedge 1},$$

and thus

$$\mathrm{E}\left|f_k^*(\boldsymbol{\xi}_{iJ}) - f_k^*(\widehat{\boldsymbol{\xi}}_{iJ})\right| \lesssim N_i^{-\tau\widetilde{\beta}},\ \text{for any } k = 1,\ldots,K.$$

Therefore, it suffices to show that

$$\mathrm{E}\left|f_k^*(\boldsymbol{\xi}_J) - f_k^*(\widehat{\boldsymbol{\xi}}_J)\right| \lesssim \underline{N}^{-\tau\widetilde{\beta}}, \forall k = 1,\ldots,K.$$

According to Lemma 7.7 in Wang and Cao (2024), we have

$$\left|E\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right)\right| \lesssim \zeta(J)+\underline{N}^{-\tau\widetilde{\beta}}, \text{ for any } k=1,\ldots,K.$$

According to Assumptions 2 and 3, we have for any $k=1,\ldots,K$,

$$\begin{aligned}
& \mathbb{P}\left(f_k^*(\widehat{\boldsymbol{\xi}}_J)\le x\right)\\
= \;& \mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J)\le x+f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right)\\
\le \;& \mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J)\le x+\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|\right)\\
\le \;& \mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J)\le x+\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|,\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|\le x\right)\\
& +\mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J)\le x+\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|,\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|> x\right)\\
\le \;& \mathbb{P}\left(f_k^*(\boldsymbol{\xi}_J)\le 2x\right)+\mathbb{P}\left(\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|> x\right)\\
\lesssim \;& x^{\alpha_k}+\zeta(J)+x\,\mathrm{E}\left|f_k^*(\boldsymbol{\xi}_J)-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right|\\
\lesssim \;& x^{\alpha_k}+\zeta(J)+x\zeta(J)+x\underline{N}^{-\tau\widetilde{\beta}}\\
\lesssim \;& x^{\alpha_k}+\zeta(J)+x\underline{N}^{-\tau\widetilde{\beta}}.
\end{aligned}$$

Next, we adapt the proof of Theorem 1 to bound the decomposed risk:

$$\begin{aligned}
& \mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^{\intercal}\left\{C_0\mathbf{1}_K\wedge\log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^*(\widehat{\boldsymbol{\xi}}_J)}\right)\right\}\right] \qquad \text{(S2)}\\
\le \;& \sum_{k=1}^K \mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^*(\boldsymbol{\xi}_J)\right)^2}{f_k^*(\widehat{\boldsymbol{\xi}}_J)}\right]\\
& +\sum_{k=1}^K \mathrm{E}\left[C_0\wedge\left|\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^*(\widehat{\boldsymbol{\xi}}_J)\right)\log\left(1+\frac{f_k^*(\widehat{\boldsymbol{\xi}}_J)-\widehat{f}^*(\widehat{\boldsymbol{\xi}}_J)}{\widehat{f}^*(\widehat{\boldsymbol{\xi}}_J)}\right)\right|\mathbb{I}\left(\mathcal{Q}_M\right)\right]\\
& +\mathrm{E}\left\{C_0\wedge\left[\boldsymbol{f}^{*\intercal}(\widehat{\boldsymbol{\xi}}_J)\log\left(\frac{\boldsymbol{f}^*(\widehat{\boldsymbol{\xi}}_J)}{\widehat{\boldsymbol{f}}^*(\widehat{\boldsymbol{\xi}}_J)}\right)\mathbb{I}\left(\mathcal{Q}_M\right)\right]\right\}+2C_0K\mathbb{P}\left(\mathcal{Q}_M^c\right).
\end{aligned}$$

By the upper bound of $\mathbb{P}\left(f_k^*(\widehat{\boldsymbol{\xi}}_J)\right)$ in (S2), when J is relatively large, we can bound

the first term in (S2). When J is relatively large, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)^2}{f_k^{*}(\widehat{\boldsymbol{\xi}}_J)}\right]\\
=&\sum_{k=1}^{K}\mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)^2}{f_k^{*}(\widehat{\boldsymbol{\xi}}_J)}\mathbb{I}\left(f_k^{*}(\widehat{\boldsymbol{\xi}}_J)>\epsilon\right)\right]\\
&+\sum_{k=1}^{K}\mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)^2}{f_k^{*}(\widehat{\boldsymbol{\xi}}_J)}\mathbb{I}\left(f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\le\epsilon\right)\right]\\
<&\sum_{k=1}^{K}\mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)^2}{\epsilon}\right]+\sum_{k=1}^{K}\mathrm{E}\left[\frac{\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)^2}{f_k^{*}(\widehat{\boldsymbol{\xi}}_J)}\right]\mathbb{P}\left(f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\le\epsilon\right)\\
\lesssim&\ \zeta(J)+\underline{N}^{-\tau\widetilde{\beta}}.
\end{aligned}$$

Similarly as indicated in proof of Theorem 1, by the fact that given M, within the concentrated set $\mathcal{Q}_M$, there exists an M^{**} such that

$$\left[\widehat{f}^{(k)}(\boldsymbol{\xi}_J)\right]^{-1}\mathbb{I}\left(\mathcal{Q}_M\right)\le M^{**}$$

for any $\boldsymbol{\xi}_J$, the second term in (S2) can be upper bounded:

$$\begin{aligned}
&\sum_{k=1}^{K}\mathrm{E}\left[C_0\wedge\left|\left(f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right)\log\left(1+\frac{f_k^{*}(\widehat{\boldsymbol{\xi}}_J)-\widehat{f}_k(\widehat{\boldsymbol{\xi}}_J)}{\widehat{f}_k(\widehat{\boldsymbol{\xi}}_J)}\right)\right|\mathbb{I}\left(\mathcal{Q}_M\right)\right]\\
\le&\sum_{k=1}^{K}\mathrm{E}\left[\left|f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right|\log\left(1+\frac{\left|f_k^{*}(\widehat{\boldsymbol{\xi}}_J)-\widehat{f}_k(\widehat{\boldsymbol{\xi}}_J)\right|\wedge C_0'}{\widehat{f}_k(\widehat{\boldsymbol{\xi}}_J)}\right)\mathbb{I}\left(\mathcal{Q}_M\right)\right]\\
\le&\sum_{k=1}^{K}\mathrm{E}\left[\left|f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right|\frac{\left|f_k^{*}(\widehat{\boldsymbol{\xi}}_J)-\widehat{f}_{(k)}(\widehat{\boldsymbol{\xi}}_J)\right|\wedge C_0'}{\widehat{f}_{(k)}(\widehat{\boldsymbol{\xi}}_J)}\mathbb{I}\left(\mathcal{Q}_M\right)\right]\\
\lesssim&\sum_{k=1}^{K}\mathrm{E}\left(\left|f_k^{\circledast}(\boldsymbol{\xi})-f_k^{*}(\widehat{\boldsymbol{\xi}}_J)\right|M^{**}\right)\\
\lesssim&\ \zeta(J)+\underline{N}^{-\tau\widetilde{\beta}}.
\end{aligned}$$

Consequently, under Assumption S2, we obtain that

$$\begin{aligned} & \mathrm{E}\left[\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})^{\intercal}\left(C_0 \wedge \log\left(\frac{\boldsymbol{f}^{\circledast}(\boldsymbol{\xi})}{\widehat{\boldsymbol{f}}^{*}(\widehat{\boldsymbol{\xi}}_J)}\right)\right)\right] \\ \lesssim\ & n^{-\theta}\log^3 n + \underline{N}^{-\tau\widetilde{\beta}} + \widetilde{\delta}_n(J) + R^{-1}\sum_{r=1}^{R}\mathrm{tr}\left(\boldsymbol{A}^{*(r)} - \boldsymbol{I}_K\right), \end{aligned}$$

where the phase transition point is at $N^* \asymp \left(n^{\theta}/\log^3 n\right)^{1/\theta'}$, and $\theta' = \tau\widetilde{\beta}$. When the classifier is well trained by minimizing the empirical loss, and $\widetilde{\delta}_J$ is sufficiently small, the proof can be completed.

