A. Experiment Details

Robust Generative Classifier (RoG): Lee et al. (2019) induces a generative classifier on top of the hidden feature spaces of a model pre-trained on noisy data. They assume the hidden features are Gaussian-distributed and estimate its parameters using the minimum covariance determinant estimator. We always use $I_{\text{max}} = 2$, and train the ensemble over 40 epochs over the filtered validation data using Adam with learning rate 0.001. 50% of the validation data was always kept and 2% of the dataset was used for validation except for smaller UCI datasets, where 10% was used. For ResNet-20 we use the convolutional and identity blocks within the final resnet block as hidden layers for RoG and use all intermediate dense layers for all other architectures.

B. Proofs

B.1. Supporting theoretical results

The following bounds $r_k(x)$ uniformly in $x \in \mathcal{X}$.

Lemma 1 (Lemma 2 of Jiang (2019))**.** *The following holds with probability at least* $1 - \delta/2$*. If*

$$2^8 \cdot D \log^2(4/\delta) \cdot \log n \leq k \leq \frac{1}{2} \cdot \omega \cdot p_{X,0} \cdot v_D \cdot r_0^D \cdot n,$$

then $\sup_{x\in\mathcal{X}} r_k(x) \leq \left(\frac{2k}{\omega \cdot v_D \cdot n \cdot p_{X,0}}\right)^{1/D}$*, where* v_D *is the volume of the unit ball in* $\mathbb{R}^D$*.*

The next result bounds the number of distinct k-NN sets over $\mathcal{X}$.

Lemma 2 (Lemma 3 of Jiang (2019))**.** *Let* M *be the number of distinct* k*-NN sets over* $\mathcal{X}$*, that is,* $M := |\{N_k(x) : x \in \mathcal{X}\}|$*. Then* $M \leq D \cdot n^D$*.*

B.2. Minimum k-NN spread

We propose a more general notion of how spread out a set of points is than S_2 which will be used in the theoretical analysis. This will allow us to more precisely characterize how difficult a configuration of incorrectly labeled examples will be to work with in the k-NN context. For example, if such examples are spread out far apart, then there will be many correctly labeled examples nearby for the k-NN approach to identify the incorrectly labeled examples. On the other hand, if the corrupted examples are all close together, then it will be more difficult to identify them without many uncorrupted examples in that region. To this end, we define the minimum k-NN spread:

Definition 4 (minimum k-NN spread)**.**

$$S_k(C) := \min_{x\in C} r_k(x, C),$$

where $r_k(x, C)$ *denotes the distance from* x *to the* k*-th closest neighbor in* C*.*

Note that this definition is consistent with the earlier definition of S_2.

B.3. Proof of Theorem 1

Proof of Theorem 1. Let $\tau, \gamma, \epsilon > 0$ be quantities that will be determined later. Suppose that for some $x \in \mathcal{X}^\Delta$, we have $r_k(x) \leq \tau$ and $S_{\lfloor(\frac{1}{2}-\gamma)\cdot k\rfloor}(C) \geq \tau$. Then, at least $\frac{1}{2} + \gamma$ fraction of the points within x's k-nearest neighbors are not in the corrupted set C. Let $A_x := N_k(x)\backslash C$, that is, the k-nearest neighbors of x that are not in C. Then it is clear that A_x is a k_0-nearest neighbor set of x relative to $X\backslash C$ for some $k_0 \geq \lceil(\frac{1}{2} + \gamma) \cdot k\rceil$. We have that $A_x \subseteq \mathcal{X}^\Delta \oplus \tau$ where $A \oplus r := \{x \in \mathcal{X} : \inf_{a\in A} |x - a| \leq r\}$. Let us consider without loss of generality that $\eta(x) \geq \frac{1}{2} + \Delta$ (call this set $\mathcal{X}^{\Delta,+}$). The case $\mathcal{X}^{\Delta,-} := \{x \in \mathcal{X}^\Delta : \eta(x) \leq \frac{1}{2} - \Delta\}$ follows by symmetry. Thus, we have $\eta(x') \geq \frac{1}{2} + \Delta - C_\alpha\tau^\alpha$ for all $x' \in A_x$. By Hoeffding's inequality, we have

$$\mathbb{P}\left(\frac{1}{|A_x|}\sum_{x'\in A_x} y(x') < \frac{1}{2} + \Delta - C_\alpha\tau^\alpha - \epsilon\right) \leq \exp(-2\epsilon^2 \cdot k_0),$$

where $y(x)$ is the label corresponding to sample x. By Lemma 2, we have that there are at most $D \cdot n^D$ such k_0-nearest neighbor sets across all k_0 in $X \backslash C$. That is, this is also a bound on the number of distinct A_x for $x \in \mathcal{X}$. Therefore, if we set

$$\epsilon = \sqrt{\frac{D \log n + \log(4D/\delta)}{(1+2\gamma) \cdot k}},$$

then by union bound, we have that

$$\mathbb{P}\left(\inf_{x \in \mathcal{X}^{\Delta,+}} \frac{1}{|A_x|} \sum_{x' \in A_x} y(x') < \frac{1}{2} + \Delta - C_\alpha \tau^\alpha - \epsilon \right) \leq \frac{\delta}{4}.$$

and thus, with probability at least $1-\delta/4$, we have that $\frac{1}{|A_x|}\sum_{x' \in A_x} y(x') \geq \frac{1}{2} + \Delta - C_\alpha \tau^\alpha - \epsilon$ *uniformly* over $x \in \mathcal{X}^{\Delta,+}$. Similarly, with probability at least $1-\delta/4$ we have that $\frac{1}{|A_x|}\sum_{x' \in A_x} y(x') \leq \frac{1}{2} - \Delta + C_\alpha \tau^\alpha + \epsilon$ *uniformly* over $x \in \mathcal{X}^{\Delta,-}$.

Hence, in order for k-nearest neighbor prediction to predict the Bayes-optimal label on $\mathcal{X}^\Delta$, it suffices that

$$k_0 \left(\frac{1}{2} + \Delta - C_\alpha \tau^\alpha - \epsilon \right) \geq \frac{k}{2}.$$

Since $k_0 \geq (\frac{1}{2} + \gamma) \cdot k$, we have that the above holds if

$$\Delta \geq C_\alpha \tau^\alpha + \epsilon + \frac{1-2\gamma}{2(1+2\gamma)}.$$

We now choose the values of τ, γ, ϵ to upper bound each of the terms on the R.H.S. by $\Delta/3$ so that the above holds.

