The Effect of Data Poisoning on Counterfactual Explanations – Appendix

No Author Given

No Institute Given

1 Proofs

1.1 Proof of Theorem 1

Proof. Sketch: For any $\vec{x}_{\text{orig}}, h(\vec{x}_{\text{orig}}) = y_{\text{orig}}$, assume uniqueness of the solution $\vec{x}'$ – i.e. the closest sample to $\vec{x}_{\text{orig}}$ on the decision boundary:

$$\begin{aligned}&\underset{\vec{x}' \in \mathbb{R}^d}{\arg\min}\, \|\vec{x}' - \vec{x}_{\text{orig}}\|_p \text{ s.t.}\\ &\exists i \neq j : (\vec{x}_i, y_i), (\vec{x}_j, y_j) \in \mathcal{D}, y_i \neq y_j,\ \text{with } \|\vec{x}' - \vec{x}_i\|_p = \|\vec{x}' - \vec{x}_j\|_p\end{aligned} \tag{1}$$

where we (w.l.o.g.) assume the use of the p-norm as the distance function in the 1-nearest neighbor classifier.

Adding $(\vec{x}', y_{\text{orig}})$ to the training data $\mathcal{D}$ implies that $\vec{x}'$ is no longer the solution to Eq. (1). Therefore, the new closest sample on the decision boundary must have a larger distance to $\vec{x}_{\text{orig}}$ than $\vec{x}'$, otherwise it would have been $\vec{x}'$ before! □

1.2 Proof of Theorem 2

Proof. Sketch: From the triangle-inequality and $\lambda > \|\vec{x}_i - \vec{x}_j\|_2$ it follows that:

$$\|\vec{x}_i - \vec{x}_j\|_2 + \delta'_j \geq \underbrace{\delta_i + \lambda}_{\delta'_i} \quad \leftrightarrow \quad \delta'_j \geq \delta_i + \lambda - \|\vec{x}_i - \vec{x}_j\|_2 \tag{2}$$

Because of $\delta_j > \delta_i$, we know that $\delta_j = \alpha\delta_i$ for some $\alpha > 1$. This allows us to rewrite Eq. (2):

$$\delta'_j \geq \underbrace{\frac{\delta_j}{\alpha}}_{\delta_i} + \lambda - \|\vec{x}_i - \vec{x}_j\|_2 \tag{3}$$

The desired results follows from choosing $\lambda \geq 2\alpha\delta_j + \|\vec{x}_i - \vec{x}_j\|_2$ yields:

$$\begin{aligned}\delta'_j &\geq \frac{\delta_j}{\alpha} + \lambda - \|\vec{x}_i - \vec{x}_j\|_2\\ &\geq \frac{\delta_j}{\alpha} + 2\alpha\delta_j + \|\vec{x}_i - \vec{x}_j\|_2 - \|\vec{x}_i - \vec{x}_j\|_2\\ &= \delta_j\end{aligned} \tag{4}$$

□

2 Experiments

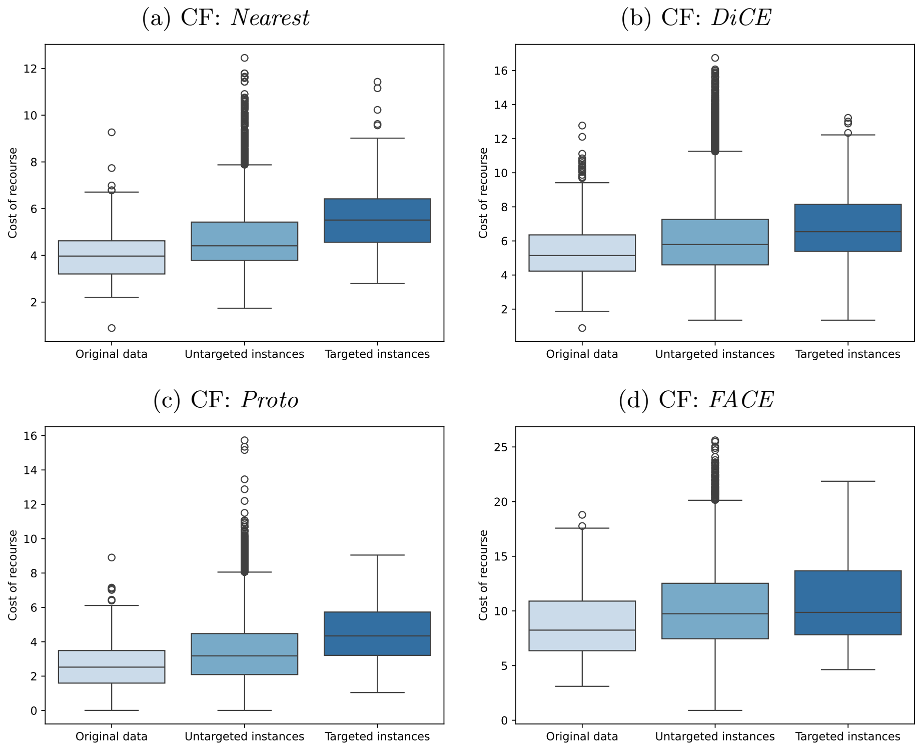
2.1 Details on the Classifiers

- **RandomForest:** 10 decision tree classifiers each with a maximum depth of 7.
- **DNN:** 3-layer neural network with ReLU activation functions.

2.2 Local Poisoning Attack

Classifier	Data set	Nearest ↑	DiCE ↑	FACE ↑	Proto ↑
DNN	Diabetes	1.59	1.42	1.24	1.89

Table 1: Difference in the cost of recourse: no vs. local poisoning. Positive numbers denote an increase in the cost of recourse. We report the median (over all folds) rounded to two decimal places.