S2 Additional numerical analysis results

S2.1 Imbalanced annotations

In practice, it is common for each annotator to label only a subset of the data rather than the entire dataset. For instance, in our real data application on lung disease, each subject receives a limited number of diagnoses from a pool of doctors. Motivated by this setting, we extend our model to address the imbalanced annotator problem.

Suppose there are R annotators, where each annotator r labels a subset of data denoted by $I_r \subset [n]$, with $|I_r| = n_r$ for $r \in [R]$. We adapt the same procedure used in the original minimization problem (3.4) to accommodate this imbalanced setting. The resulting optimization problem is given by

$$(\widehat{\boldsymbol{f}}^{*}, \widehat{\boldsymbol{A}}^{(r)}) = \underset{\boldsymbol{f}\in\mathcal{D},\boldsymbol{A}^{(r)}\in\mathcal{A}}{\arg\min} -\frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_r}\sum_{i\in I_r}\sum_{k=1}^{K}\mathbb{I}(\widetilde{Y}_i^{(r)} = k)\log\left([\boldsymbol{A}^{(r)}\boldsymbol{f}(\boldsymbol{\xi}_J^{(i)})]_k\right) + \lambda\,\mathrm{tr}\left(\sum_{r=1}^{R}\boldsymbol{A}^{(r)}\right).$$

We adopt the same data-generating strategies used in Models 1 – 8, and randomly assign the index sets $\{I_r\}_{r\in[R]}$ such that $\bigcup_{r\in[R]} I_r = [n]$, and $n_r = \lceil 0.5n \rceil$ for all $r \in [R]$.

This design ensures that each annotator evaluates no more than half of the data. Below, we present Figures S1 to S4, which display the classification accuracy and F-norm error of the confusion matrices for all methods, evaluated under varying sampling frequencies across Models 1 to 8, based on the imbalanced design described above. Consistent with the results from the balanced design, our proposed method, `afDNN`, consistently outperforms all baseline methods. Notably, across all levels of contamination, `afDNN` demonstrates performance comparable to that of the oracle classifier, highlighting its effectiveness in accurately estimating the CM and correctly classifying new subjects. Specifically, the proposed `afDNN` maintains strong performance under low contamination levels ($\epsilon = 0.1, 0.3$), as measured by both classification accuracy and CM estimation. When the noise level increases to $\epsilon = 0.5$, our method still outperforms the alternatives, though it remains below the oracle model. This decline is mainly due to the increasing bias in CM estimation and the sparsity of annotations per annotator.

S2.2 Additional Results for Models 5 to 12

Below, Figures S5 to S11 depict classification accuracy and F-norm error in Models 5 to 12.

S3 ADNI dataset

The PET images underwent spatial normalization and comprehensive post-processing to ensure consistency across subjects. The AD group, which had between three and six clinical visits, contributed scans from their third visit, while the EMCI cohort, with

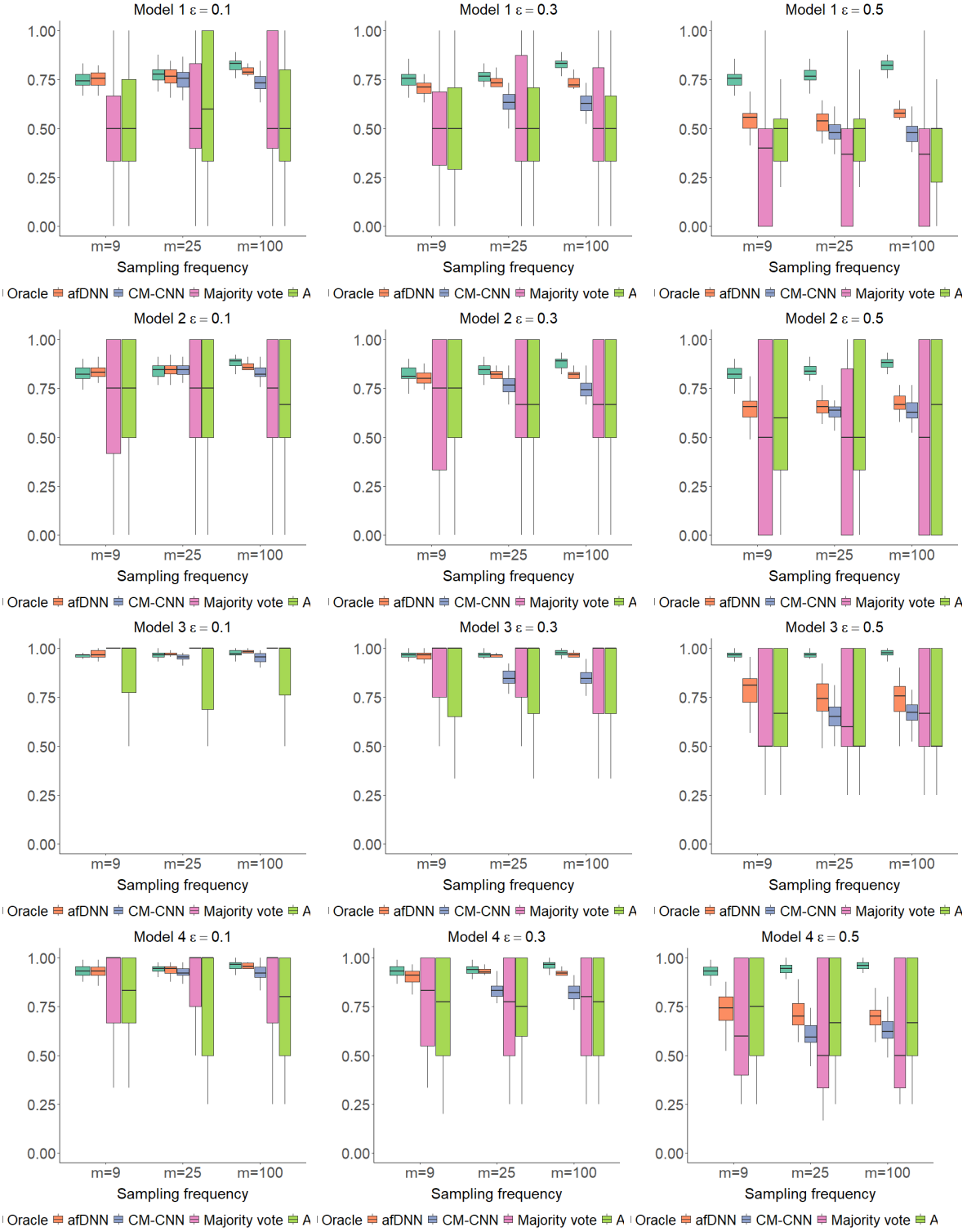

Figure S1: Classification accuracy for all methods across different sampling frequencies for Models 1 – 4 with imbalanced design. From top to bottom, the rows represent the results for Model 1 – 4, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

