APPENDIX A HARDWARE

All experiments were conducted on a high-performance workstation running Ubuntu 22.04.5 LTS, equipped with an Intel® Core™ i9-10980XE CPU @ 3.00GHz (18 cores, 36 threads), 251 GB of memory, and two NVIDIA GPUs: a Quadro RTX 6000 for compute-intensive tasks and a GeForce GT 1030 for display. The system was configured with NVIDIA driver version 535.247.01.

APPENDIX B HYPERPARAMETER OPTIMIZATION

a: Black box classifier

The hyperparameter is determined by doing a cross-validation on the training set with 5 folds. The best hyperparameters are selected based on the average accuracy across the folds. The hyperparameter search is done using a random search approach, where some combinations of the specified options are evaluated. While not of great importance, the hyperparameter search is done to ensure that the models are not overfitting or underfitting the data. The hyperparameters and their possible values, specific for each black box model, can be found in Table 5.

Model	Hyperparameter	Values
MLP	hidden neurons	32, 64, (32, 32), (64, 32)
	learning rate	0.001, 0.01, 0.1
	max epochs	150, 300, 500
LipMLP	hidden neurons	32, 64, (32, 32), (64, 32)
	lipschitz lambda	0.0001, 0.001, 0.01
	learning rate	0.001, 0.01, 0.1
	max epochs	150, 300, 500
RF	num estimators	50, 100, 150
	max depth	6, 10, 30, None
	min samples split	2, 5, 10
	min samples leaf	1, 2, 4
XGBoost	num estimators	50, 100, 150
	max depth	3, 6, 9, None
	learning rate	0.01, 0.05, 0.1
	min child weight	1, 3, 5

TABLE 5: Hyperparameter configurations for different models, including MLP, LipMLP, RF, and XGBoost. The choices for each hyperparameter are specified.

In the Table 6, there are the best hyperparameters selected for each model and dataset. The hyperparameters are chosen based on the average accuracy across the folds.

Across all datasets, LipMLP consistently converges with a low learning rate (0.001) and compact architectures, suggesting that the Lipschitz constraint already acts as a strong regularizer. In contrast, MLP requires more varied learning rates and larger architectures to achieve comparable performance.

b: Counterfactual Generators: DiCE and LoRE

We used fixed hyperparameters for the generation of the counterfactuals when using DiCE and LoRE. For DiCE we used the proximity weight equal to 0.5 and as a method `kdtree`. For LoRE we used the `genetic` option for the neighborhood synthesis.

c: Counterfactual Generators: ILS

ILS is optimized using a hyperparameter search using a stratified 3-fold random search, looking for 20 different random combinations within the hyperparameter space composed of:

- latent space dimensions: 2, 3, 4,
- batch size: 4, 8, 16, 32, 64, 128,
- learning rate: 0.0001, 0.001, 0.005, 0.008, 0.01
- sigma: 0.5, 1.0, 2.0

for a maximum of 2000 epochs with an early stop of 50 for the **German**, **Two-Moons**, and **Wisconsin** datasets, while 70 for **Adult**. What is optimized during the hyperparameter search is the KL-loss function of ILS, on the calibration set. We found that the ILS' best hyperparameters do not vary when varying the black-box model, but they remain the same across the dataset. For the **German** we found the best latent dimension to be 4, the best batch size to be 4, the learning rate equal to 0.0001, and σ=0.5; for the **Adult** dataset and **Wisconsin**, the best hyperparameters remained the same. ILS represent the space $(\mathcal{X}, \mathcal{Y})$ in an analogous way, performing best when the number of dimensions is higher, while approaching the solution very slowly, having the lowest possible learning rate and the lowest batch size as well.

We do report that a different combination was obtained for the **Two-Moons**, which was indeed different: latent dimension=3, batch size=16, learning rate=0.001, and σ=1.0.

d: Counterfactual Generators: GS

GS [42] generates counterfactuals by sampling candidate points on hyperspheres of increasing radius centered on the input instance, progressively expanding the search until a point with a different predicted class is found. The closest such point is then returned as the counterfactual. In our experiments we used 200 candidates per layer, an initial radius of 0.05, a radius decrease factor of 2.0, and sparse perturbations enabled.