We can bound the last term by $\Delta/3$ by setting:

$$\gamma = \frac{1}{2} \cdot \frac{3-2\Delta}{3+2\Delta}.$$

Next, taking

$$k \geq \frac{3(3+2\Delta)}{2\Delta^2} \left(D \log n + \log(4D/\delta) \right),$$

we have that $\epsilon \leq \Delta/3$. Now, by Lemma 1, we have that setting

$$\tau = \left(\frac{2k}{\omega \cdot v_D \cdot n \cdot p_{X,0}} \right)^{1/D}$$

gives us that $r_k(x) \leq \tau$ for all $x \in \mathcal{X}$ with probability at least $1-\delta/2$. It thus suffices to take

$$k \leq \frac{1}{2} \left(\frac{\Delta}{3 \cdot C_\alpha} \right)^{D/\alpha} \cdot \omega \cdot v_D \cdot p_{X,0} \cdot n$$

so that $C_\alpha \tau^\alpha \leq \Delta/3$. Now in order for $S_{\lfloor (\frac{1}{2}-\gamma) \cdot k \rfloor}(C) \geq \tau$, it suffices to have $S_2(C) \geq \tau$. This can be accomplished by having the following hold:

$$k \leq \frac{1}{2} \cdot S_2(C)^D \cdot \omega \cdot v_D \cdot p_{X,0} \cdot n.$$

Thus, there exists positive constants K_l and K_u depending only on $\mathcal{F}$ such that if

$$K_l \cdot \frac{1}{\Delta^2} \cdot (\log^2(1/\delta) \cdot \log n) \leq k \leq K_u \cdot \min\{S_2(C)^D, \Delta^{D/\alpha}\} \cdot n,$$

then the desired conditions hold. □

B.4. Proof of Theorem 2

Proof of Theorem 2. The proof begins in the same way as the proof of Theorem 1. As before, let $\tau, \gamma, \epsilon > 0$ be quantities that will be determined later. Like before, we are reduced to showing

$$\Delta \geq C_\alpha \tau^\alpha + \epsilon + \frac{1-2\gamma}{2(1+2\gamma)},$$

as long as the conditions for Lemma 1 and 2 hold and $S_2(x) \geq \tau$ and $r_k(x) \leq \tau$ where we choose

$$\begin{aligned}
\epsilon &= \sqrt{\frac{D\log n + \log(4D/\delta)}{(1+2\gamma)\cdot k}}, \\
\gamma &= \frac{1}{2}\cdot\frac{3-2\Delta}{3+2\Delta}, \\
\tau &= \left(\frac{2k}{\omega\cdot v_D\cdot n\cdot p_{X,0}}\right)^{1/D}.
\end{aligned}$$

These conditions hold for some K_u and K_l depending on $\mathcal{F}$. Then we are reduced to having

$$\frac{2}{3}\Delta \geq \sqrt{\frac{D\log n + \log(4D/\delta)}{(1+2\gamma)\cdot k}} + C_\alpha\left(\frac{2k}{\omega\cdot v_D\cdot n\cdot p_{X,0}}\right)^{\alpha/D}.$$

Since $\gamma \geq 0$, it suffices to have

$$\frac{2}{3}\Delta \geq \sqrt{\frac{D\log n + \log(4D/\delta)}{k}} + C_\alpha\left(\frac{2k}{\omega\cdot v_D\cdot n\cdot p_{X,0}}\right)^{\alpha/D}.$$

The desired form for Δ clearly follows for some choice of K depending only on $D, \omega, p_{X,0}, C_\alpha$, all of which depend only on $\mathcal{F}$.

Finally, we must ensure that $\lfloor(\frac{1}{2}-\gamma)\cdot k\rfloor \geq 2$ so that $S_{\lfloor(\frac{1}{2}-\gamma)\cdot k\rfloor}(C) \geq S_2(x)$. Given the expression for γ, it is equivalent to have $\lfloor\left(\frac{2\Delta}{3+2\Delta}\right)\cdot k\rfloor \geq 2$. It suffices to show that $k \geq \frac{3(3+2\Delta)}{2\Delta}$. Given the form of Δ in terms of n and k, we see that it suffices to have that

$$k \geq \frac{9}{2}\cdot K\cdot\left(\sqrt{\frac{\log n + \log(1/\delta)}{k}} + \left(\frac{k}{n}\right)^{\alpha/D}\right)^{-1} + 3,$$

which holds when $k \geq K_0 \cdot n^{2\alpha/(2\alpha+D)}$ for some K_0 depending only on $\mathcal{F}$, as desired. □

B.5. Proof of Theorem 3

Proof of Theorem 3. The first part follows from Theorem 2. For the second part, we have by Theorem 2 that if $x \in \mathcal{X}^\Delta$, then the k-NN classifier and the Bayes-optimal classifier match with probability $1-\delta$ uniformly. Thus, we have

$$\begin{aligned}
R_X - R^* &\leq \mathbb{P}(x \notin \mathcal{X}^\Delta)(\mathbb{E}_\mathcal{F}[g_k(x) \neq y | x \notin \mathcal{X}^\Delta] - \mathbb{E}_\mathcal{F}[g^*(x) \neq y | x \notin \mathcal{X}^\Delta]) \\
&\leq C_\beta \cdot \Delta^\beta \cdot (\mathbb{E}_\mathcal{F}[g_k(x) \neq y | x \notin \mathcal{X}^\Delta] - \mathbb{E}_\mathcal{F}[g^*(x) \neq y | x \notin \mathcal{X}^\Delta]) \\
&\leq C_\beta \cdot \Delta^\beta \cdot 2\Delta.
\end{aligned}$$

The result follows immediately from Theorem 2. □

C. Hard Flip Permutations

For Fashion MNIST we hard flip by swapping the following classes: TSHIRT ↔ SHIRT, TROUSER ↔ DRESS, PULLOVER ↔ COAT, SANDAL ↔ BAG, SNEAKER ↔ ANKLEBOOT. For CIFAR10 we swap the pairs: TRUCK ↔ AUTOMOBILE, BIRD ↔ AIRPLANE, DEER ↔ HORSE, CAT ↔ DOG, FROG ↔ SHIP. For CIFAR100, we hard flip circularly (i.e. $\pi(i) = (i+1) \mod K$) within each of the 20 superclasses. For all other datasets, we hard flip circularly.

D. Area-Under-Curve Table

We provide the Area-Under-the-Curve tables for the case when clean auxiliary data is available. These were omitted from the main text due to space constraints.

Flip								
Dataset	% Clean	Forward	Clean	Distill	GLC	k-NN	k-NN Classify	Full
Letters	5	4.83	3.16	2.96	2.35	**2.1**	2.52	2.92
	10	4.45	2.52	2.12	1.92	**1.8**	2.06	2.49
	20	3.85	2.07	1.78	1.56	1.58	**1.42**	1.93
Phonemes	5	7.91	1.93	3.21	2.16	**1.34**	3.97	4.31
	10	7.95	1.54	2.75	1.69	**1.22**	3.64	3.96
	20	7.89	1.34	1.97	1.35	**1.16**	2.87	3.28
Wilt	5	5.27	0.56	3.89	0.53	**0.52**	5.15	4.95
	10	5.63	0.44	3.14	0.43	**0.41**	4.86	4.84
	20	5.18	0.34	2.67	0.35	**0.34**	4.23	4.32
Seeds	5	5.13	4.33	5.56	4.39	**3.64**	5.11	4.88
	10	5.02	3.38	4.58	3.19	**2.65**	4.9	4.73
	20	4.41	2.56	3.57	2.65	**2.09**	4.23	3.86
Iris	5	5.25	3.38	5.15	4.03	**3.02**	4.32	4.43
	10	5.06	2.53	4.58	2.24	**2.17**	3.98	4.08
	20	4.46	1.51	3.45	**1.38**	1.46	3.36	3.53
Parkinsons	5	5.17	3.55	4.49	4.1	**3.36**	5.34	5.35
	10	5.43	3.38	4.25	3.44	**3.32**	5.27	5.13
	20	5.19	3.02	3.94	3.05	**2.95**	4.99	5.06
MNIST	5	3.6	0.69	1.91	0.5	**0.44**	3.46	3.49
	10	3.22	0.5	1.5	0.42	**0.35**	3.1	3.14
	20	2.48	0.35	0.86	0.34	**0.27**	2.36	2.41
Fashion MNIST	5	3.55	1.87	2.55	**1.59**	1.6	3.3	3.31
	10	3.14	1.71	2.13	**1.53**	1.54	2.92	2.99
	20	2.38	1.56	1.54	1.46	**1.46**	2.17	2.27
CIFAR10	5	7.14	7.2	7.12	5.52	**5.35**	6.71	7.12
	10	6.82	6.56	6.72	5.62	**5.32**	6.48	6.83
	20	6.62	5.59	5.9	5.16	**4.85**	6.1	6.56
CIFAR100	5	10.79	10.24	10.03	9.64	9.66	**9.2**	9.29
	10	10.81	9.89	9.68	9.46	9.63	**9.1**	9.25
	20	10.8	9.44	9.13	8.97	9.17	**8.92**	9.09
SVHN	5	5.6	3.65	4.09	2.17	**2.05**	5.23	5.52
	10	5.5	2.28	3.72	2.29	**1.48**	4.96	5.29
	20	4.85	1.83	3.09	1.9	**1.19**	4.34	4.82

Table 2. AUC for Flip corruption type.