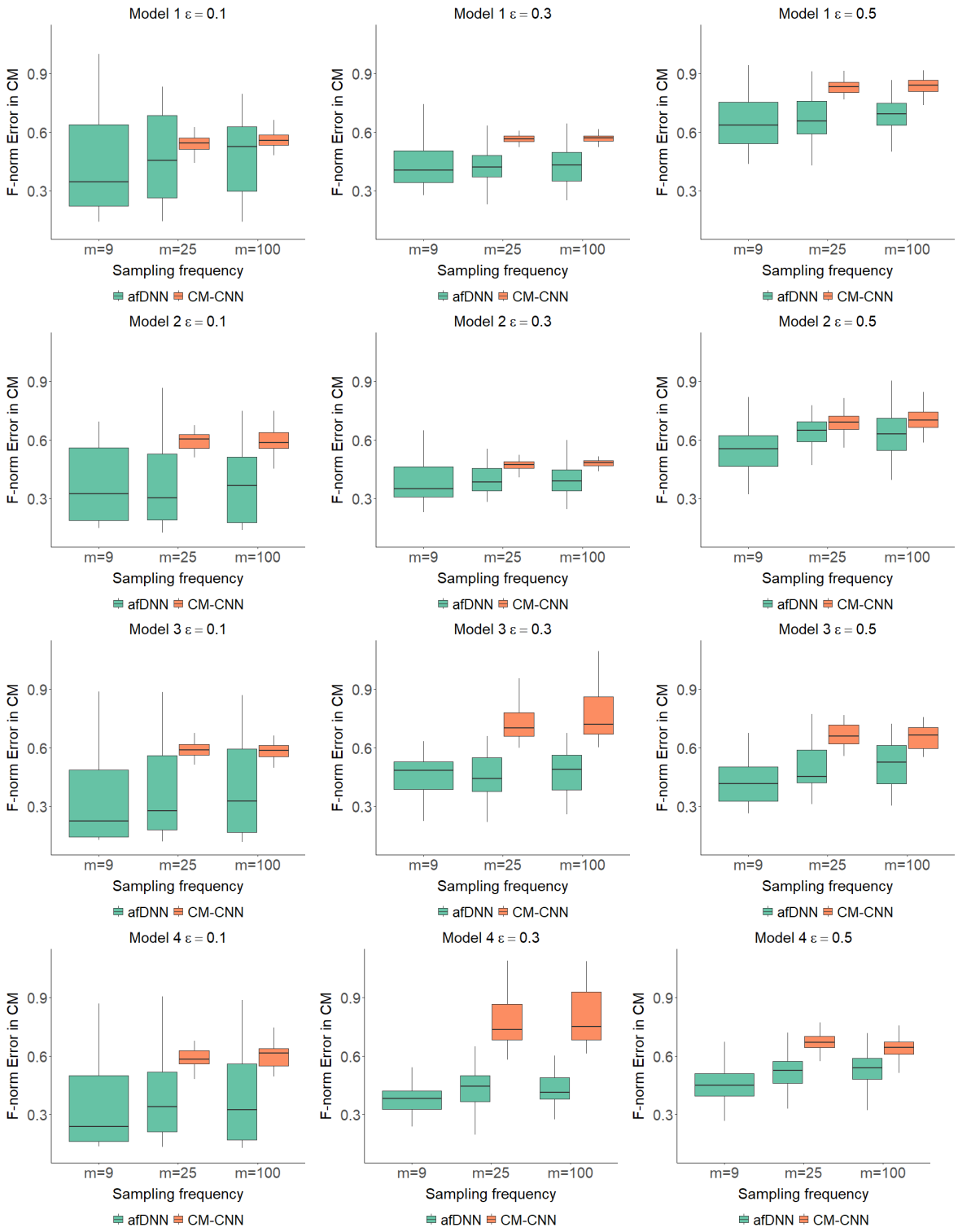

Figure S2: F-norm error in confusion matrices for all methods across different sampling frequencies for Models 1 – 4 with imbalanced design. From top to bottom, the rows represent the results for Model 1 – 4, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

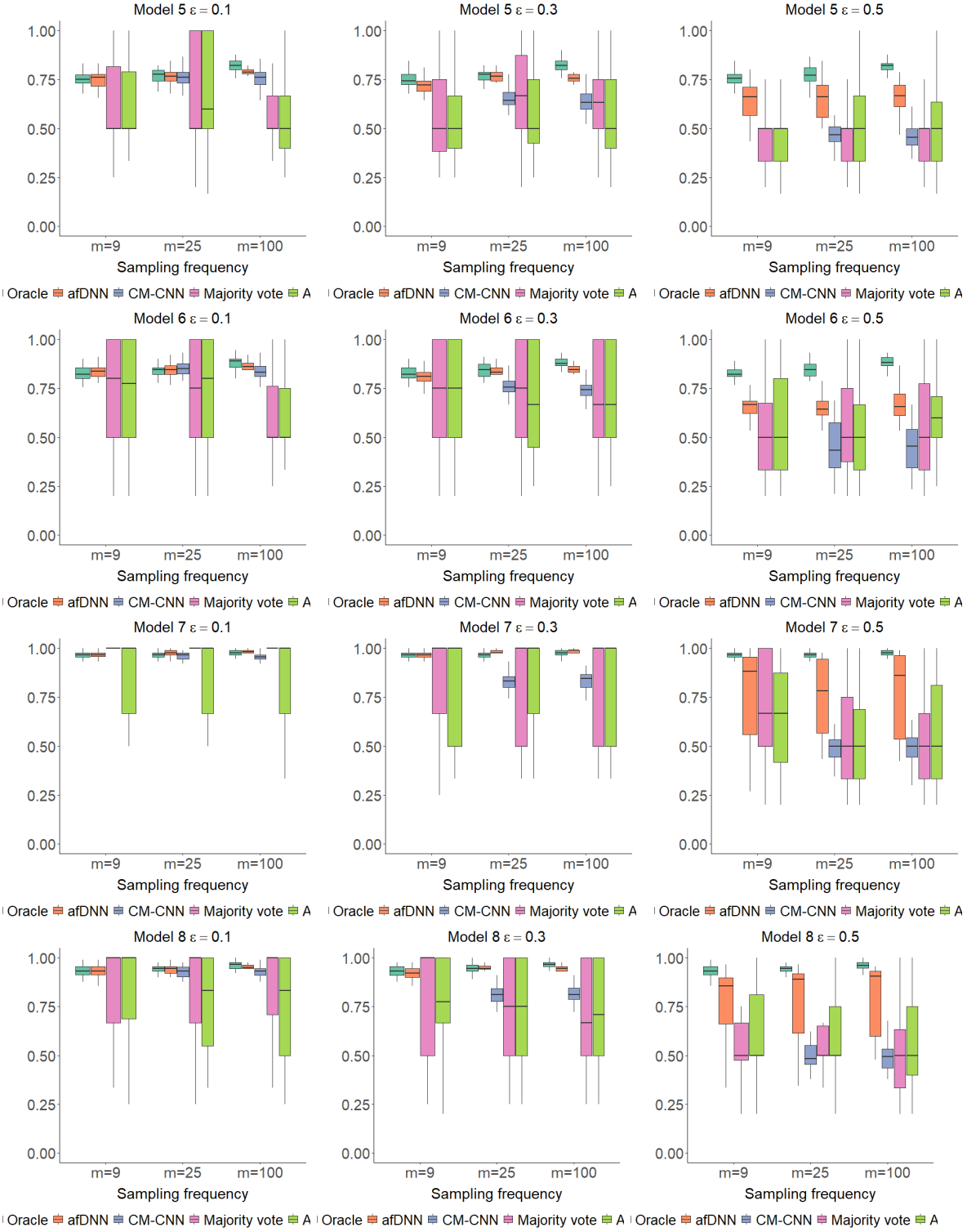

Figure S3: Classification accuracy for all methods across different sampling frequencies for Models 5 – 8 with imbalanced design. From top to bottom, the rows represent the results for Model 5 – 8, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