Looking at the execution times in Table 7, we can see that DiCE is the slowest method, especially on larger datasets like **Adult**. Other methods like LoRE and ILS have more reasonable execution times, while GS is generally the fastest across all models and datasets. This is due to the fact that DiCE relies on a more complex optimization process to generate counterfactuals, while GS uses a more efficient search strategy. The execution times also vary significantly depending on the black box model used, with RF and XGBoost generally taking longer than MLP and LipMLP, likely due to their more complex structures (e.g., ensemble methods, tree-based models, as opposed to neural networks). Overall, these results highlight the importance of considering execution time when choosing a counterfactual generation method, especially for larger datasets and more complex models.

APPENDIX C DATASET-SPECIFIC CRITICAL DIFFERENCE DIAGRAMS

Here we report the dataset-specific critical difference diagrams with respect to the two performance metrics employed

Model	Hyperparameters	German	Adult	Wisconsin	Two-Moons
MLP	hidden neurons	(64)	(32)	(64)	(32)
	learning rate	0.001	0.01	0.01	0.1
	max epochs	150	300	300	500
LipMLP	hidden neurons	(32)	(64, 32)	(32)	(32)
	lipschitz lambda	0.0001	0.0001	0.001	0.001
	learning rate	0.001	0.001	0.001	0.001
	max epochs	150	150	300	300
RF	num estimators	50	150	50	50
	max depth	10	30	10	10
	min samples split	2	5	5	5
	min samples leaf	1	4	1	1
XGBoost	num estimators	150	150	150	50
	max depth	6	6	9	None
	learning rate	0.1	0.1	0.1	0.01
	min child weight	1	1	3	3

TABLE 6: Best hyperparameters identified for each model and dataset through our optimization process.

	MLP				LipMLP				RF				XGBoost			
	DiCE	LoRE	ILS	GS	DiCE	LoRE	ILS	GS	DiCE	LoRE	ILS	GS	DiCE	LoRE	ILS	GS
German	1.83h	42.87m	22.64m	8.59m	10.38h	31.67m	10.67h	8.74m	6.27h	49.5m	4.1h	57.3m	1.35h	3.54h	59.5m	1.73h
Adult	45.97h*	1.46h	5.42h	1.54m	61.82h	1.01h	4.29h	6.53m	196.02h*	1.73h	9.35h	2.77h	126.35h*	1.72h	8.3h	4.09h
Wisconsin	14.84h	26.01m	20.84m	2.99m	43.15h	20.54m	1.79h	9.76m	30.12h	28.18m	18.17h	18.83m	24.34h	30.4m	22.95h	23.44m
Two-Moons	3.89s	38.73m	16.82s	1.28m	7.7h	25.15m	16.13s	3.3m	12.07s	38.7m	19.93s	18.29m	4.67s	36.78m	7.99s	10.27m

TABLE 7: Execution times for each combination of black box model and counterfactual generation method across all datasets. Times marked with * indicate runs using a subsampled version of the dataset.

to evaluate rejection quality, namely Non-Rejected Accuracy (*NA*) and Classification Quality (*CQ*).

Figure 6c shows results of *NA* for **Adult**: PlugInRule leads with rank 3.6, followed by a strong group of ILS_l methods. In particular, ILS_l with *Cosine* distance and *avg* aggregation policy ranks second (4.1) with no statistical difference with the baseline. PlugInRuleAUC performs worst overall (rank 12), significantly below all other approaches. With respect to *CQ* (see Figure 6d), a large clique of top-ranking methods includes the baseline PlugInRule (5.5) together with GS paired with $L2$ (5.5) and *Cosine* (5.6), but also ILS_l similarly paired with *Cosine* (5.7) and $L2$ (5.8) and also LoRE with the *Canberra* distance (5.9).