(a) CF: *Nearest*

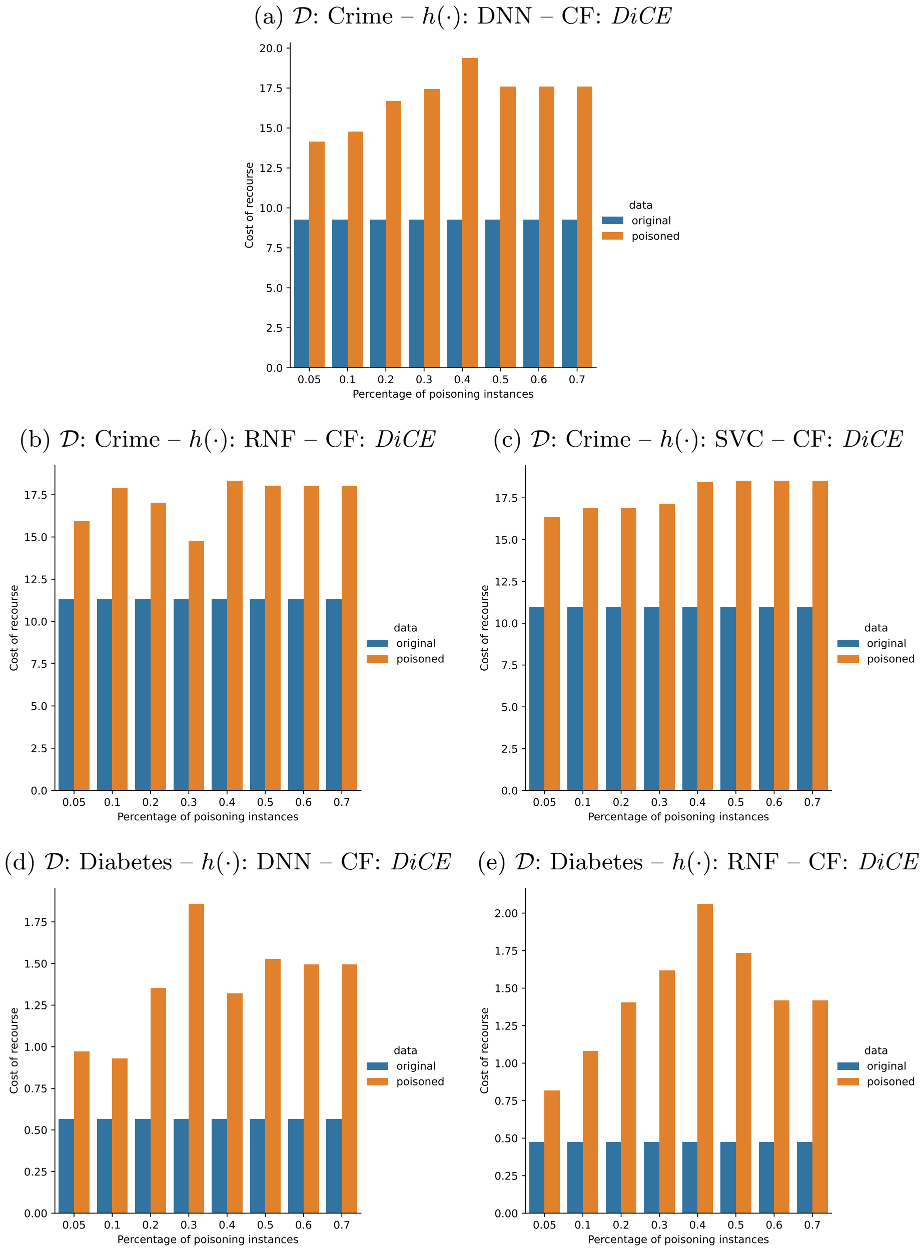
(b) CF: *DiCE*

(c) CF: *Proto*

(d) CF: *FACE*

Fig. 1: *Local* data poisoning: Cost of recourse (over all test samples) in the case of the diabetes data set and a DNN classifier. Cost of recourse without any data poisoning, of untargeted instances and targeted instances in a local data poisoning.

2.3 Sub-group Poisoning Attack



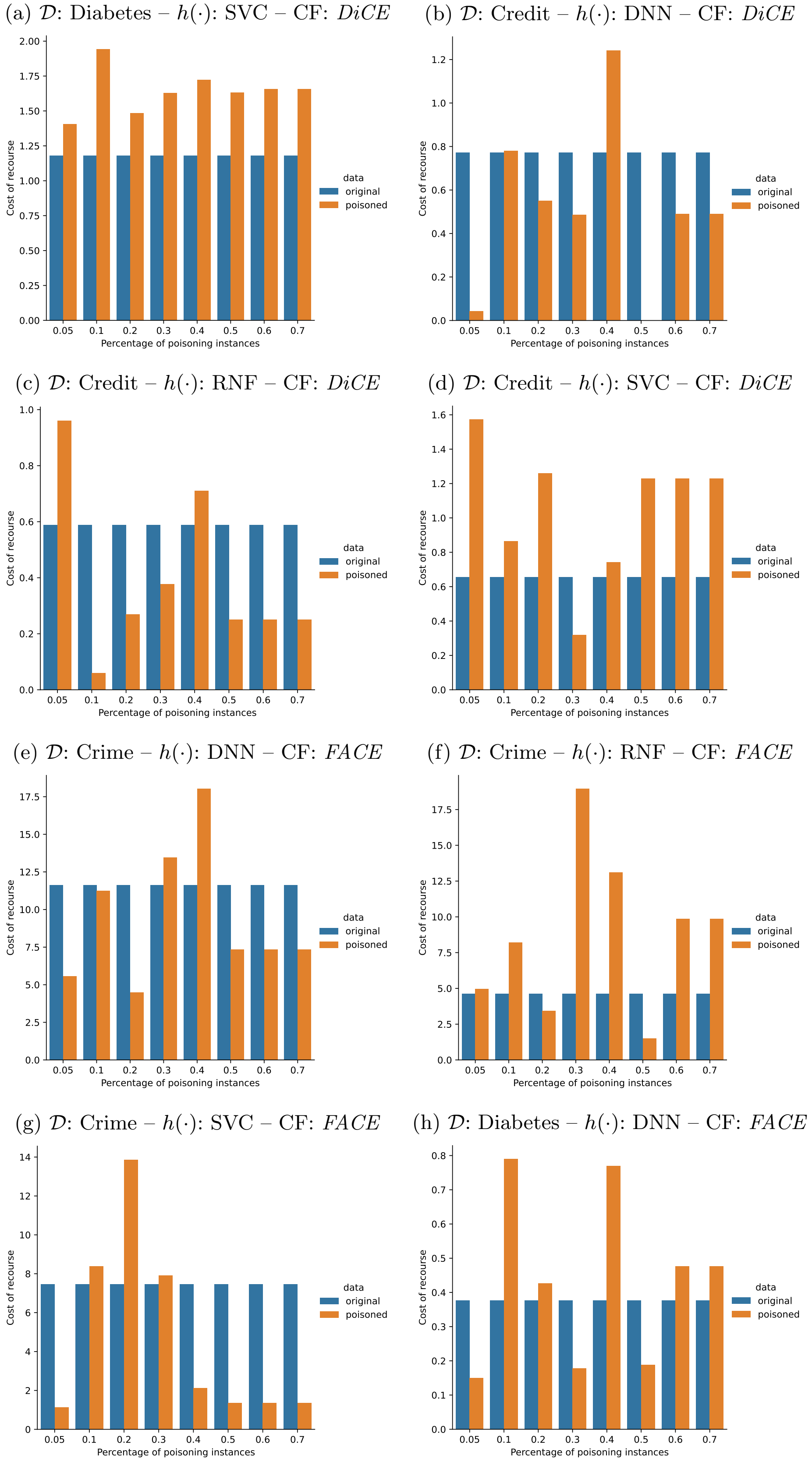
(a) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *DiCE*

(b) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *DiCE*

(c) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *DiCE*

(d) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *DiCE*

(e) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *DiCE*


(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *DiCE*

(b) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *DiCE*

(c) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *DiCE*

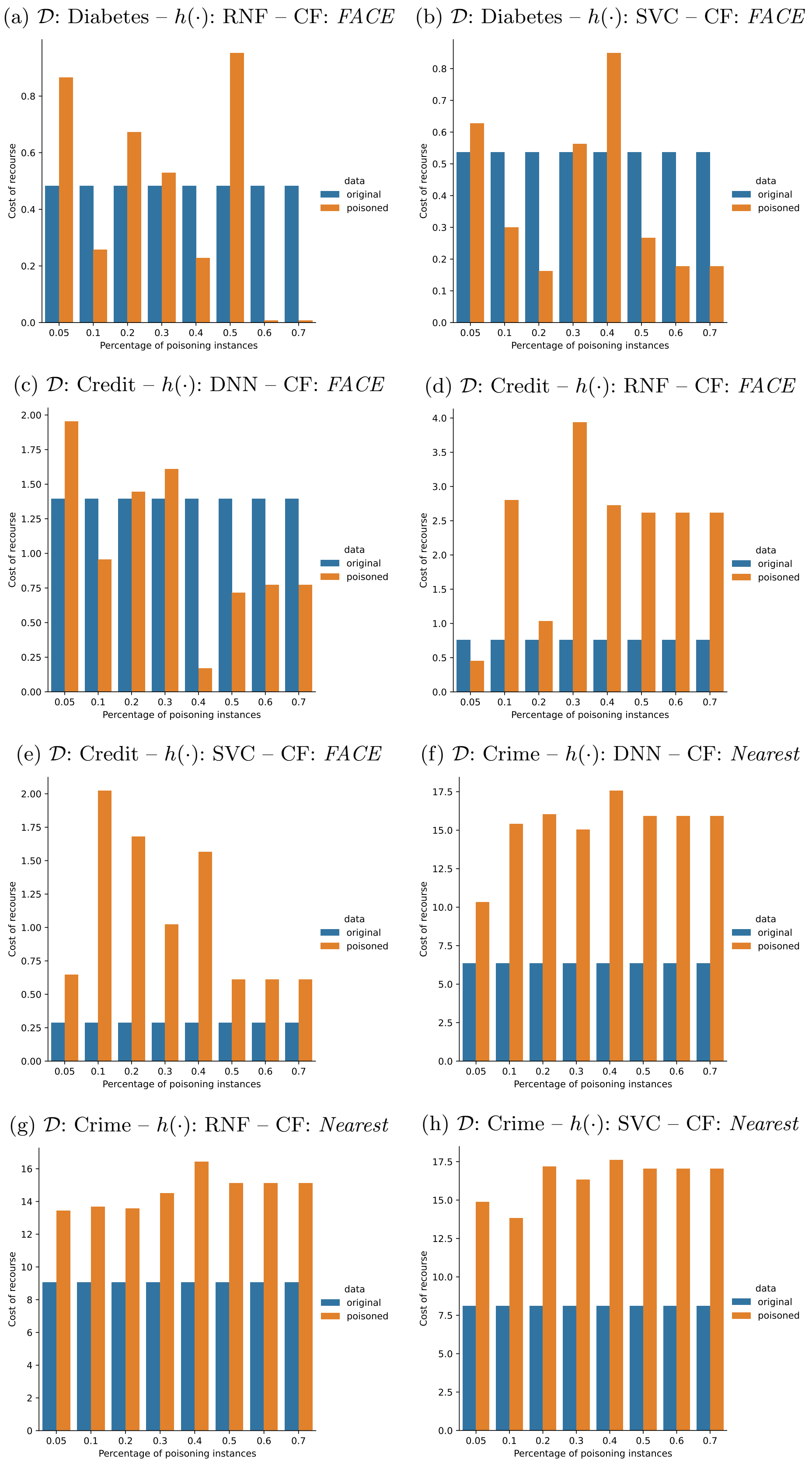
(d) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *DiCE*