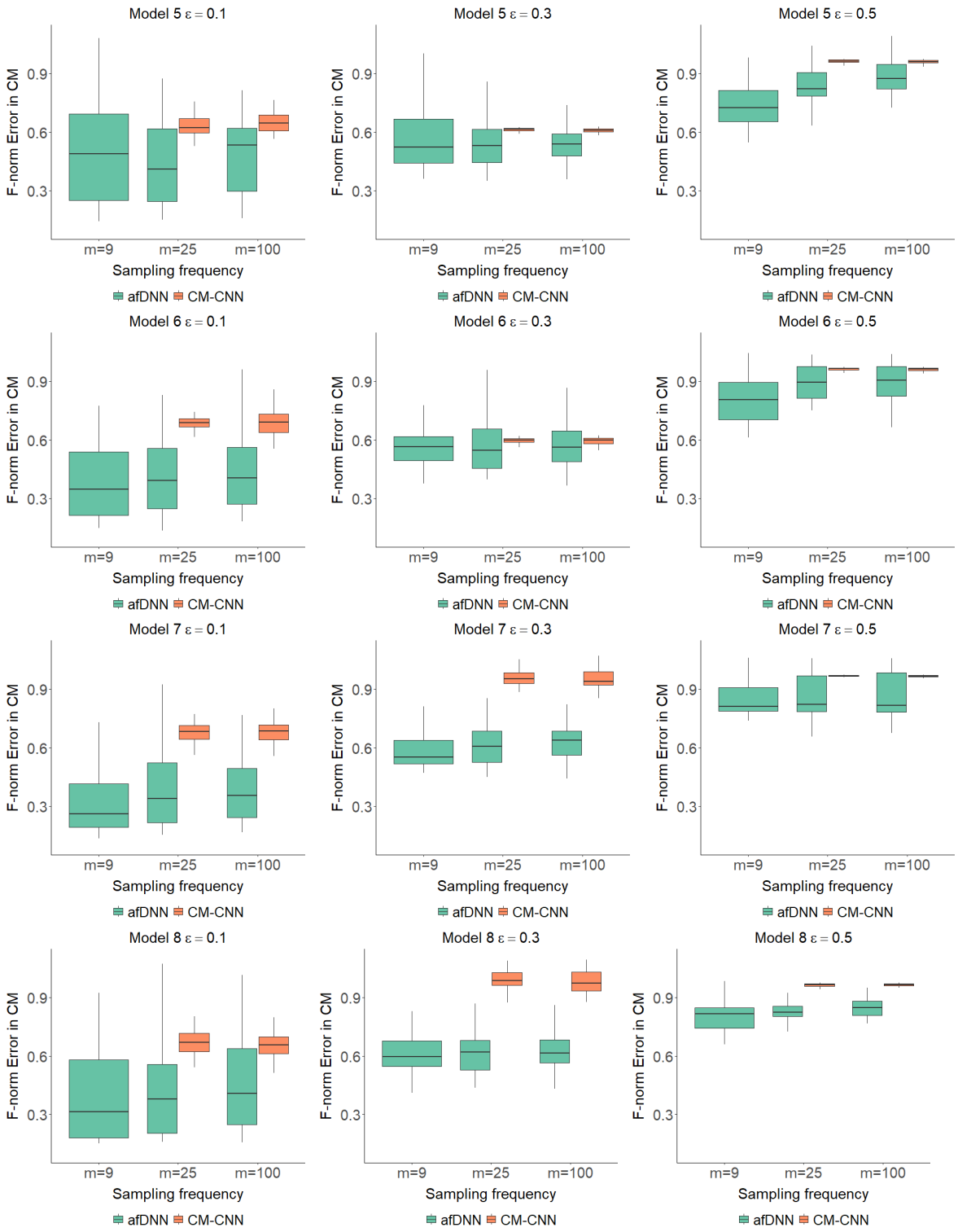

Figure S4: F-norm error in confusion matrices for all methods across different sampling frequencies for Models $5-8$ with imbalanced design. From top to bottom, the rows represent the results for Model $5-8$, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

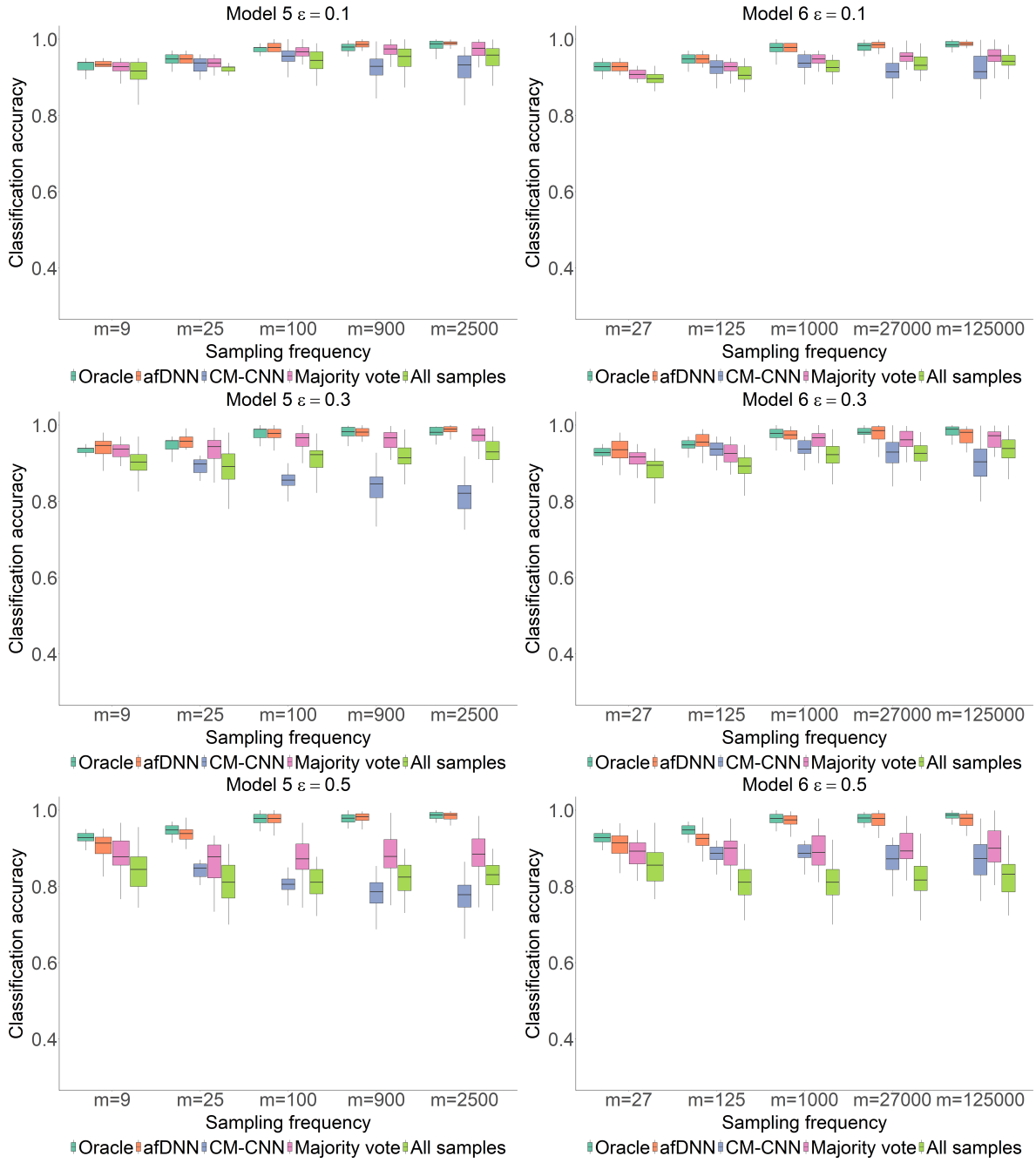

Figure S5: Classification accuracy for all methods across different sampling frequencies for Models 5 – 6. From left to right, the rows represent the results for Model 5 – 6, respectively. From top to bottom, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