The **German** dataset results in Figure 6a show a different picture: GS with *Centered* L2 distance leads and all aggregation policies (ranks 4.8, 5.4 and 5.6), followed by PlugInRule (rank 5.6) and further GS variants using *Cosine* and $L2$ distances. PlugInRuleAUC ranks last (rank 8.8). The connecting lines indicate that methods within the top group are not significantly different from each other, though differences between the best and worst methods are highly significant (min pairwise $p = 1.043 \times 10^{-8}$). A very similar configuration results when analysing the critical difference diagram of *CQ* (see Figure 6b), where the top methods, together with PlugInRule (rank 5) are different configurations of GS with *Contered* $L2$ (5.6 with *min*, 5.9 with *max*, 6.3 with *avg*) and *Cosine* (5.9 with *max*, 6 with *min*, 6.1 with *avg*)

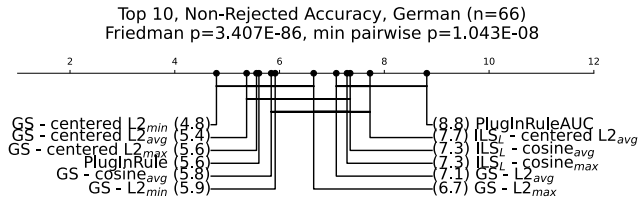
The **Wisconsin** dataset results for *NA* (see Figure 6e) are dominated by GS across a wide variety of statistically indistinguishable distance functions, including *Canberra* (ranks 5.2 with *min* and 5.3 with *avg*), *Cosine* (ranks 5.7 with *max*, 5.9 with *min* and 6 with *avg*), and *Centered* $L2$ (ranks 5.8 with *avg*, 5.9 with *min*). A comparable performance is also achieved by LoRE with *Cosine* (rank 5.6). PlugInRule falls in the lower tier (rank 7.3), while PlugInRuleAUC is again the worst-performing method (rank 11), significantly separated from all others. Looking at the *CQ* performance, we observe very different rankings compared to *NA*: the top tier is unexpectedly occupied by PlugInRuleAUC (3.7) together with ILS_l paired with *Centered* $L2$ (ranks 5.7 with *max*, 6.2 with *avg*). ILS also shows comparable performances with *Centered* $L2$ (ranks 6.1 with *min*, 6.8 with *avg*). In this case, the baseline PlugInRule lied much behind.

Finally, the **Two-Moons** results for *NA* in Figure 6g reveal that LoRE with $L2$ distance and *min* aggregation policy achieves the best rank (4), followed by GS variants with $L2$ (ranks 4.2 with *min*, 4.3 with *avg* and 4.4 with *max*) and *Canberra* (ranks 7.2 with *max*, 7.3 with *min* and 7.5 with *avg*) distances. PlugInRule ranks in the middle of the diagram (4.6). The top methods form a single connected clique, indicating no statistically significant differences among them. With respect to the *CQ* performance, Figure 6h shows a top tier of `SC-CE` configurations with the same rank of 4.5, namely: GS, ILS and LoRE, all with the *Centered* $L2$ distance. Both baselines exhibit worse performance.

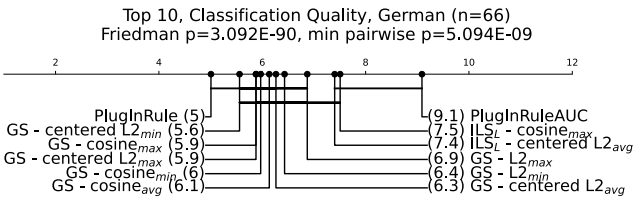
APPENDIX D COUNTERFACTUAL PROPERTIES

We evaluate the generators across six complementary properties to characterize effectiveness, plausibility, and variability of the produced recourse.