(e) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *FACE*

(f) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *FACE*

(g) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *FACE*

(h) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *FACE*



(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *FACE*

(b) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *FACE*

(c) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *FACE*

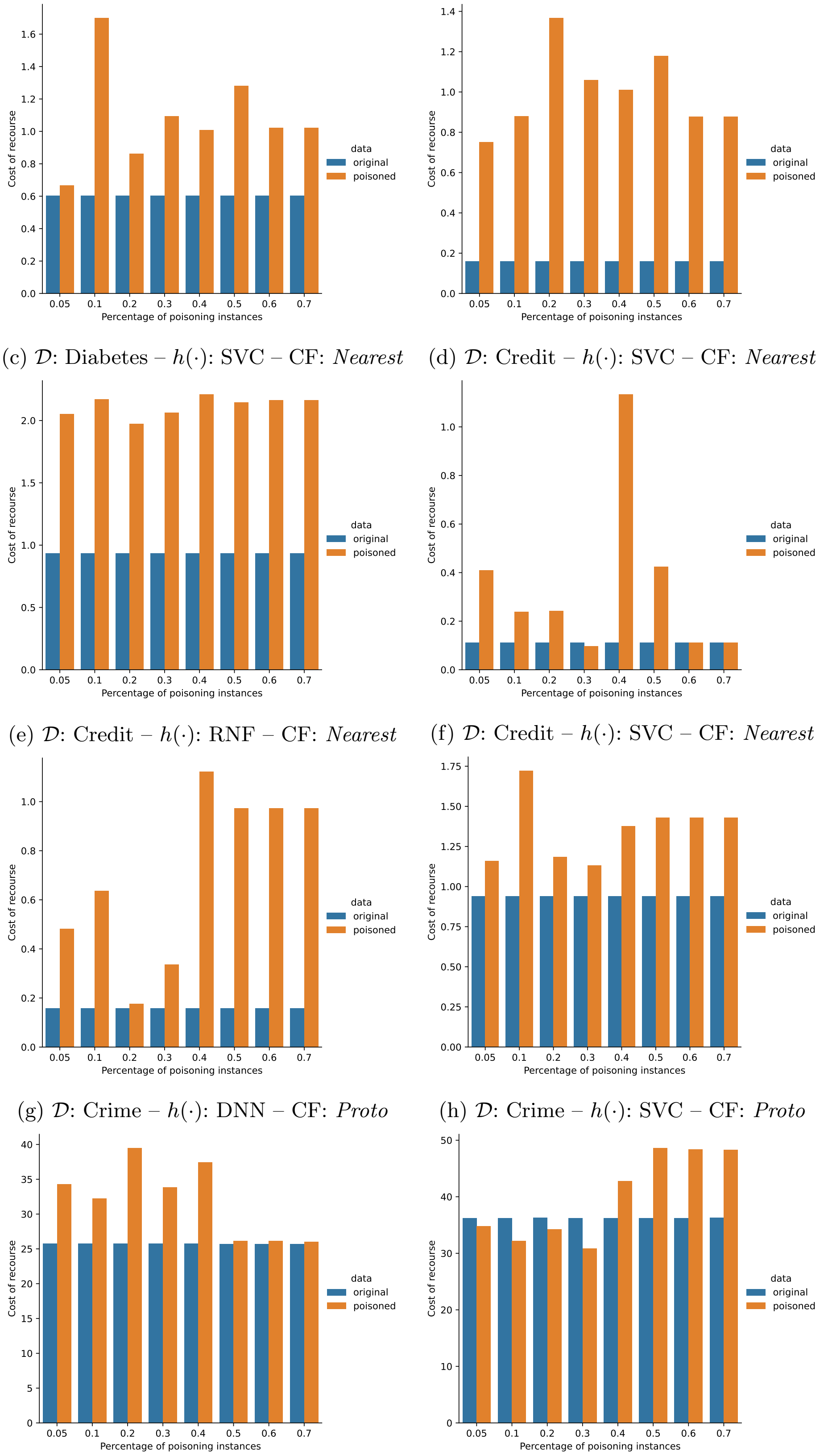
(d) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *FACE*

(e) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *FACE*

(f) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *Nearest*

(g) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *Nearest*

(h) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *Nearest*



(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *Nearest*

(b) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *Nearest*

(c) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *Nearest*

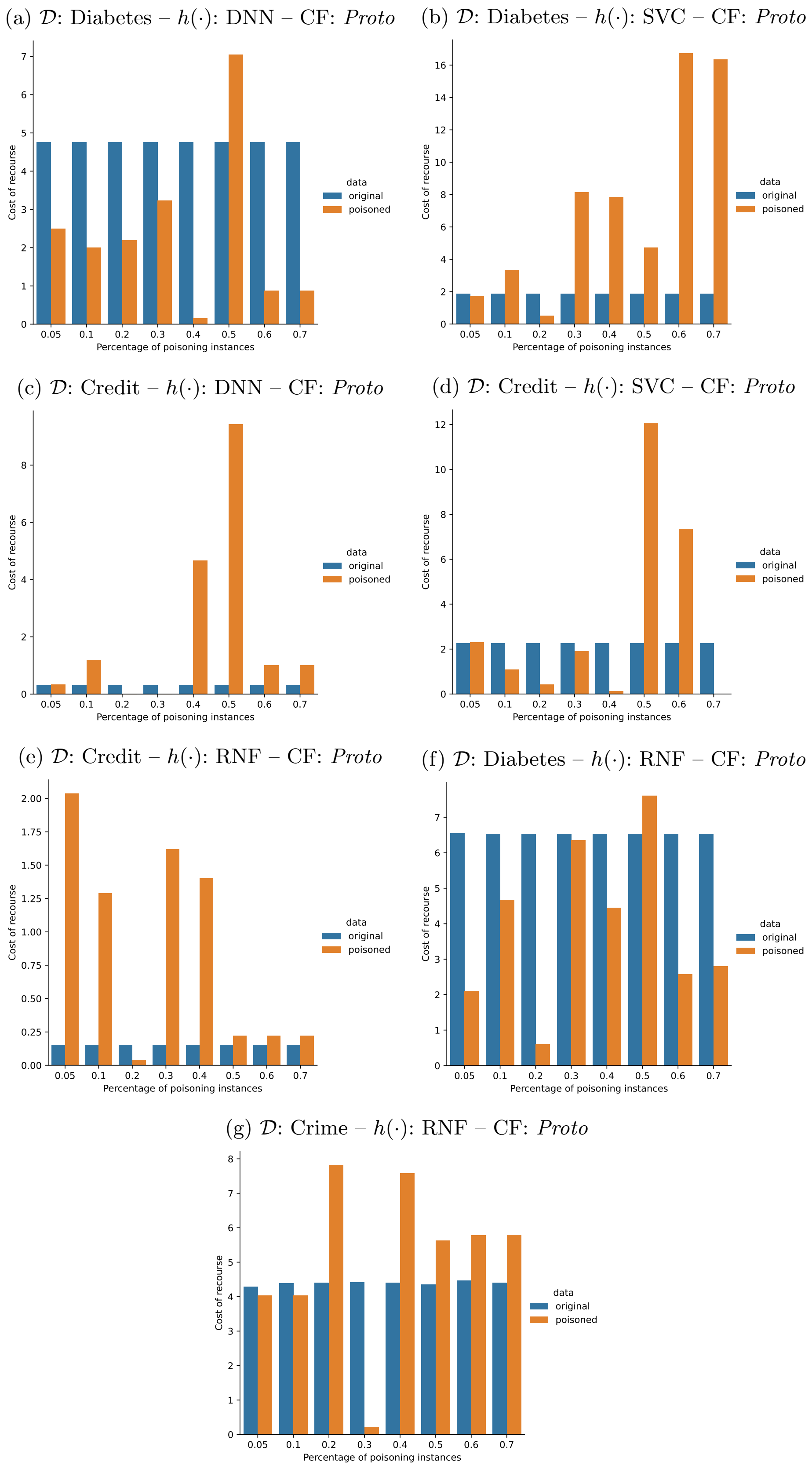
(d) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *Nearest*

