	supervised				PU learning				noisy PU learning			
digit	BSVM	BAG	PUE	p	BSVM	BAG	PUE	p	BSVM	BAG	PUE	p
0	99.4–99.5	99.5–99.6	99.4–99.6		95.9–96.6	98.4–99.1	99.0–99.3	•	91.1–93.1	93.8–97.3	98.4–98.8	• • •
1	99.6–99.7	99.7–99.7	99.7–99.7		97.9–98.7	98.9–99.4	99.3–99.6	• • •	97.3–97.9	98.3–99.1	99.4–99.6	• • •
2	95.5–96.1	96.1–96.6	96.5–97.1	• •	86.2–89.1	93.0–94.8	95.9–96.5	• • •	81.6–84.9	88.2–91.1	93.8–95.6	• • •
3	94.9–95.7	96.2–97.0	97.1–97.5	• • •	88.0–90.0	94.0–95.6	95.8–96.7	•	84.1–85.8	88.7–92.2	94.8–95.6	• • •
4	97.5–98.0	98.0–98.4	98.1–98.6		90.4–92.5	94.4–96.5	96.9–97.5	• •	88.7–90.9	92.6–95.5	95.6–96.7	• •
5	93.6–94.5	94.0–94.8	94.9–96.4	• • •	87.1–89.9	90.5–93.9	93.8–94.8	•	82.6–86.4	87.9–91.3	92.9–94.1	• • •
6	98.2–98.7	98.7–99.0	98.7–98.9		93.3–94.0	97.5–98.4	98.1–98.5		87.3–90.7	93.6–96.3	97.4–98.2	• • •
7	97.7–98.1	97.8–98.2	97.9–98.3		94.4–95.4	96.6–97.6	97.4–98.0	•	90.5–91.5	92.4–95.3	96.7–97.5	• •
8	86.1–87.6	92.8–93.8	95.0–95.2	• • •	77.7–80.7	89.9–92.2	92.8–94.1	• •	76.6–80.3	86.2–89.0	92.3–93.7	• • •
9	93.6–94.5	94.7–95.5	95.7–96.1	• •	88.5–89.9	91.8–94.0	94.2–95.2	• •	84.9–86.8	87.7–90.3	92.9–93.8	• • •

Table 1: 95% confidence intervals for the mean area under the ROC curve in percent on MNIST test set for each digit in a one-vs-all setup. AUC is computed using full label information. Models are trained in the following settings: (i) fully supervised (no false positives and no false negatives), (ii) classic PU learning setting (10% false negatives, no false positives) and (iii) a noisy PU learning setting (10% false positives and 10% false negatives). In each setting the results of a paired one-tailed Wilcoxon signed-rank test comparing the AUC of BAG and PUE with alternative hypothesis $AUC^{PUE} > AUC^{BAG}$ is included using the following result coding: • $p < 0.05$, • • $p < 0.01$ and • • • $p < 0.001$.

	supervised				PU learning				noisy PU learning			
digit	BSVM	BAG	PUE	p	BSVM	BAG	PUE	p	BSVM	BAG	PUE	p
0	96.3–97.1	96.7–97.4	96.7–97.4		74.2–78.6	89.8–94.1	94.2–95.6	•	57.5–63.5	70.5–84.2	90.7–93.2	• • •
1	98.1–98.4	98.3–98.6	98.2–98.6		88.1–91.4	94.8–97.0	96.2–97.4	• •	80.2–84.0	91.5–94.9	96.4–97.8	• • •
2	86.6–88.8	88.0–89.8	89.1–90.4		52.9–60.3	74.1–80.8	84.1–86.7	• • •	41.6–49.9	58.5–68.8	77.8–84.0	• • •
3	83.1–86.0	86.2–88.9	88.8–90.5	• •	54.2–61.0	74.2–81.5	81.8–86.1	•	43.0–49.1	56.3–66.6	78.4–82.6	• • •
4	87.7–89.2	88.7–90.6	89.5–91.8	• •	55.6–61.2	70.9–79.4	83.3–85.6	• •	51.0–57.9	66.3–75.6	78.5–83.4	• • •
5	77.1–80.5	78.0–80.6	80.6–82.5		52.9–57.8	61.8–71.2	70.4–73.3	•	39.0–47.2	52.9–61.7	69.0–71.7	• • •
6	92.6–93.9	94.1–95.1	94.4–95.1		65.8–68.2	86.1–90.0	89.4–92.4		50.7–58.4	71.2–81.8	87.1–91.2	• • •
7	91.8–93.0	92.4–93.3	93.3–93.9	• • •	71.8–76.2	85.2–89.7	90.7–92.1	• • •	56.9–60.0	65.5–78.7	88.1–90.1	• •
8	56.3–59.8	74.8–77.4	80.0–81.1	• • •	32.8–39.3	62.1–68.1	72.3–75.3	• •	29.7–36.1	49.4–59.4	70.0–73.7	• • •
9	71.3–76.1	77.0–81.0	81.1–82.9	• •	49.1–51.9	63.6–71.3	73.6–76.8	• • •	41.2–44.8	50.8–59.3	67.2–71.7	• • •

Table 2: 95% confidence intervals for the mean area under the ROC curve in percent on MNIST test set for each digit in a one-vs-all setup. AUC is computed using full label information. Models are trained in the following settings: (i) fully supervised (no false positives and no false negatives), (ii) classic PU learning setting (10% false negatives, no false positives) and (iii) a noisy PU learning setting (10% false positives and 10% false negatives). In each setting the results of a paired one-tailed Wilcoxon signed-rank test comparing the AUC of BAG and PUE with alternative hypothesis $AUC^{PUE} > AUC^{BAG}$ is included using the following result coding: • $p < 0.05$, • • $p < 0.01$ and • • • $p < 0.001$.