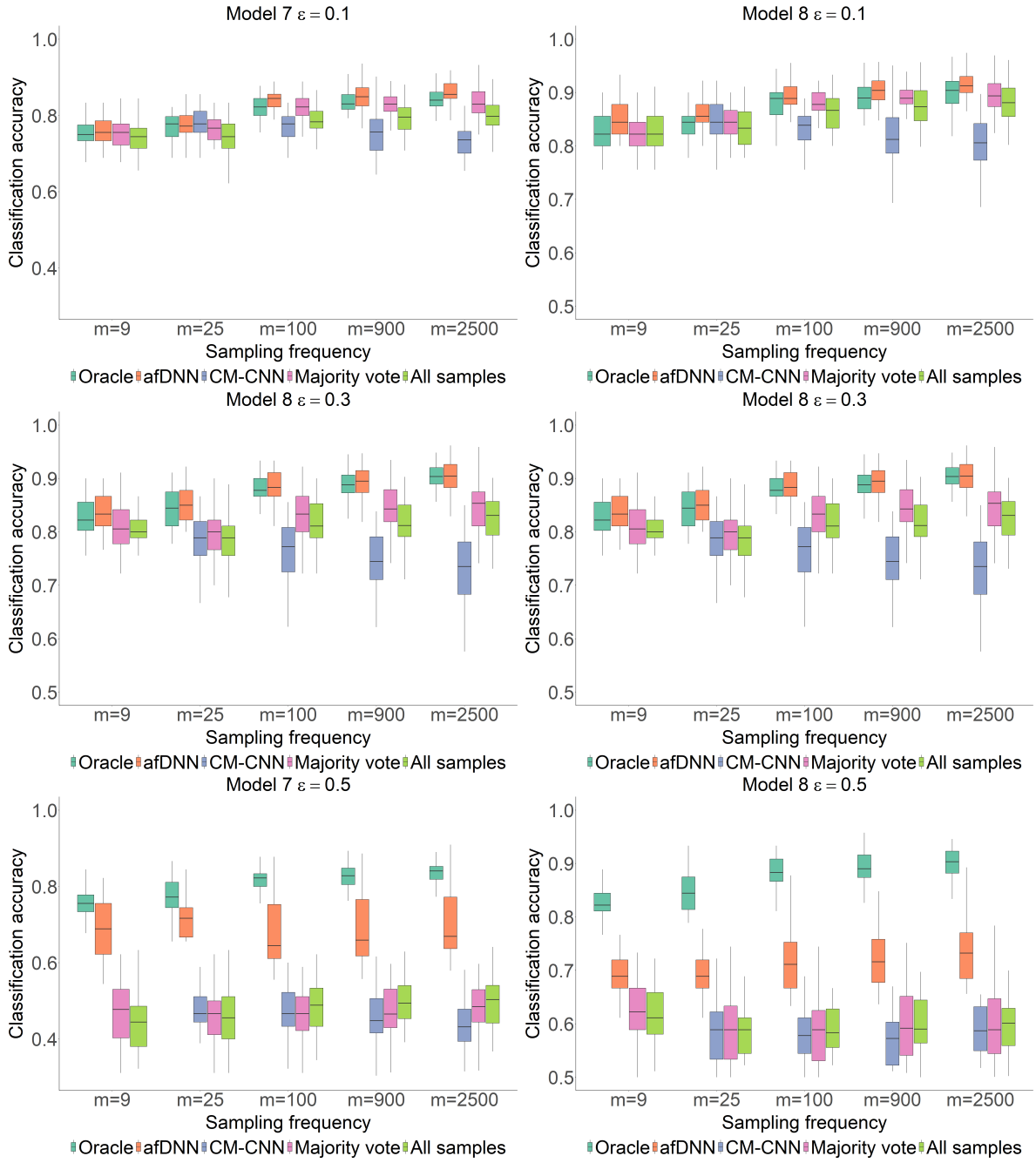

Figure S6: Classification accuracy for all methods across different sampling frequencies for Models 7 – 8. From left to right, the rows represent the results for Model 7 – 8, respectively. From top to bottom, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

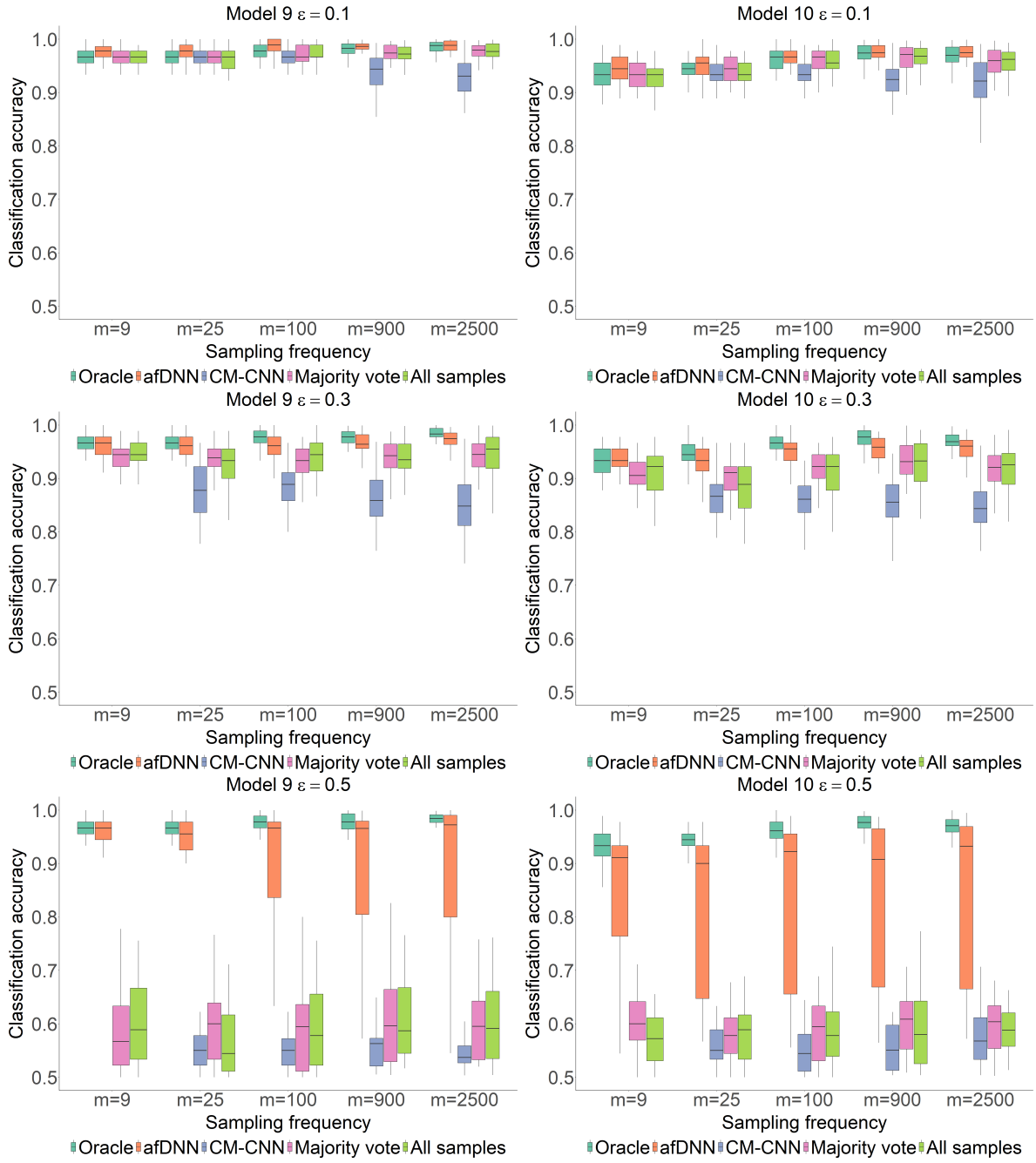

Figure S7: Classification accuracy for all methods across different sampling frequencies for Models 9 – 10. From left to right, the rows represent the results for Model 9 – 10, respectively. From top to bottom, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