(e) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *Nearest*

(f) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *Nearest*

(g) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *Proto*

(h) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *Proto*



(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *Proto*

(b) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *Proto*

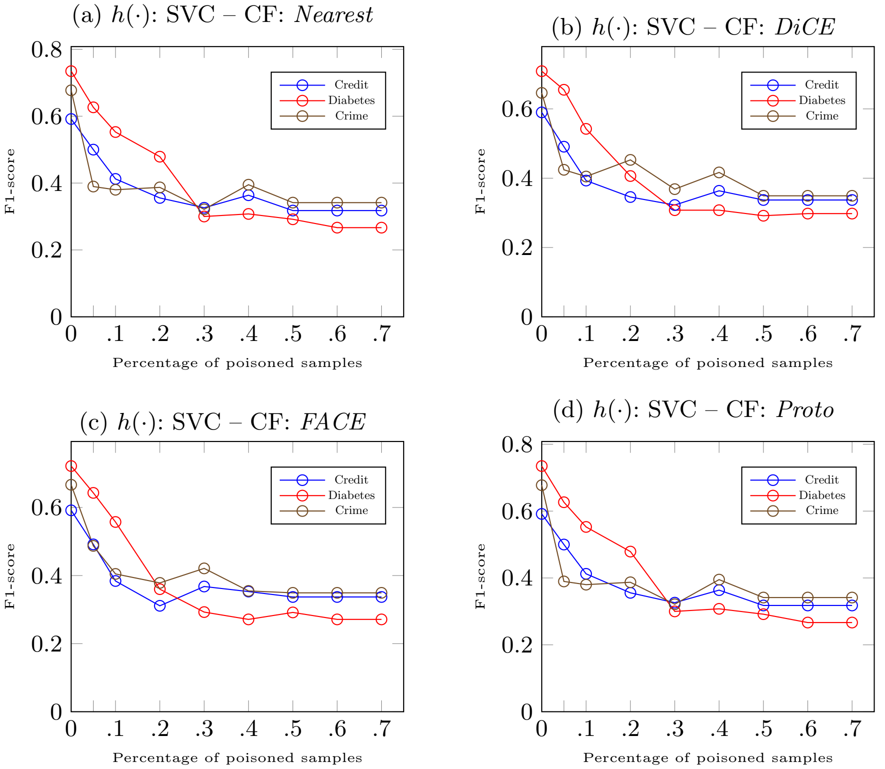
(c) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *Proto*

(d) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *Proto*

(e) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *Proto*

(f) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *Proto*

(g) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *Proto*



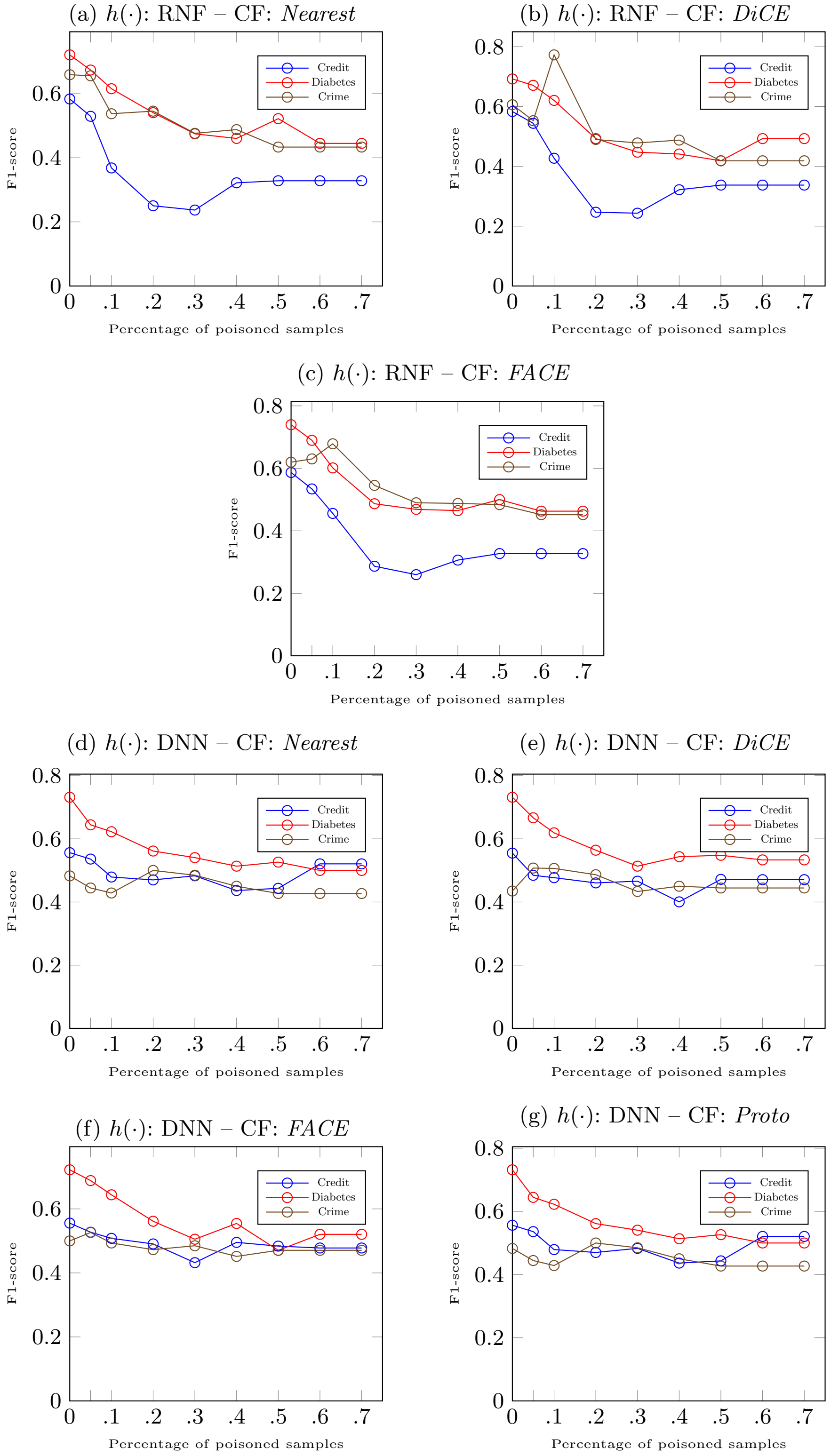
(a) $h(\cdot)$: SVC – CF: *Nearest*

(b) $h(\cdot)$: SVC – CF: *DiCE*

(c) $h(\cdot)$: SVC – CF: *FACE*

(d) $h(\cdot)$: SVC – CF: *Proto*

Fig. 7: Sub-group data poisoning attack: Median (over all folds) F1-score of the classifier for different percentages of poisoned samples (0% to 70%).



(a) $h(\cdot)$: RNF – CF: *Nearest*

(b) $h(\cdot)$: RNF – CF: *DiCE*

(c) $h(\cdot)$: RNF – CF: *FACE*

(d) $h(\cdot)$: DNN – CF: *Nearest*

(e) $h(\cdot)$: DNN – CF: *DiCE*

(f) $h(\cdot)$: DNN – CF: *FACE*

(g) $h(\cdot)$: DNN – CF: *Proto*

Fig. 8: Sub-group data poisoning attack: Median (over all folds) F1-score of the classifier for different percentages of poisoned instances (0% to 70%).