Hard Flip								
Dataset	% Clean	Forward	Clean	Distill	GLC	k-NN	k-NN Classify	Full
Letters	5	3.77	3.17	2.1	1.83	**1.82**	1.82	2.52
	10	3.63	2.52	1.85	1.6	**1.6**	1.64	2.36
	20	3.35	2.05	1.62	**1.38**	1.41	1.42	2.04
Phonemes	5	6.6	1.92	1.77	1.96	**1.2**	1.78	2.7
	10	6.73	1.55	1.61	1.6	**1.15**	1.62	2.58
	20	6.36	1.33	1.45	1.3	**1.14**	1.38	2.28
Wilt	5	4.6	0.55	0.73	0.58	**0.39**	0.98	1.9
	10	4.78	0.43	0.66	0.43	**0.31**	0.78	1.81
	20	5.63	0.34	0.61	0.36	**0.3**	0.57	1.49
Seeds	5	2.94	4.06	3.63	3.91	2.99	**2.84**	3.01
	10	2.75	3.42	2.96	3.14	**2.25**	2.69	2.86
	20	2.85	2.53	2.25	2.57	**1.62**	2.43	2.49
Iris	5	2.9	3.29	3.13	4.1	2.32	**1.14**	1.35
	10	2.6	2.34	2.02	1.86	1.15	**0.91**	1.16
	20	2.51	1.84	1.48	1.34	**0.58**	0.65	1
Parkinsons	5	4.98	3.71	3.56	4.59	3.68	**3.28**	3.63
	10	5.24	3.26	**3.1**	3.37	3.11	3.12	3.82
	20	5.1	2.98	3.01	2.98	2.96	**2.91**	3.47
MNIST	5	2.03	0.69	0.78	**0.22**	0.29	0.65	2.14
	10	1.86	0.5	0.67	**0.21**	0.26	0.48	1.98
	20	1.54	0.35	0.53	**0.2**	0.22	0.36	1.67
Fashion MNIST	5	2.3	1.87	1.62	**1.44**	1.48	2.31	2.4
	10	2.14	1.71	1.52	**1.41**	1.46	2.19	2.22
	20	1.92	1.56	1.43	**1.38**	1.41	2.04	1.97
CIFAR10	5	5.08	7.13	5.85	**3.76**	4.27	4.33	4.96
	10	4.89	6.53	5.3	**3.91**	4.4	4.21	4.89
	20	4.77	5.45	4.68	**3.51**	3.82	4.03	4.75
CIFAR100	5	10.64	10.23	9.88	8.58	8.98	**7.48**	8.04
	10	10.65	9.89	9.38	8.56	9.19	**7.44**	7.99
	20	10.66	9.41	8.65	8.07	8.81	**7.33**	7.9
SVHN	5	3.04	4.17	2.16	**0.89**	1.35	2.32	3.04
	10	2.98	2.41	2.01	**0.88**	1.04	2.16	2.98
	20	2.7	1.85	1.54	0.98	**0.98**	1.83	2.74

Table 3. AUC for Hard Flip corruption type.

E. Additional Plots

We provide the plots that were omitted from the main text due to space constraints.

E.1. Plots for Experiments with Clean Auxiliary Data

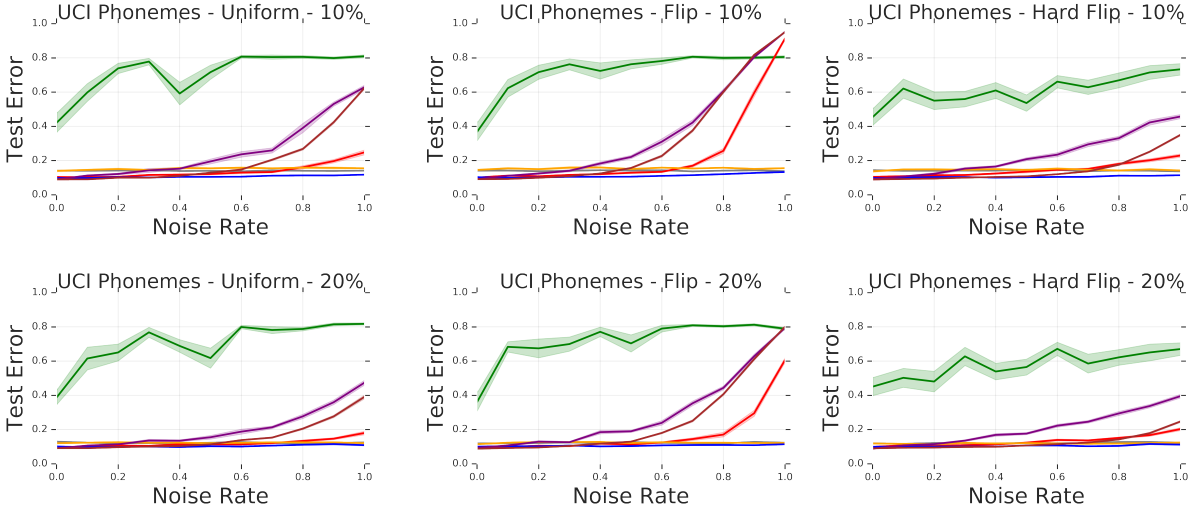


Figure 6. Plots for UCI Phonemes dataset at 10, 20% clean data and all corruption types.