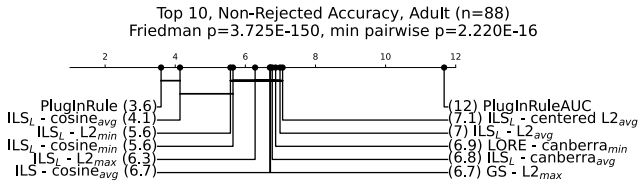
Fidelity. Fidelity is generally high, with GS achieving 1.00 in all reported configurations and ILS remaining near-perfect except on **Two-Moons** (RF: 0.84, XGBoost: 0.69). In contrast, LoRE shows clear drops on non-linear settings (e.g., **Two-Moons**: 0.47-0.48 for RF/MLP), while DiCE exhibits isolated failures (e.g., 0.79 on **German** with XGBoost).



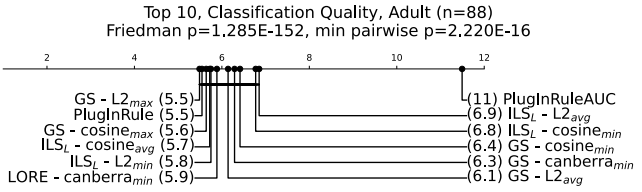
(a) **German**- *NA*



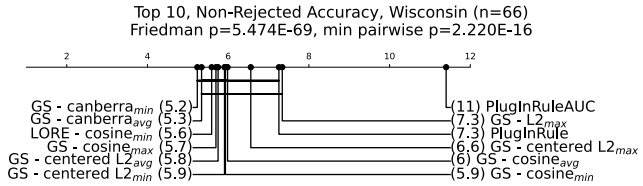
(b) **German**- *CQ*



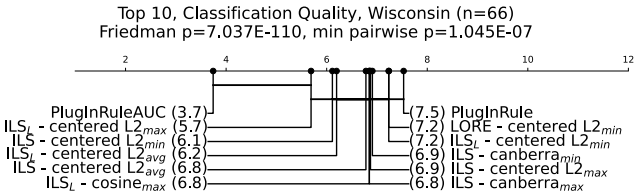
(c) **Adult**- *NA*



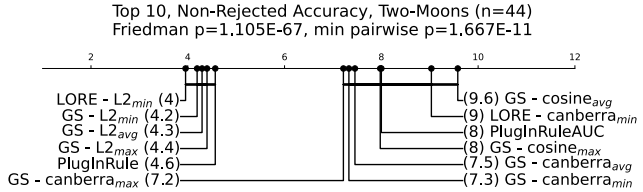
(d) **Adult**- *CQ*



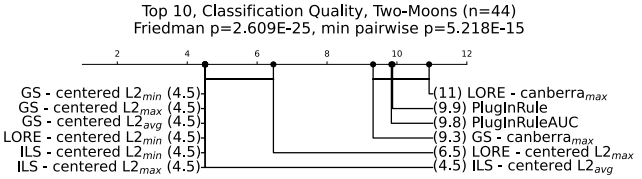
(e) **Wisconsin**- *NA*



(f) **Wisconsin**- *CQ*



(g) **Two-Moons**- *NA*



(h) **Two-Moons**- *CQ*

FIGURE 6: Critical difference diagram showing the best-performing rejection policies in terms of Non-Rejected Accuracy (*NA*) and Classification Quality (*CQ*), divided by dataset.

TABLE 8: Counterfactual Quality Metrics: Fidelity

Dataset	Model	DiCE	LoRE	ILS	GS
German	LipMLP	1.00	1.00	1.00	1.00
	MLP	1.00	1.00	1.00	1.00
	RF	1.00	1.00	1.00	1.00
	XGBoost	0.79	1.00	1.00	1.00
Adult	LipMLP	N/A	1.00	1.00	1.00
	MLP	1.00	1.00	1.00	1.00
	RF	1.00	1.00	1.00	1.00
	XGBoost	0.82	1.00	1.00	1.00
Wisconsin	LipMLP	1.00	0.92	1.00	1.00
	MLP	1.00	0.96	1.00	1.00
	RF	1.00	0.83	1.00	1.00
	XGBoost	0.93	0.88	1.00	1.00
Two-Moons	LipMLP	1.00	1.00	0.99	1.00
	MLP	1.00	0.48	1.00	1.00
	RF	1.00	0.47	0.84	1.00
	XGBoost	1.00	1.00	0.69	1.00

Discriminative Power. On **Adult**, **Wisconsin**, and **Two-Moons**, DiCE and GS typically provide the strongest separation (often around 0.75-0.90), while LoRE is competitive but generally lower. ILS is consistently weaker on these datasets (mostly 0.49-0.64), although it remains competitive on **German** (up to 0.63).