2.4 Global Poisoning Attack

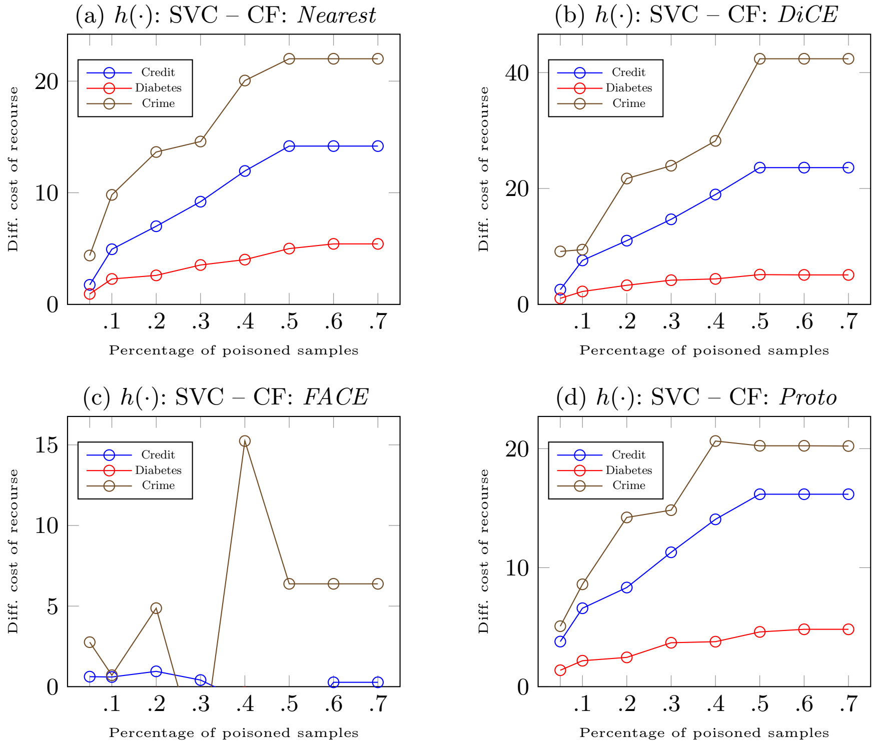
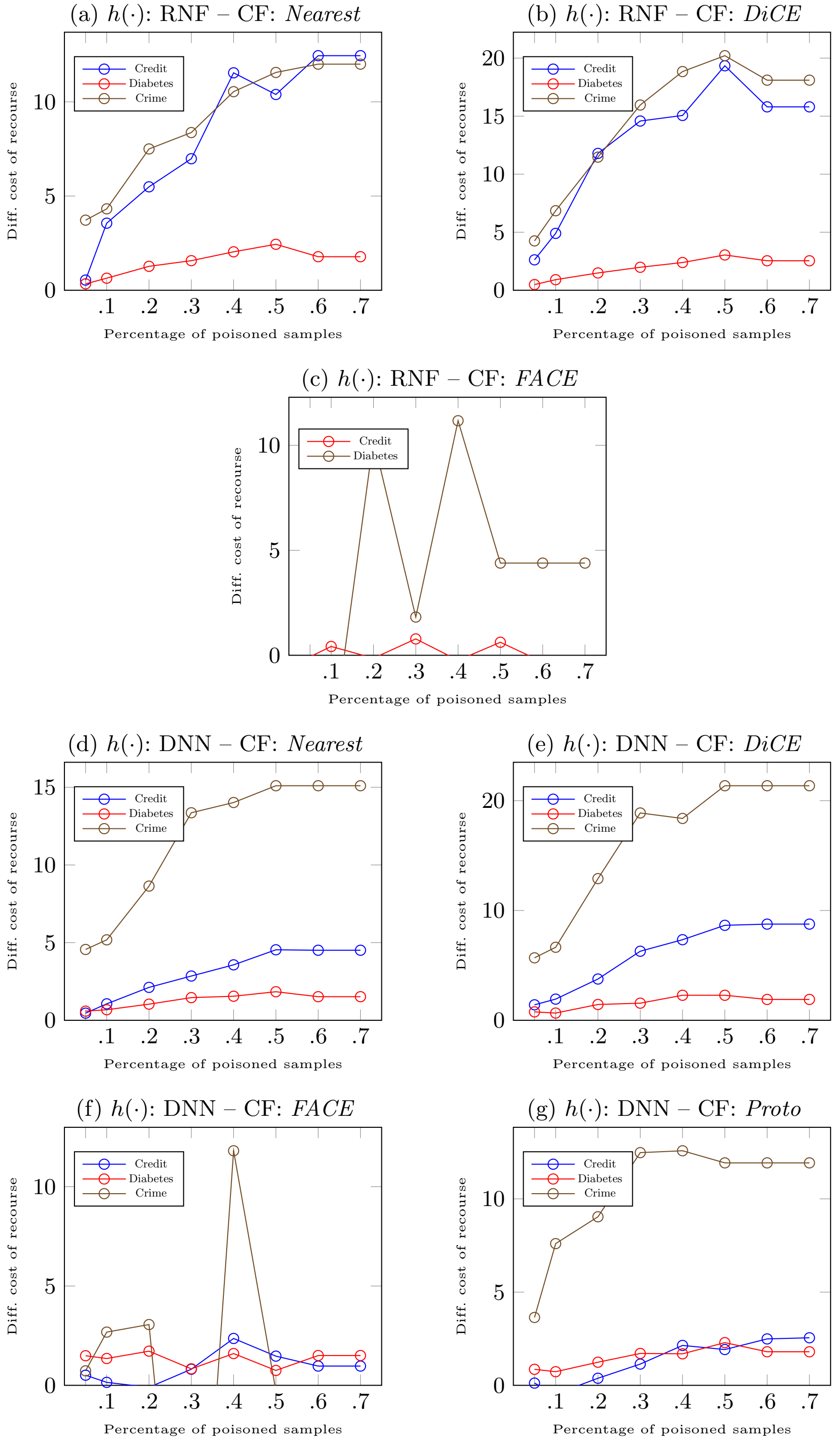


Fig. 9: Global data poisoning attack: Median (over all folds) difference in the cost of recourse vs. percentage of poisoned instances (5% to 70%).