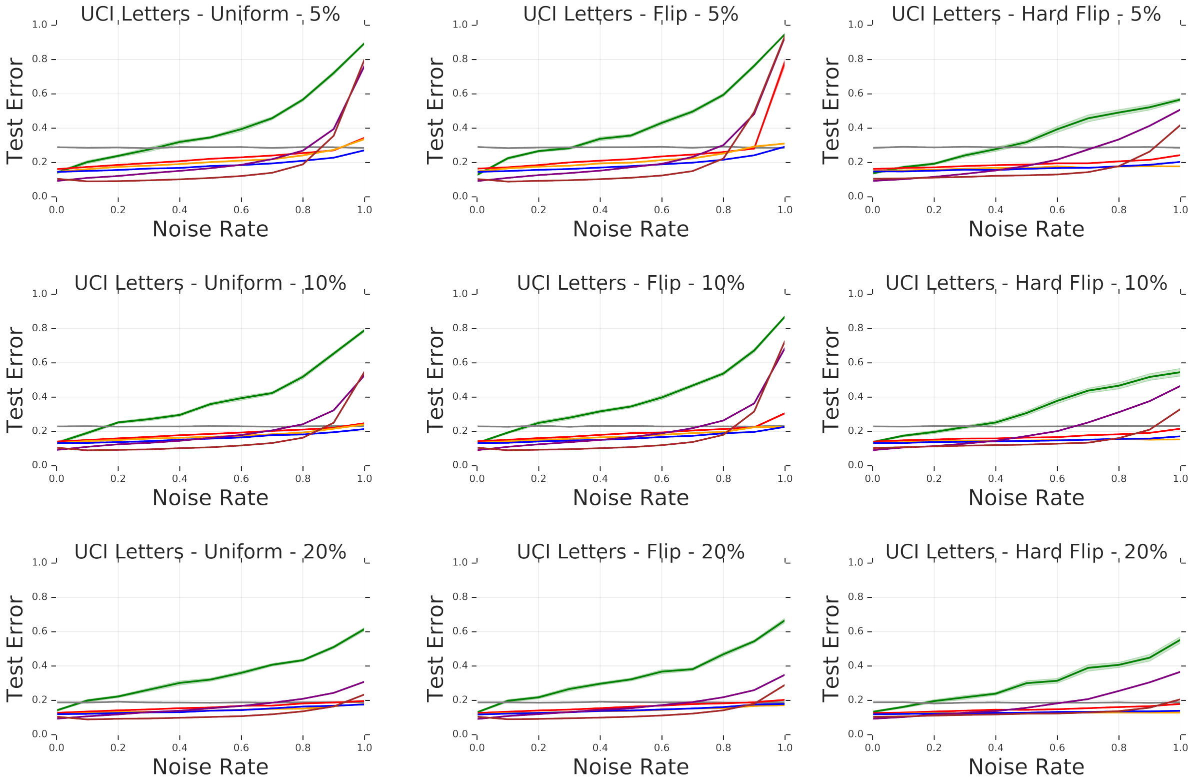


Figure 7. Plots for UCI Letters dataset at 5, 10, 20% clean data and all corruption types.

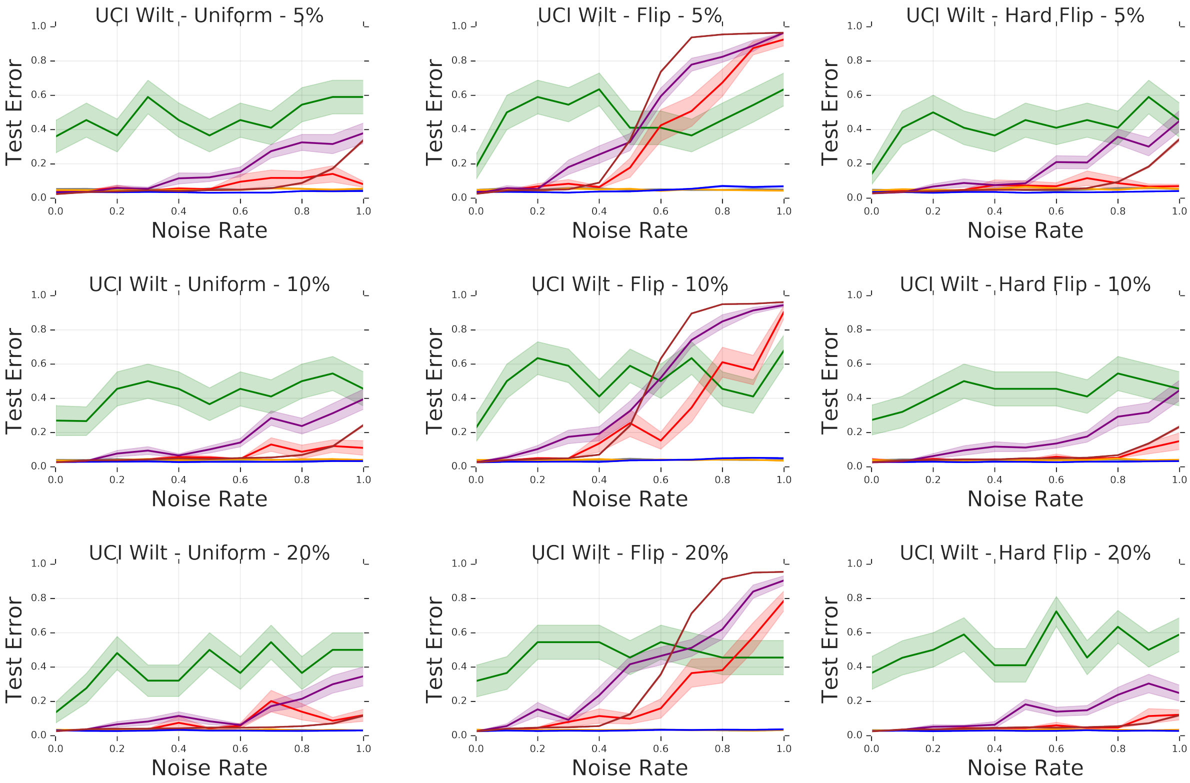


Figure 8. Plots for UCI Wilt dataset at 5, 10, 20% clean data and all corruption types.

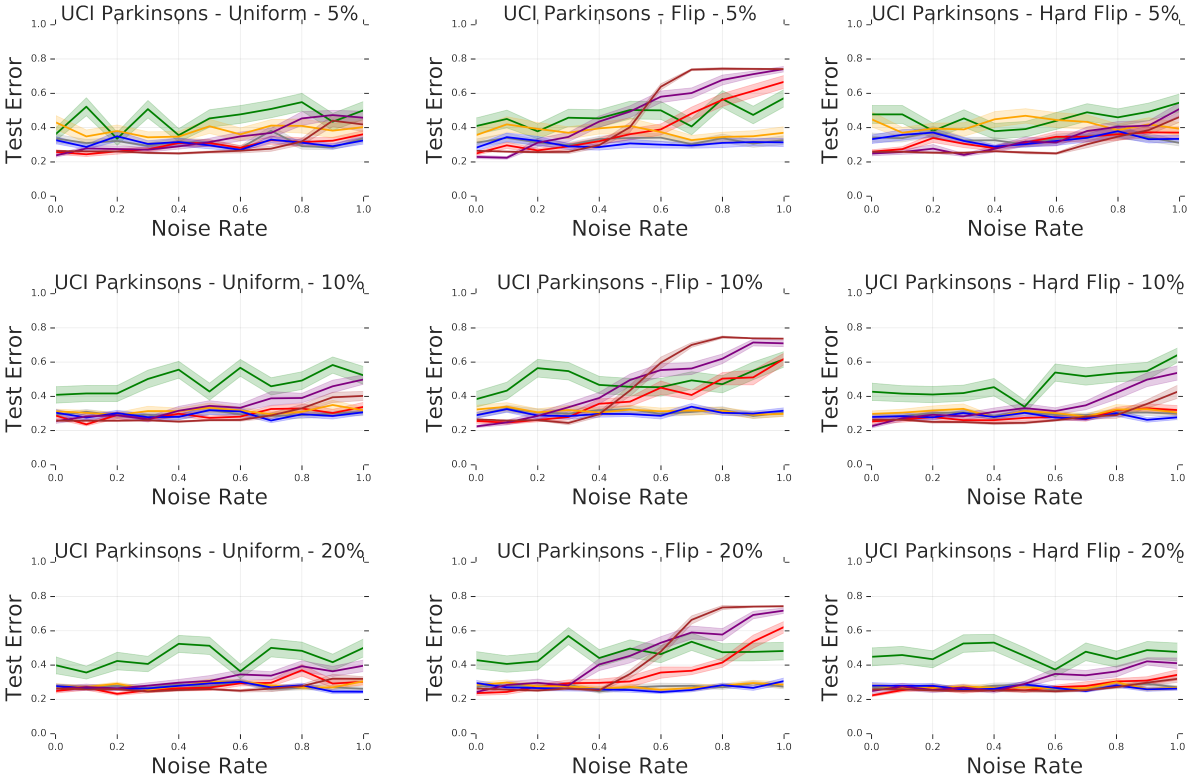


Figure 9. Plots for UCI Parkinsons dataset at 5, 10, 20% clean data and all corruption types.