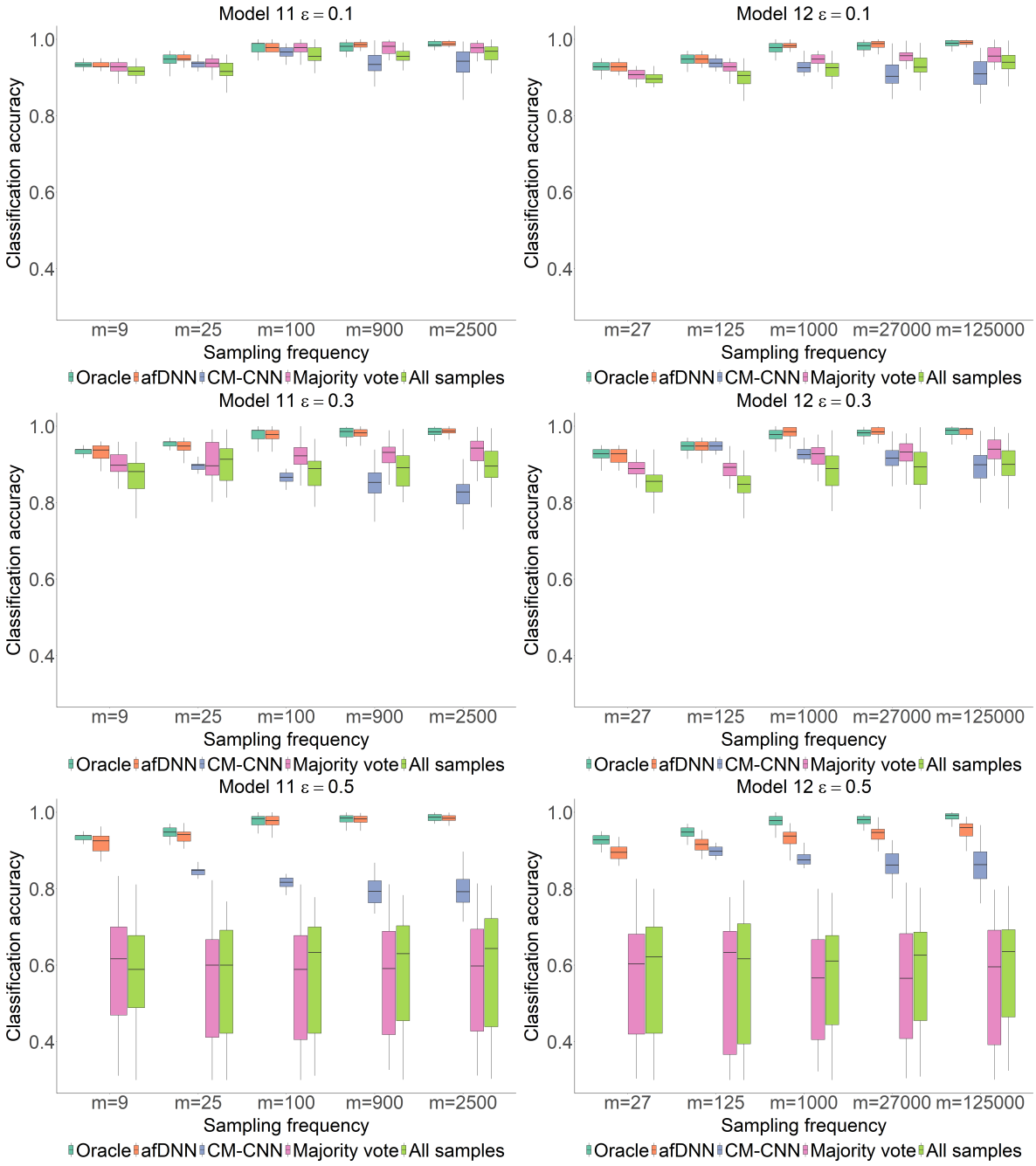

Figure S8: Classification accuracy for all methods across different sampling frequencies for Models 11 – 12. From left to right, the rows represent the results for Model 11 – 12, respectively. From top to bottom, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

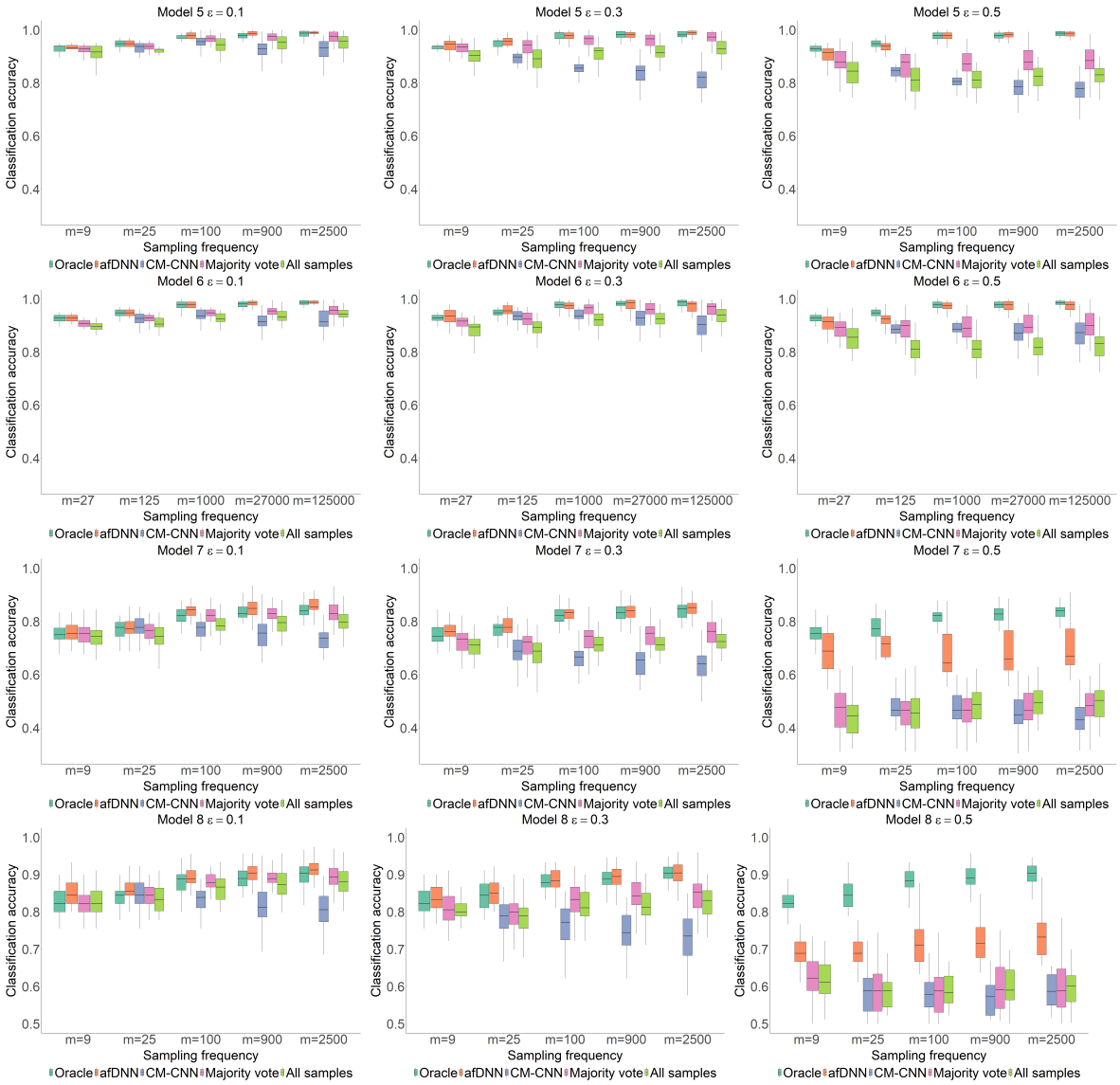

Figure S9: Classification accuracy for all methods across different sampling frequencies for Models 5 – 8. From top to bottom, the rows represent the results for Model 5 – 8, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