(a) $h(\cdot)$: RNF – CF: *Nearest*

(b) $h(\cdot)$: RNF – CF: *DiCE*

(c) $h(\cdot)$: RNF – CF: *FACE*

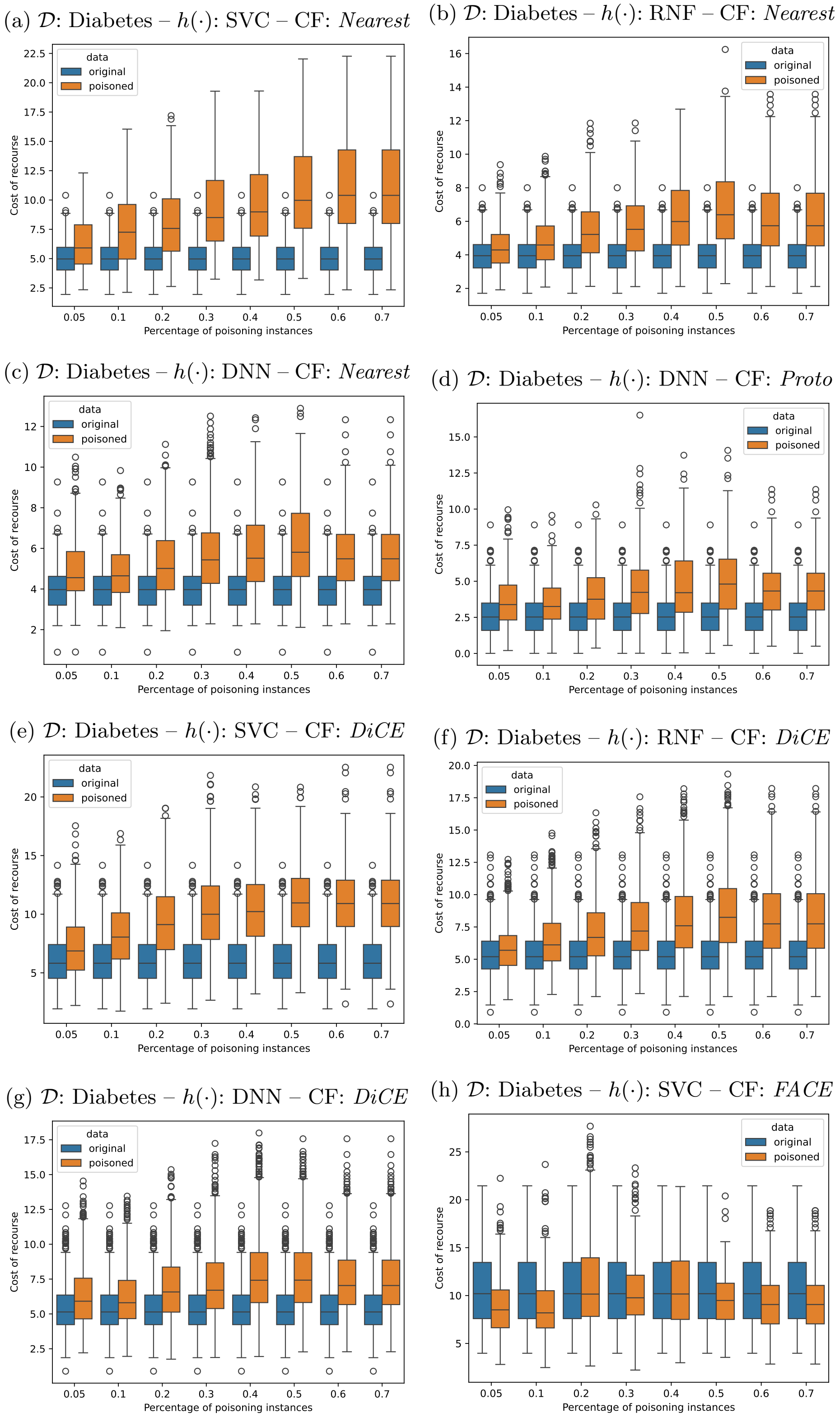
(d) $h(\cdot)$: DNN – CF: *Nearest*

(e) $h(\cdot)$: DNN – CF: *DiCE*

(f) $h(\cdot)$: DNN – CF: *FACE*

(g) $h(\cdot)$: DNN – CF: *Proto*

Fig. 10: Global data poisoning attack: Median (over all folds) difference in the cost of recourse vs. percentage of poisoned instances (5% to 70%).



(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *Nearest*

(b) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *Nearest*

(c) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *Nearest*

(d) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *Proto*

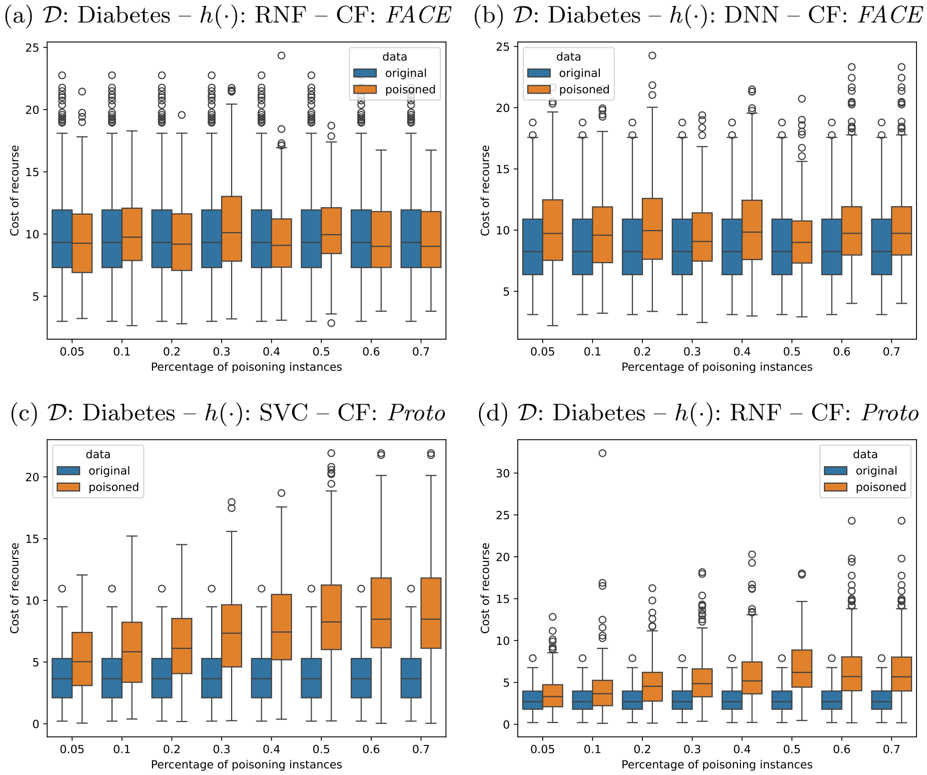
(e) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *DiCE*

(f) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *DiCE*

(g) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *DiCE*

(h) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *FACE*

Fig. 11: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



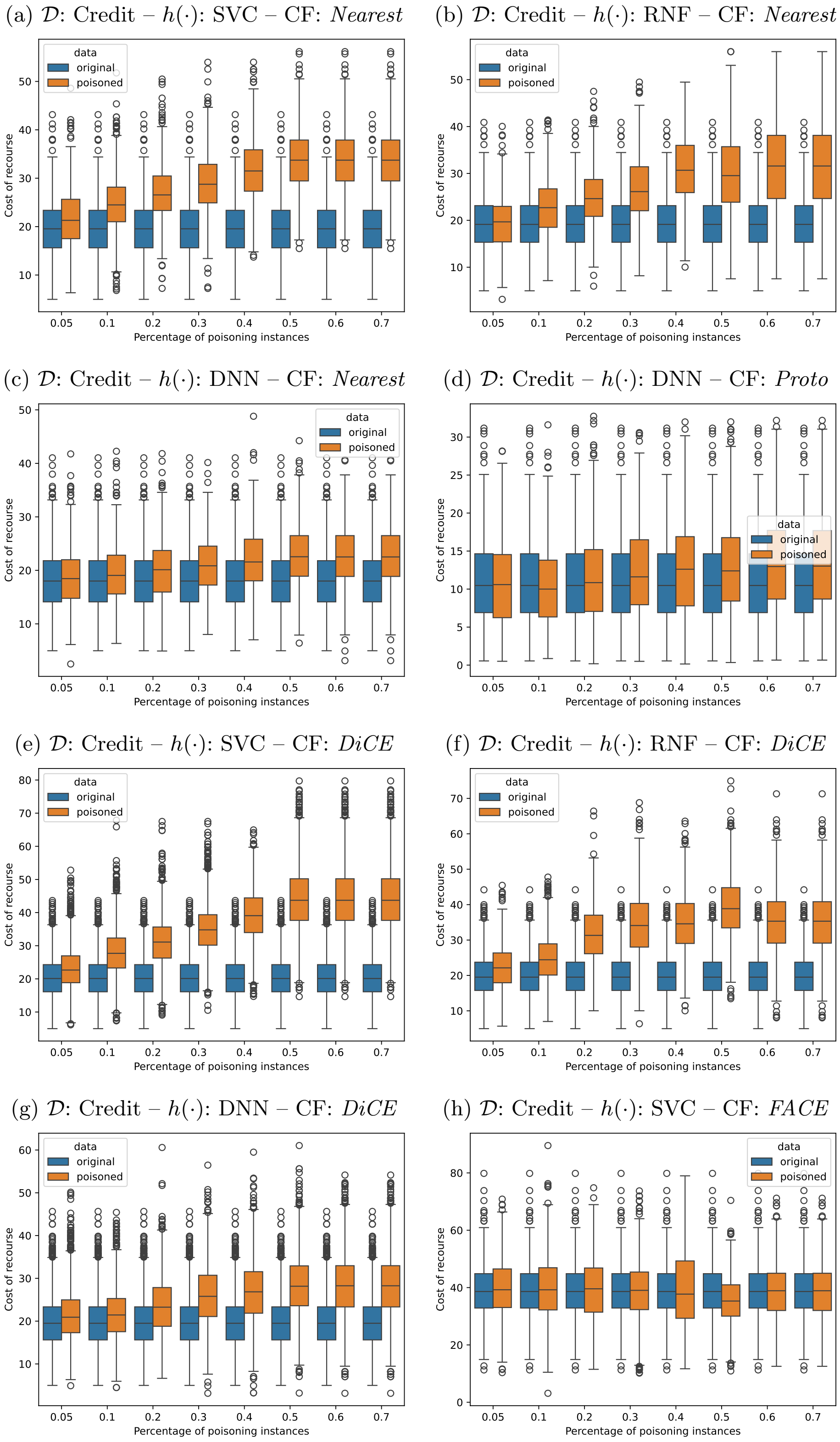
(a) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *FACE*