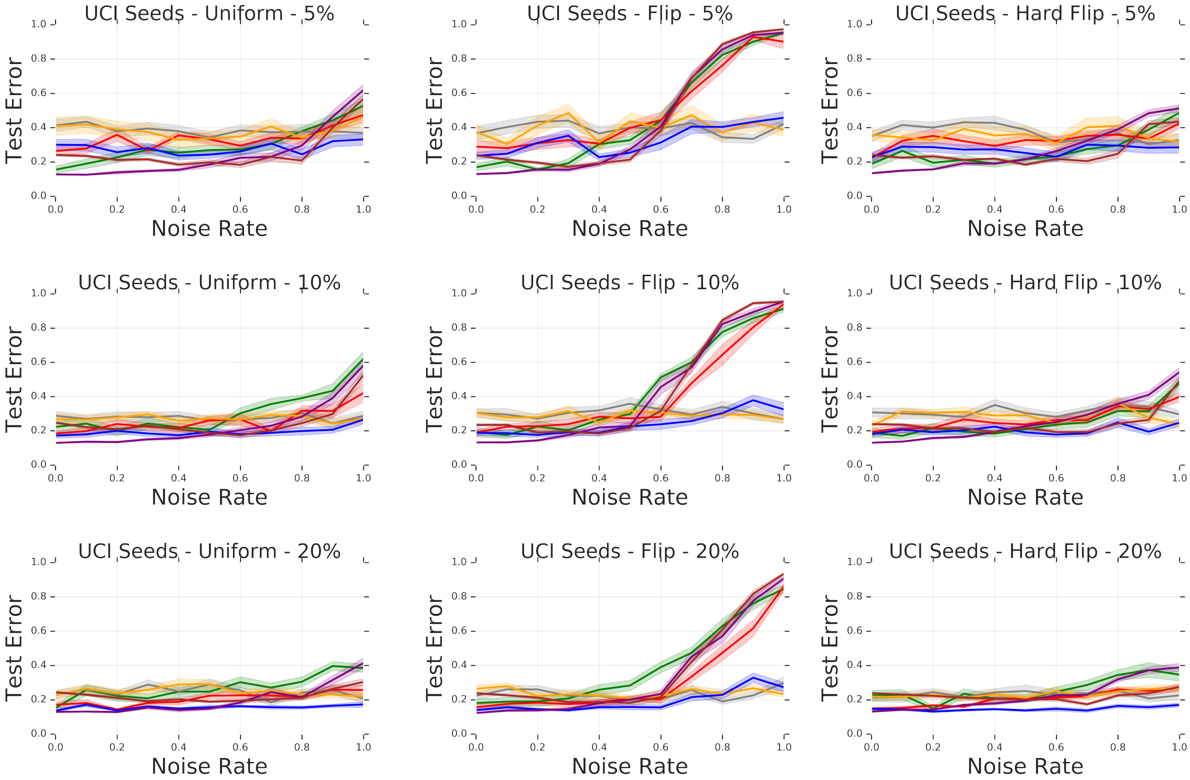


Figure 10. Plots for UCI Seeds dataset at 5, 10, 20% clean data and all corruption types.

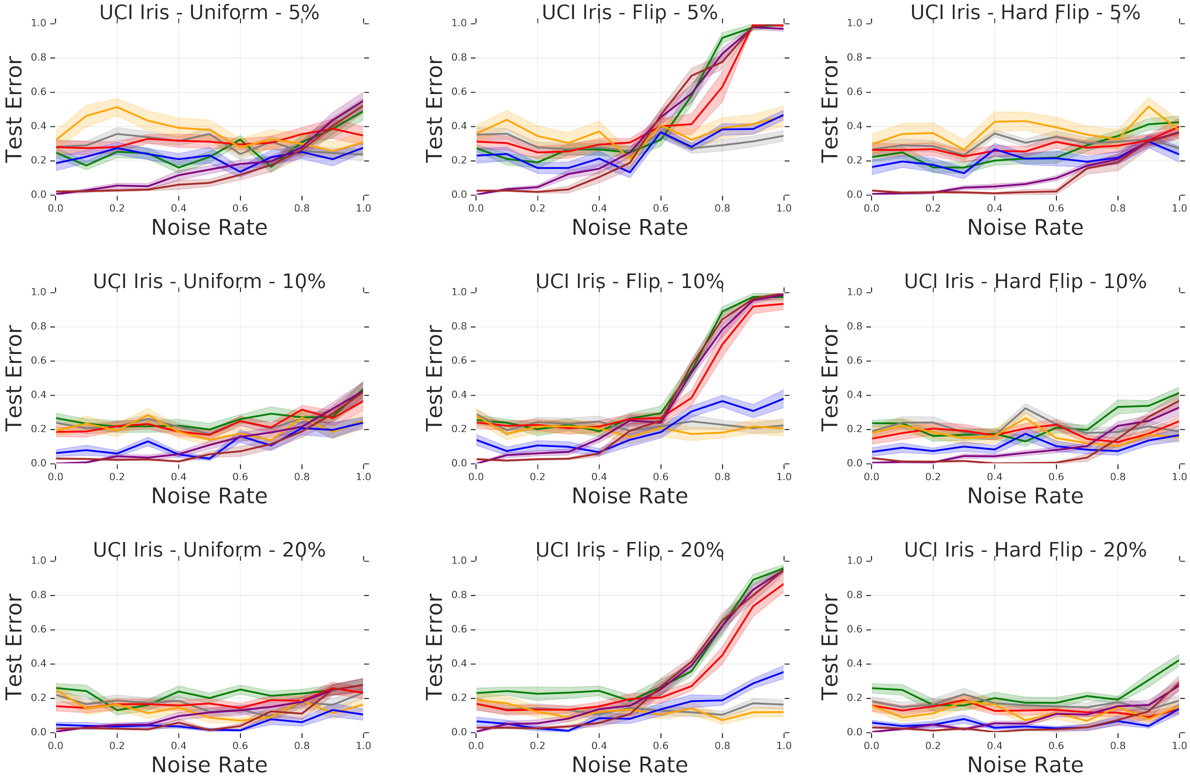


Figure 11. Plots for UCI Iris dataset at 5, 10, 20% clean data and all corruption types.