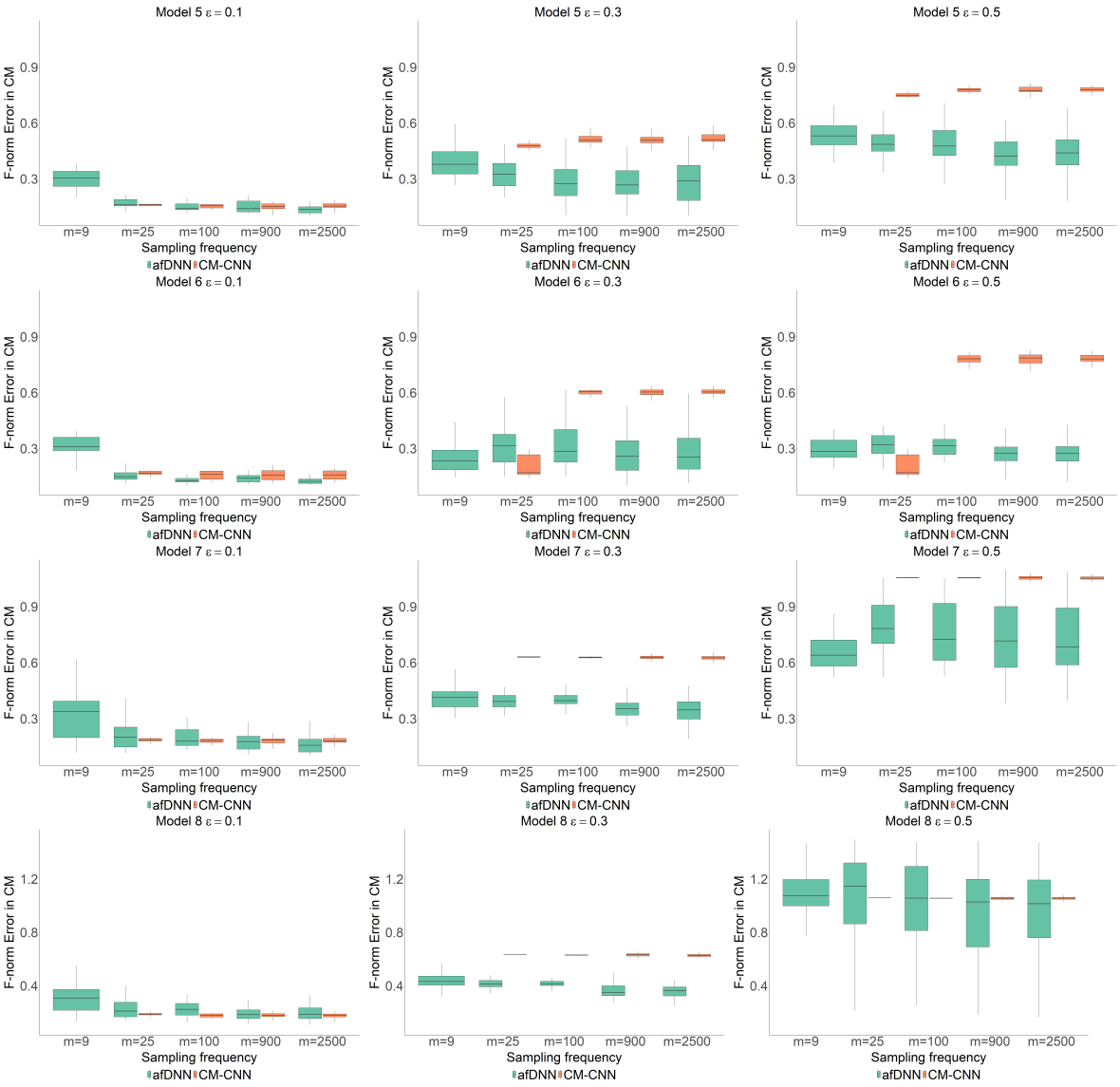

Figure S10: F-norm error in CM for all methods across different sampling frequencies for Models 5 – 8. From top to bottom, the rows represent the results for Model 5 – 8, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

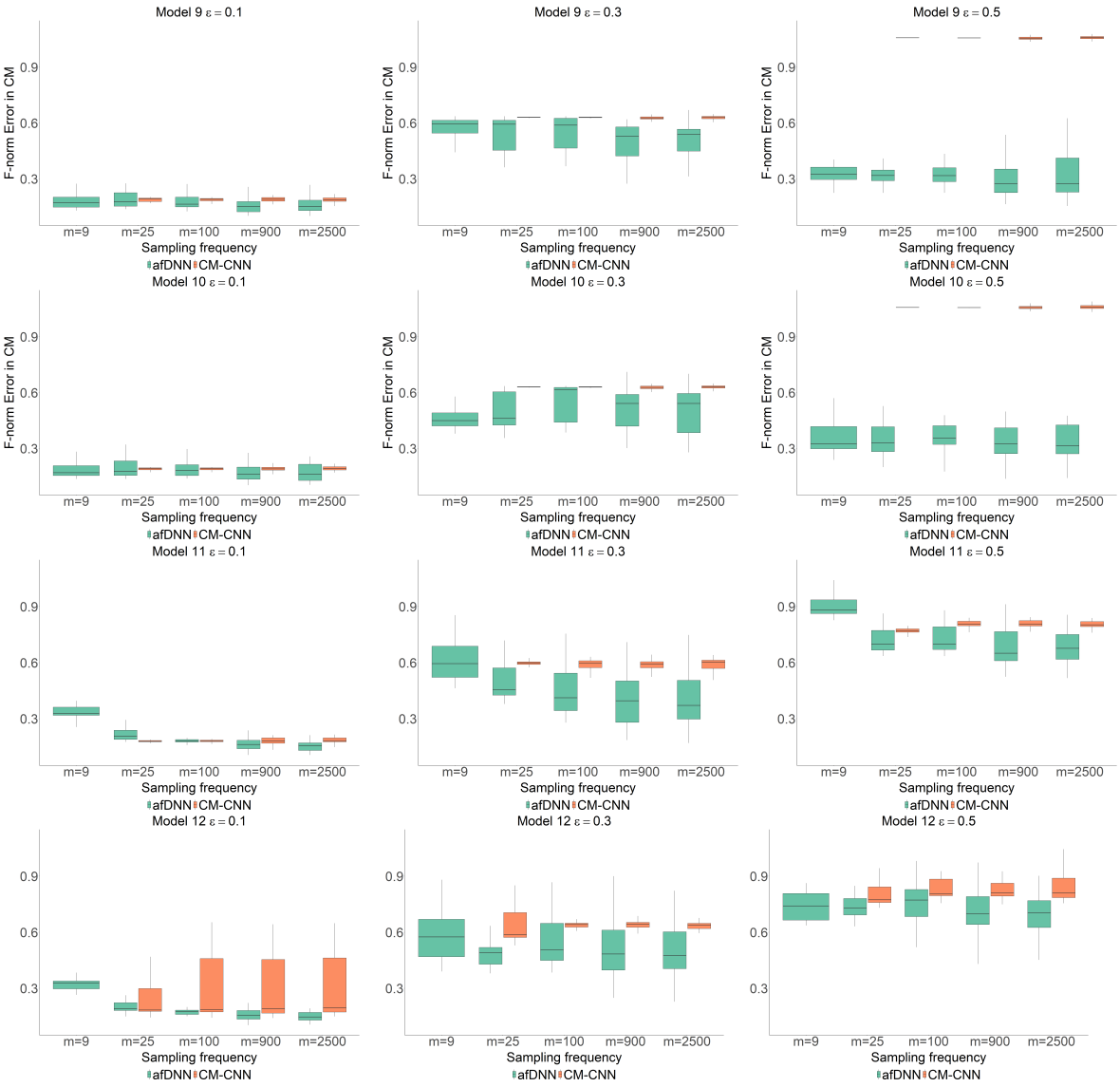

Figure S11: F-norm error in CM for all methods across different sampling frequencies for Models 9 – 12. From top to bottom, the rows represent the results for Model 9 – 12, respectively. From left to right, the columns correspond to $\epsilon = 0.1, 0.3$, and 0.5.