(b) $\mathcal{D}$: Diabetes – $h(\cdot)$: DNN – CF: *FACE*

(c) $\mathcal{D}$: Diabetes – $h(\cdot)$: SVC – CF: *Proto*

(d) $\mathcal{D}$: Diabetes – $h(\cdot)$: RNF – CF: *Proto*

Fig. 12: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



(a) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *Nearest*

(b) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *Nearest*

(c) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *Nearest*

(d) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *Proto*

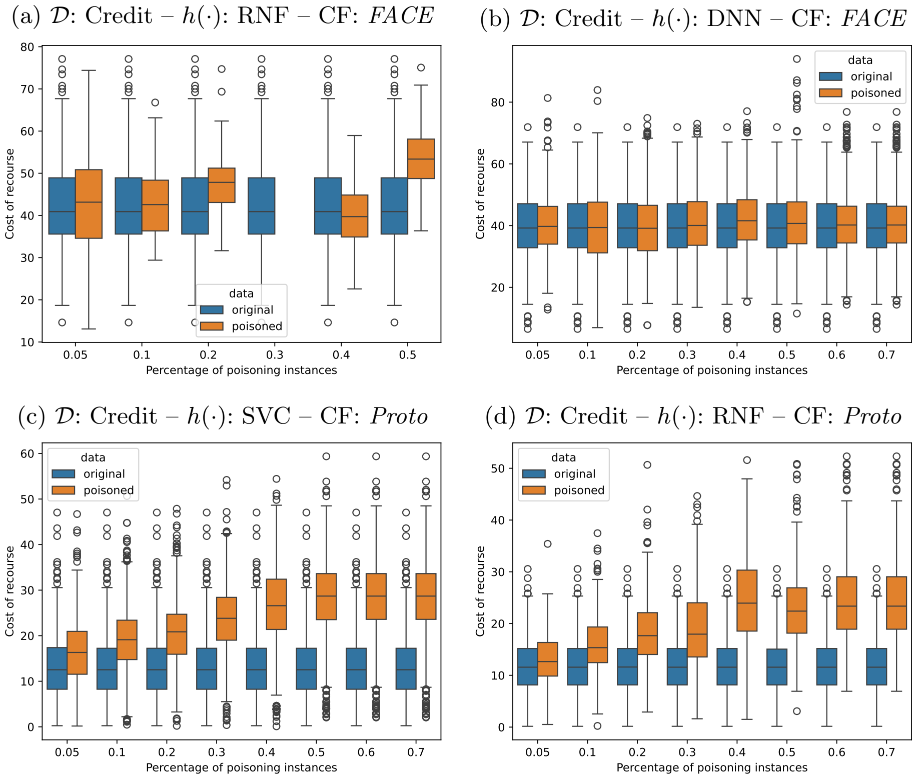
(e) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *DiCE*

(f) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *DiCE*

(g) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *DiCE*

(h) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *FACE*

Fig. 13: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



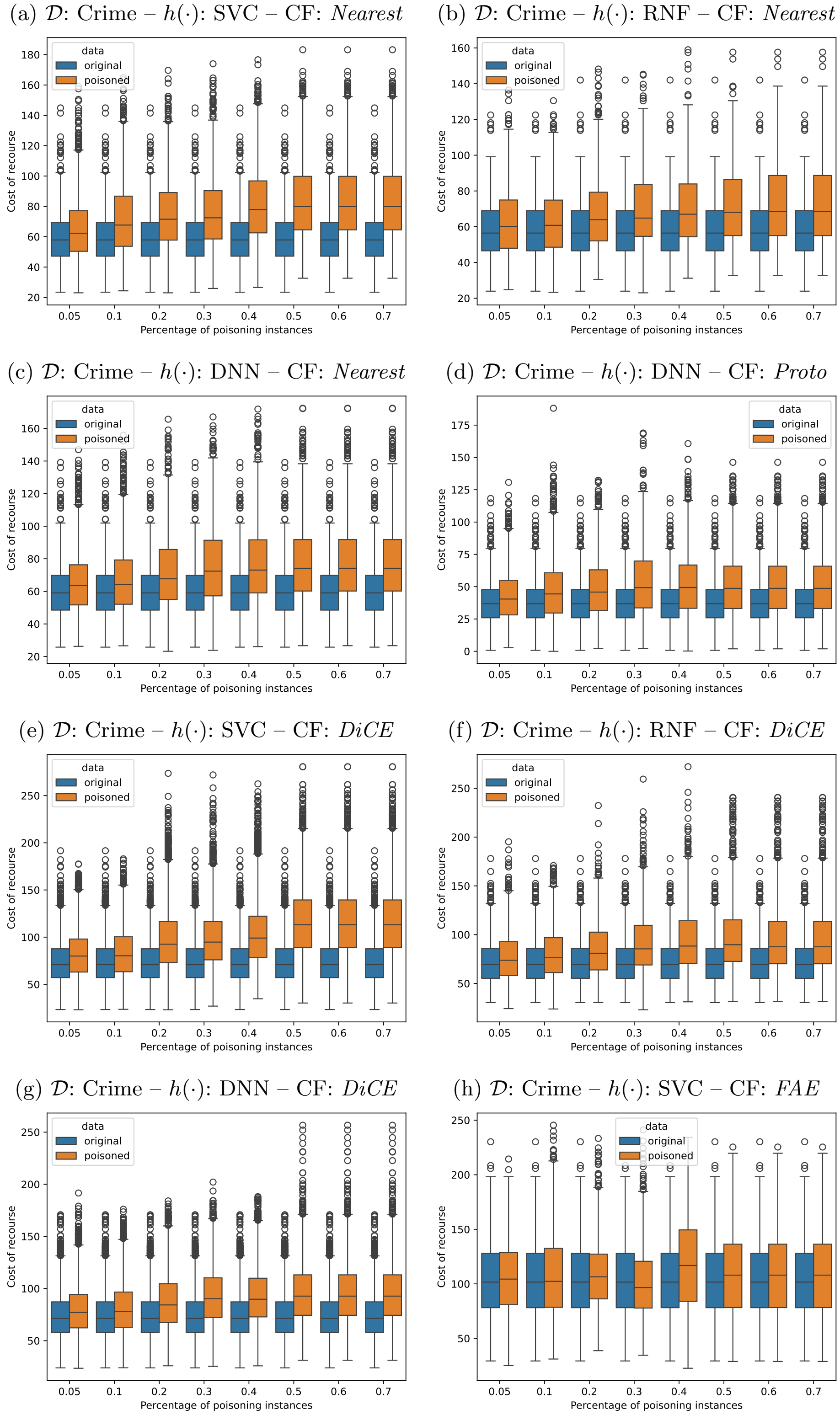
(a) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *FACE*

(b) $\mathcal{D}$: Credit – $h(\cdot)$: DNN – CF: *FACE*

(c) $\mathcal{D}$: Credit – $h(\cdot)$: SVC – CF: *Proto*

(d) $\mathcal{D}$: Credit – $h(\cdot)$: RNF – CF: *Proto*

Fig. 14: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



(a) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *Nearest*

(b) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *Nearest*

(c) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *Nearest*

(d) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *Proto*

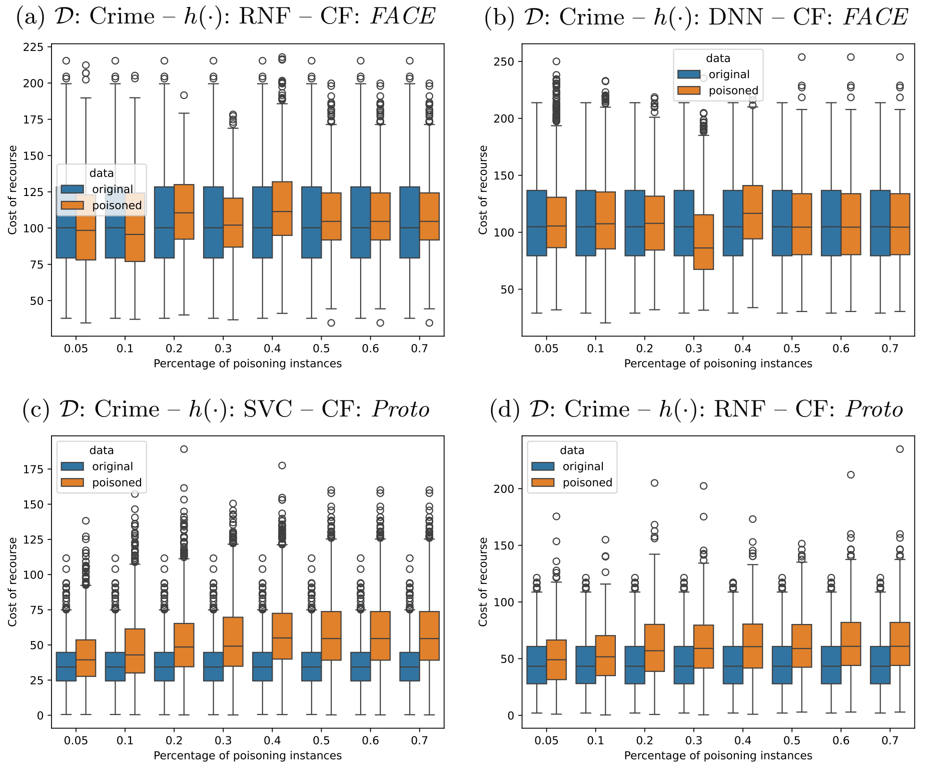
(e) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *DiCE*