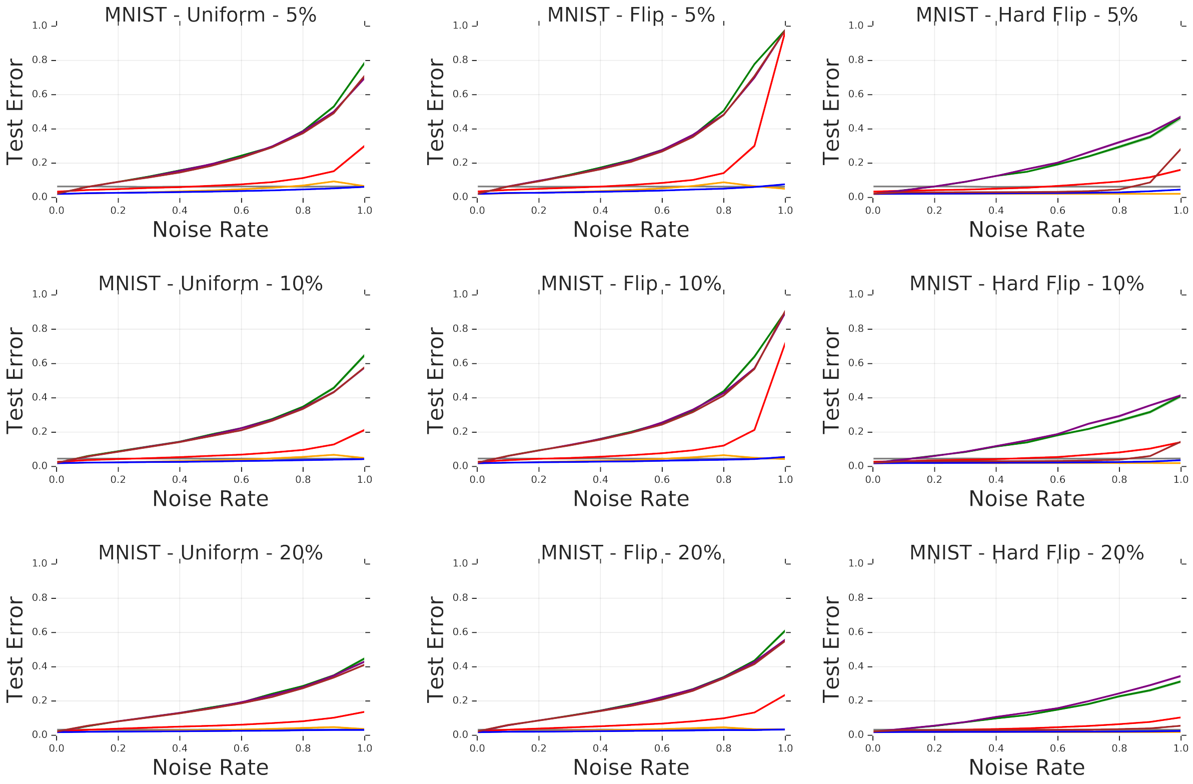


Figure 12. Plots for MNIST at 5, 10, 20% clean data and all corruption types.

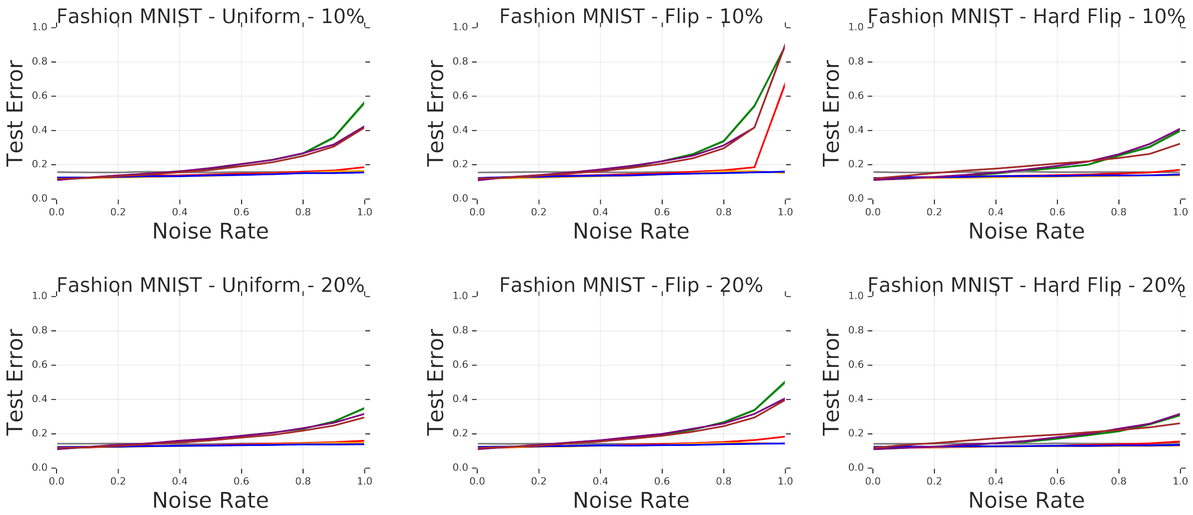
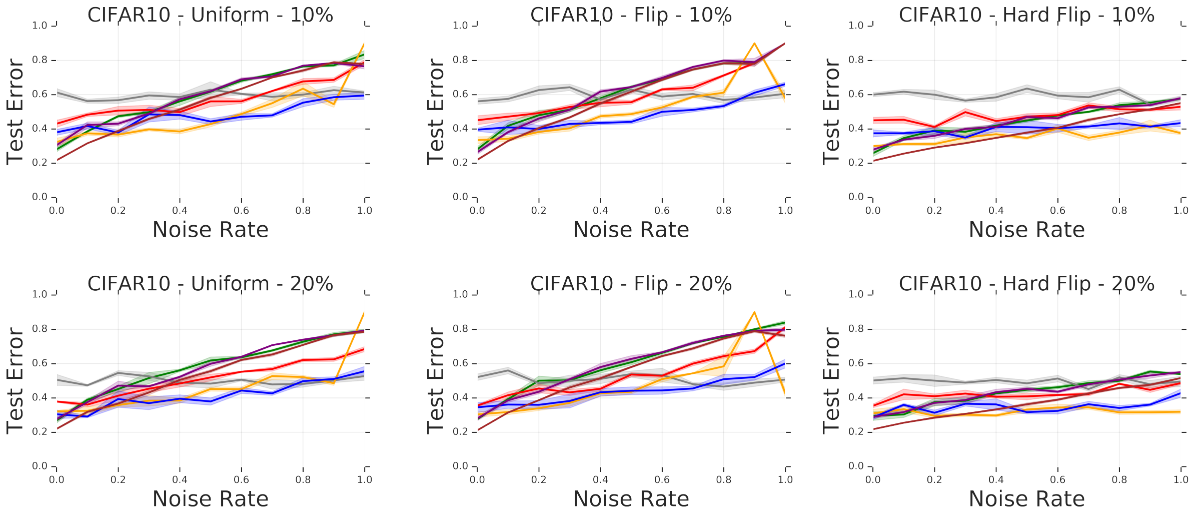


Figure 13. Plots for Fashion MNIST at 10, 20% clean data and all corruption types.