Table S1: Averaged classification accuracy with standard errors in brackets for ADNI 2D brain images with different values of ϵ for 10th, 30th and 50th slices.

Methods	$\epsilon = 0.1$			$\epsilon = 0.3$			$\epsilon = 0.5$		
	10-th	30-th	50-th	10-th	30-th	50-th	10-th	30-th	50-th
Oracle	0.50 (0.03)	0.45 (0.03)	0.48 (0.04)	0.50 (0.03)	0.45 (0.03)	0.48 (0.04)	**0.50 (0.03)**	**0.45 (0.03)**	**0.48 (0.04)**
`afDNN`	**0.52 (0.02)**	**0.49 (0.01)**	**0.52 (0.01)**	**0.51 (0.02)**	**0.49 (0.02)**	**0.51 (0.01)**	0.45 (0.03)	0.48 (0.02)	0.46 (0.02)
All sample	0.49 (0.02)	0.47 (0.03)	0.49 (0.01)	0.50 (0.01)	0.48 (0.02)	0.40 (0.01)	0.44 (0.03)	0.43 (0.03)	0.37 (0.03)
Majority vote	0.48 (0.03)	0.45 (0.01)	0.42 (0.02)	0.40 (0.02)	0.47 (0.04)	0.37 (0.03)	0.32 (0.02)	0.36 (0.00)	0.36 (0.00)
CM-CNN	0.50 (0.06)	0.46 (0.06)	0.45 (0.05)	0.43 (0.05)	0.44 (0.06)	0.44 (0.06)	0.38 (0.06)	0.42 (0.06)	0.41 (0.06)

only two visits, provided scans from the second session. Figure S12 illustrates the averaged images of the 20th, 40th and 60th slices for the AD, EMCI, and CN groups. Table S1 presents averaged classification accuracy with standard errors in brackets with different values of ϵ for 10th, 30th and 50th slices. Similar conclusions can be drawn for these three slices, as presented in the main paper: the proposed `afDNN` consistently demonstrates strong performance under low contamination levels ($\epsilon = 0.1, 0.3$), as reflected in both classification accuracy and confusion matrix estimation. As the noise level increases to $\epsilon = 0.5$, our method continues to outperform competing approaches, although it remains slightly below the oracle model. Table S2 presents the classification accuracy comparison results. All methods achieve slightly better classification accuracy in the 3D setting than the 2D slice-based analysis, suggesting that the additional spatial information is beneficial. Our proposed method continues to perform strongly and remains competitive with the oracle model that is trained using the true labels.

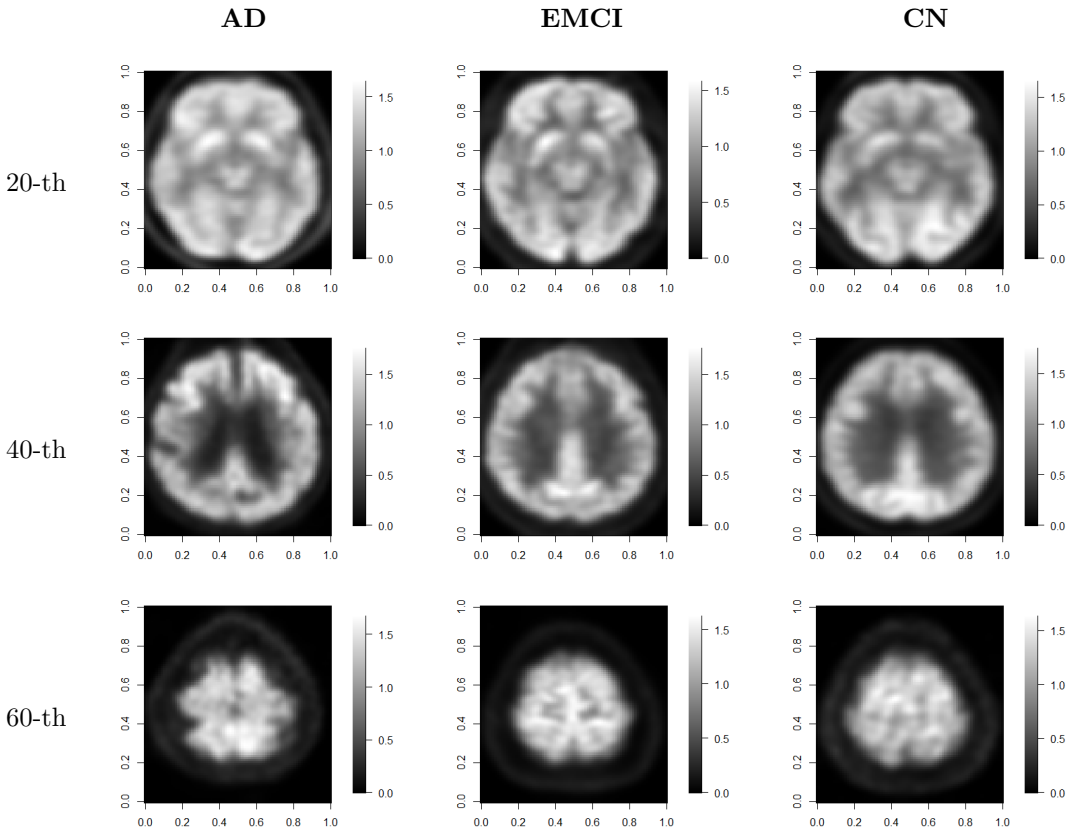

Figure S12: Sample images of the 20th, 40th and 60th slices for AD group (left column), EMCI group (middle column), and CN group (right column).

Table S2: Averaged classification accuracy with standard errors in brackets for ADNI 3D brain images with different values of ϵ.

Methods	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$
Oracle	**0.66** (0.07)	**0.66** (0.07)	**0.66** (0.07)
`afDNN`	**0.66** (0.06)	0.63 (0.06)	0.59 (0.07)
All sample	0.58 (0.05)	0.54 (0.04)	0.50 (0.04)
Majority vote	0.57 (0.05)	0.52 (0.05)	0.51 (0.04)
CM-CNN	0.60 (0.08)	0.58 (0.06)	0.57 (0.06)

Bibliography

Bos, T. and J. Schmidt-Hieber (2022). Convergence rates of deep relu networks for multiclass classification. *Electronic Journal of Statistics 16*, 2724–2773.

Tanno, R., A. Saeedi, S. Sankaranarayanan, C. D. Alexander, and N. Silberman (2019). Learning from noisy labels by regularized estimation of annotator confusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11244–11253.

Wang, S. and G. Cao (2024). Multiclass classification for multidimensional functional data through deep neural networks. *Electronic Journal of Statistics 18*, 1248–1292.