(f) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *DiCE*

(g) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *DiCE*

(h) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *FAE*

Fig. 15: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



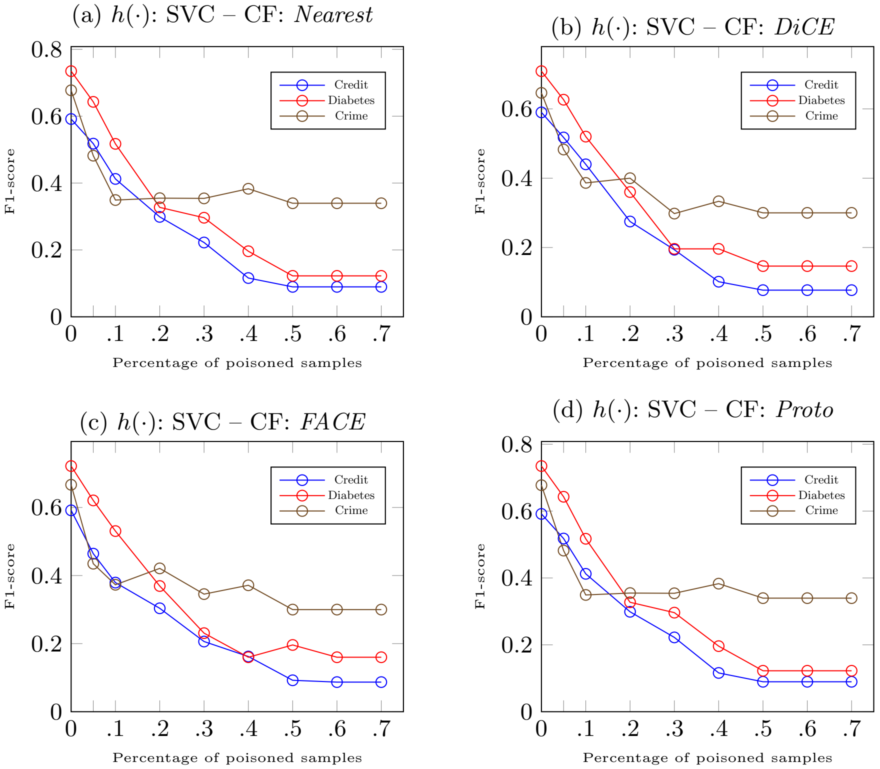
(a) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *FACE*

(b) $\mathcal{D}$: Crime – $h(\cdot)$: DNN – CF: *FACE*

(c) $\mathcal{D}$: Crime – $h(\cdot)$: SVC – CF: *Proto*

(d) $\mathcal{D}$: Crime – $h(\cdot)$: RNF – CF: *Proto*

Fig. 16: Cost of recourse (over all folds) of original data vs. poisoned data – (5% to 70% of poisoned instances).



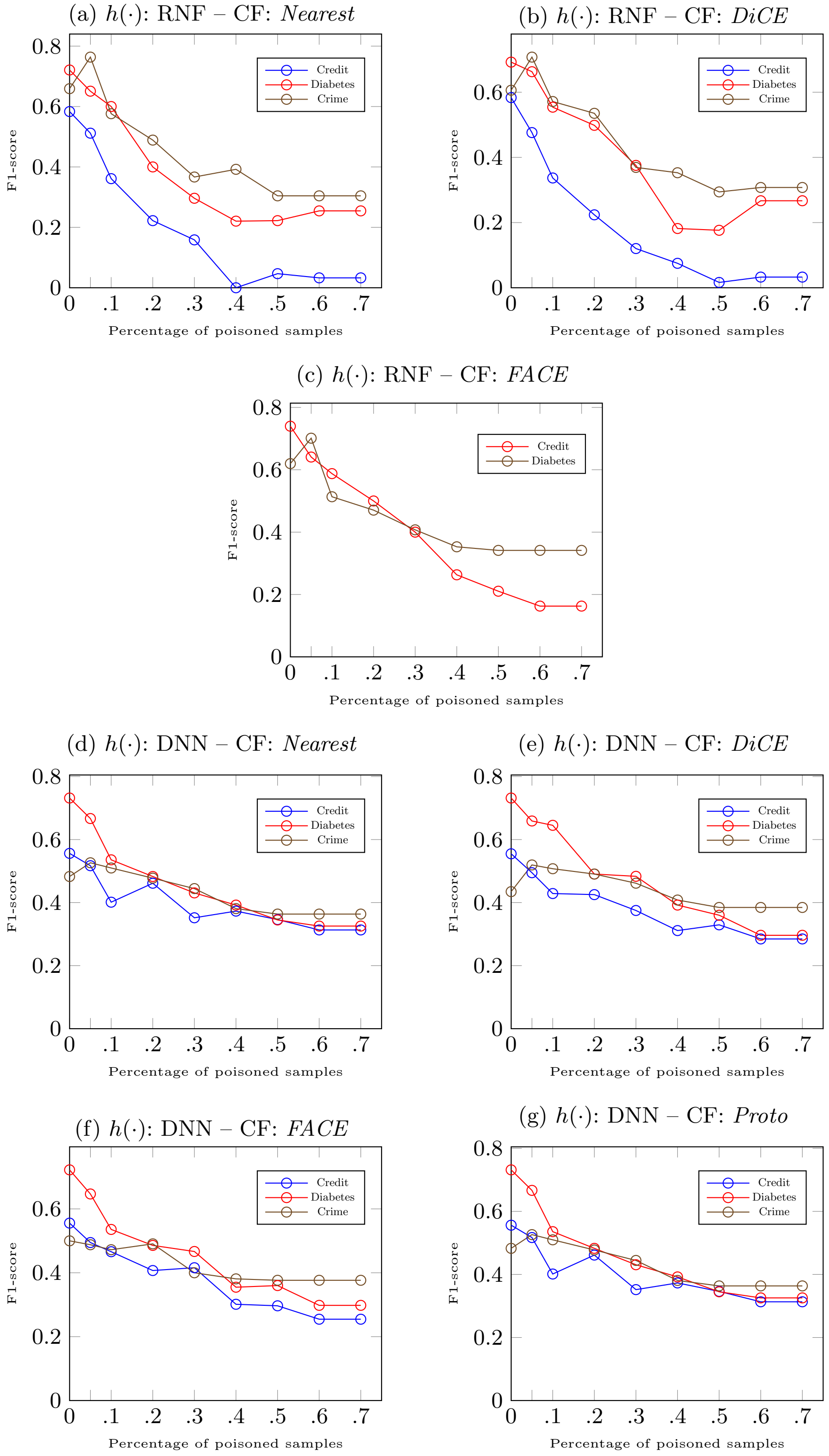
(a) $h(\cdot)$: SVC – CF: *Nearest*

(b) $h(\cdot)$: SVC – CF: *DiCE*

(c) $h(\cdot)$: SVC – CF: *FACE*

(d) $h(\cdot)$: SVC – CF: *Proto*

Fig. 17: Global data poisoning attack: Median (over all folds) F1-score of the classifier for different percentages of poisoned samples (0% to 70%).



(a) $h(\cdot)$: RNF – CF: *Nearest*

(b) $h(\cdot)$: RNF – CF: *DiCE*

(c) $h(\cdot)$: RNF – CF: *FACE*

(d) $h(\cdot)$: DNN – CF: *Nearest*

(e) $h(\cdot)$: DNN – CF: *DiCE*

(f) $h(\cdot)$: DNN – CF: *FACE*

(g) $h(\cdot)$: DNN – CF: *Proto*

Fig. 18: Global data poisoning attack: Median (over all folds) F1-score of the classifier for different percentages of poisoned samples (0% to 70%).