CIFAR10 - Uniform - 10%

CIFAR10 - Flip - 10%

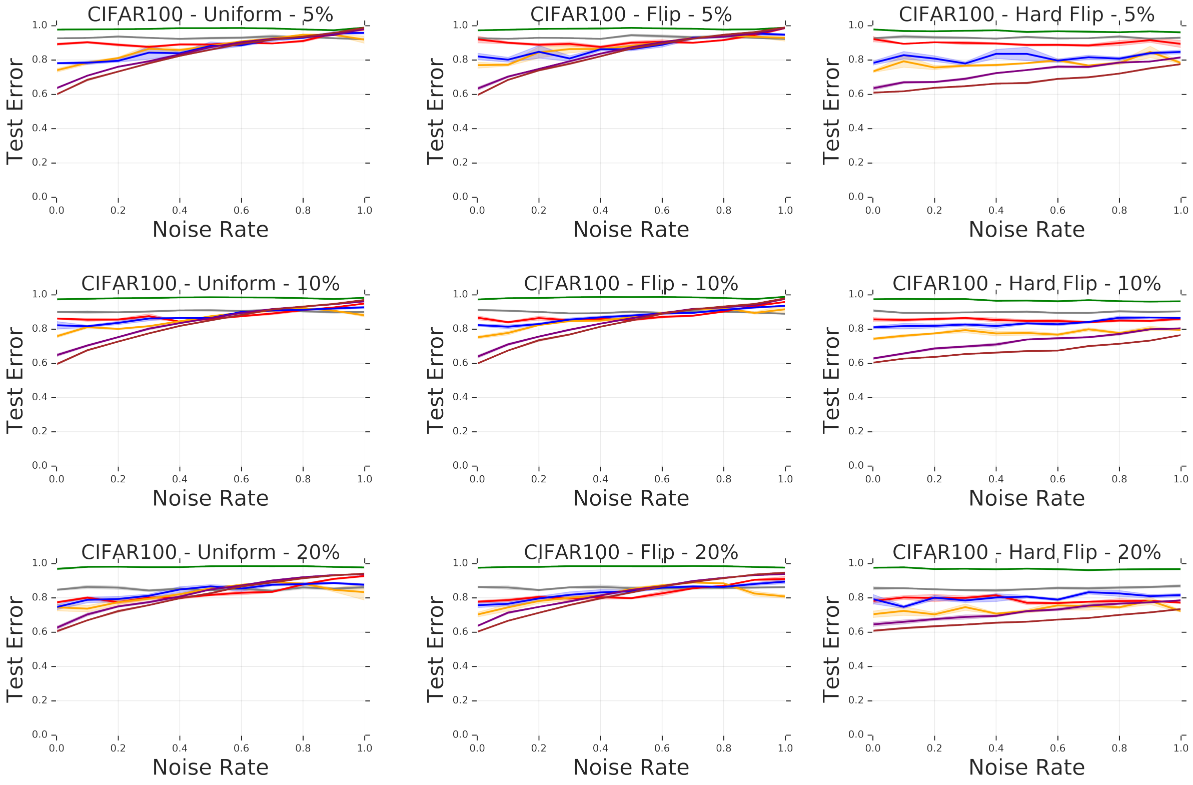
CIFAR10 - Hard Flip - 10%

CIFAR10 - Uniform - 20%

CIFAR10 - Flip - 20%

CIFAR10 - Hard Flip - 20%

Figure 14. Plots for CIFAR10 at 10, 20% clean data and all corruption types.



CIFAR100 - Uniform - 5%

CIFAR100 - Flip - 5%

CIFAR100 - Hard Flip - 5%

CIFAR100 - Uniform - 10%

CIFAR100 - Flip - 10%

CIFAR100 - Hard Flip - 10%

CIFAR100 - Uniform - 20%

CIFAR100 - Flip - 20%

CIFAR100 - Hard Flip - 20%

Figure 15. Plots for CIFAR100 at 5, 10, 20% clean data and all corruption types.

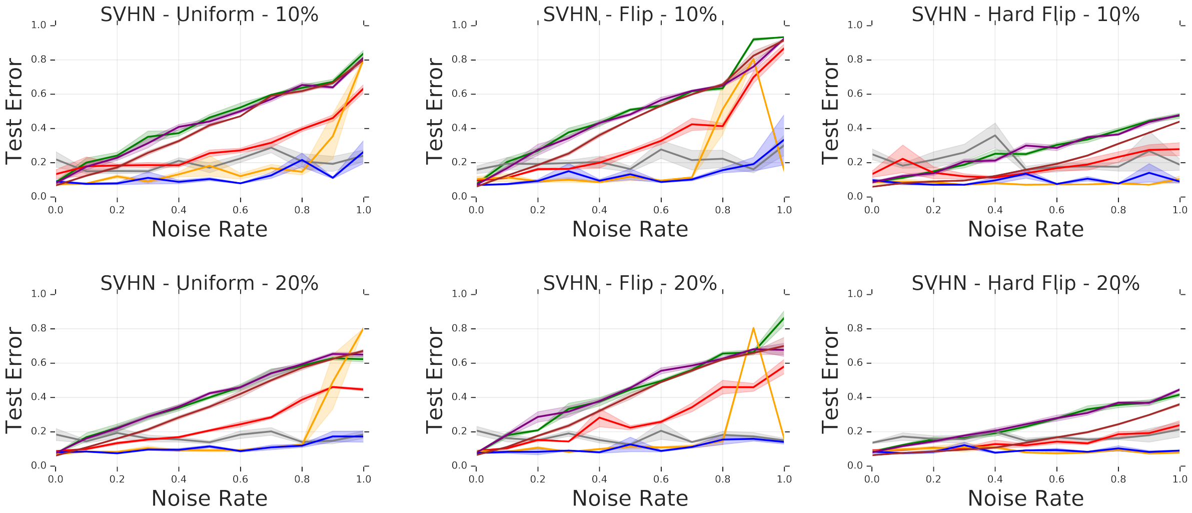


Figure 16. Plots for SVHN at 10, 20% clean data and all corruption types.

E.2. Additional Plot For Robustness to k

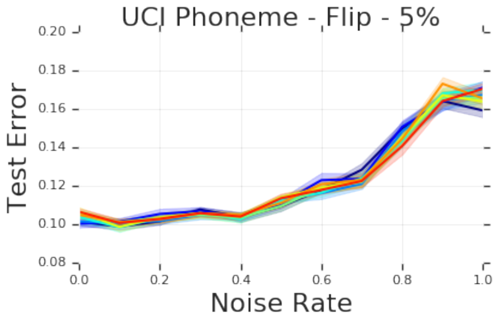


Figure 17. Performance across different values of k under Flip corruption type.

E.3. Additional Plots For Experiments Without Clean Auxiliary Data

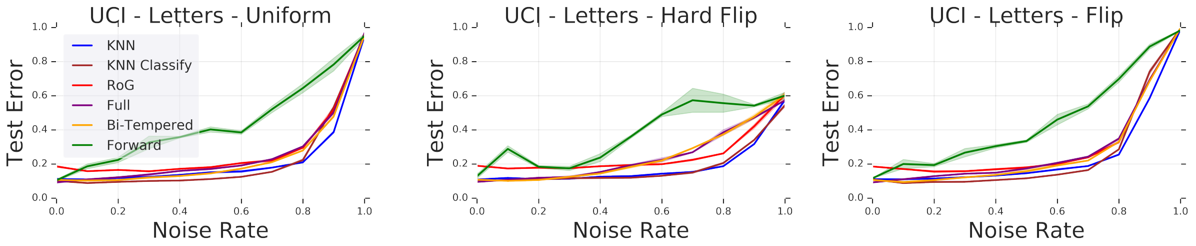


Figure 18. Plots for UCI Letters for all corruption types.

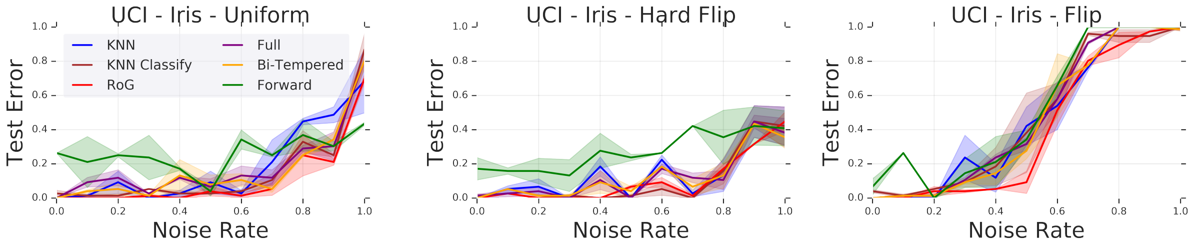


Figure 19. Plots for UCI Iris for all corruption types.

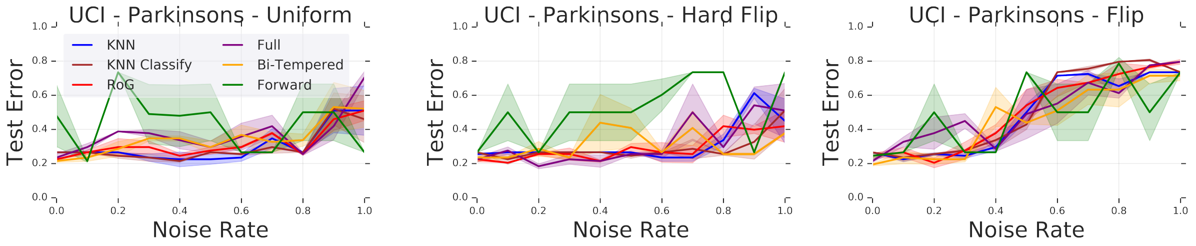


Figure 20. Plots for UCI Parkinsons for all corruption types.

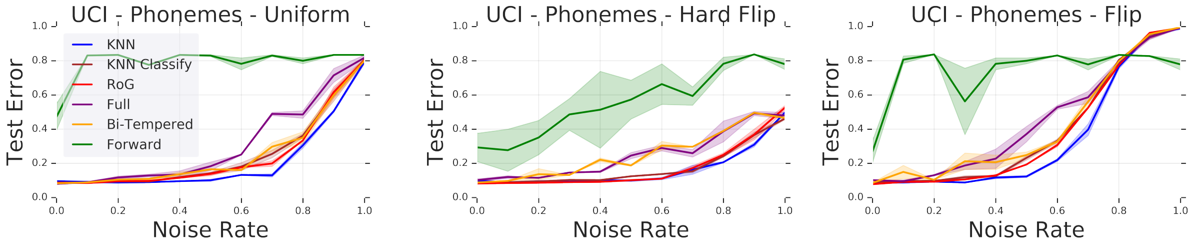


Figure 21. Plots for UCI Phonemes for all corruption types.

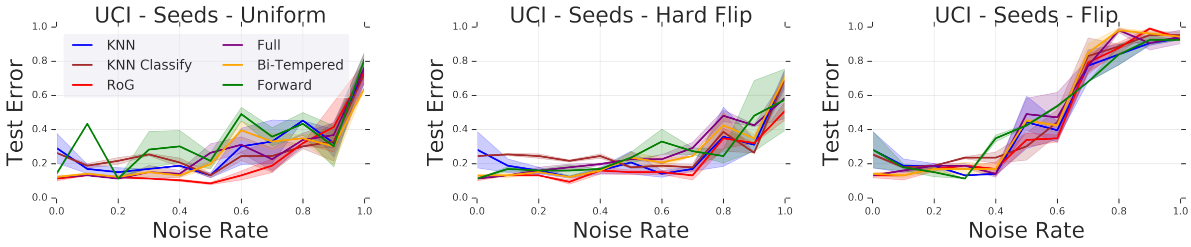


Figure 22. Plots for UCI Seeds for all corruption types.

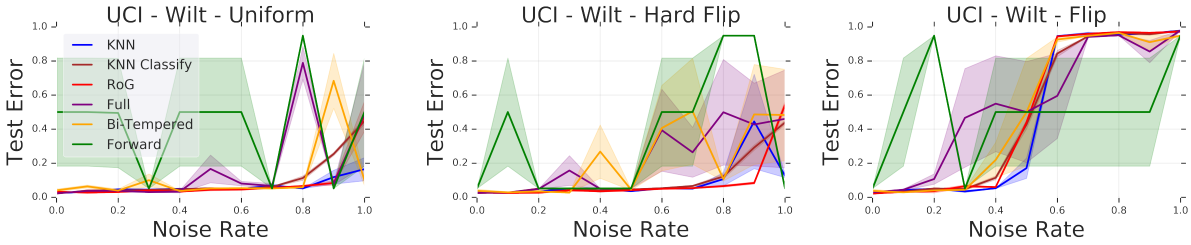


Figure 23. Plots for UCI Wilt for all corruption types.

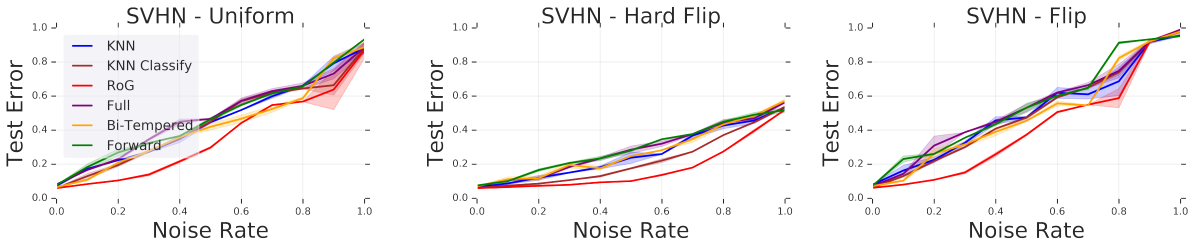


Figure 24. Plots for SVHN for all corruption types.

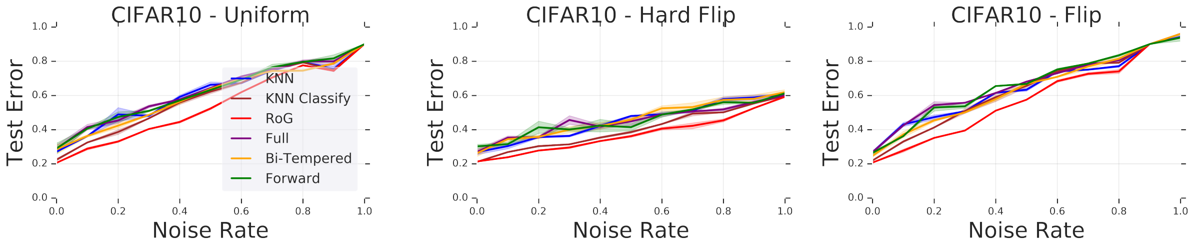


Figure 25. Plots for CIFAR10 for all corruption types.

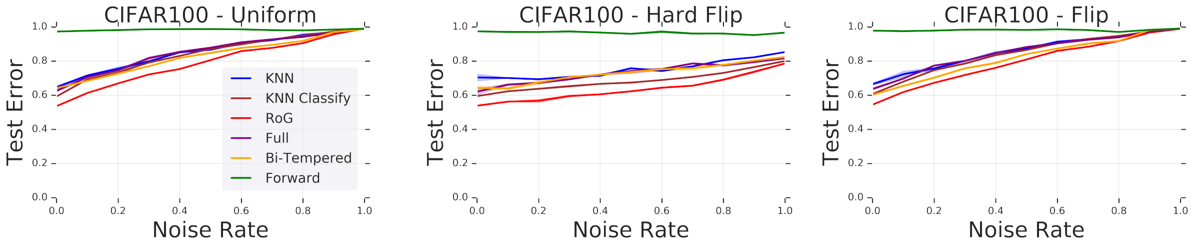


Figure 26. Plots for CIFAR100 for all corruption types.