Proximity. GS yields the smallest perturbations almost everywhere (typically 0.13-0.84), indicating highly local counterfactual moves. LoRE is intermediate (about 0.66-2.54), whereas ILS often requires much larger shifts (up to 74.94), suggesting lower proximity.

Sparsity. ILS is usually the sparsest method (often around 0.78-5.99 changed features), followed by LoRE, while GS is moderate and dataset-dependent. DiCE is generally the least sparse, with very high values on **Wisconsin** (29.99 across models).

DPP Diversity. Diversity is quantified via the Determinantal Point Process (DPP) volume of the generated counterfactual set [48], which measures the geometric spread of the set in feature space; higher values indicate more varied recourse options. Importantly, this quantity scales with both the number of generated counterfactuals and the dimensionality of the feature space, so comparisons across generators or datasets must account for these structural differences. GS

TABLE 9: Counterfactual Quality Metrics: Discriminative Power

Dataset	Model	DiCE	LoRE	ILS	GS
German	LipMLP	0.58	0.53	0.59	0.53
	MLP	0.60	0.53	0.62	0.53
	RF	0.58	0.54	0.57	0.55
	XGBoost	0.61	0.51	0.57	0.52
Adult	LipMLP	N/A	0.74	0.64	0.74
	MLP	0.77	0.73	0.56	0.69
	RF	0.76	0.73	0.57	0.76
	XGBoost	0.78	0.69	0.53	0.75
Wisconsin	LipMLP	0.82	0.78	0.51	0.83
	MLP	0.82	0.75	0.49	0.80
	RF	0.86	0.75	0.53	0.89
	XGBoost	0.84	0.79	0.51	0.83
Two-Moons	LipMLP	0.90	0.91	0.74	0.89
	MLP	0.90	0.80	0.50	0.89
	RF	0.80	0.78	0.52	0.81
	XGBoost	0.90	0.90	0.55	0.88

TABLE 10: Counterfactual Quality Metrics: L2 Distance

Dataset	Model	DiCE	LoRE	ILS	GS
German	LipMLP	5.26	2.54	48.41	0.82
	MLP	4.53	2.53	15.49	0.84
	RF	4.39	2.39	20.12	0.81
	XGBoost	5.61	1.26	18.51	0.13
Adult	LipMLP	N/A	2.40	16.90	0.30
	MLP	0.65	2.51	43.11	0.18
	RF	0.65	2.36	74.94	0.46
	XGBoost	0.78	2.51	12.94	0.39
Wisconsin	LipMLP	0.87	0.68	1.71	0.35
	MLP	0.74	0.71	1.84	0.33
	RF	0.77	0.66	1.55	0.37
	XGBoost	0.77	0.67	3.25	0.25
Two-Moons	LipMLP	1.06	1.46	1.44	0.64
	MLP	0.83	1.80	13.65	0.63
	RF	0.72	2.07	3.06	0.53
	XGBoost	0.84	1.76	0.89	0.61

TABLE 11: Counterfactual Quality Metrics: L0 Sparsity

Dataset	Model	DiCE	LoRE	ILS	GS
German	LipMLP	12.50	6.74	2.31	7.08
	MLP	11.80	6.79	2.08	6.59
	RF	11.61	6.16	4.25	3.19
	XGBoost	12.91	3.68	3.69	2.53
Adult	LipMLP	N/A	6.73	2.05	4.82
	MLP	11.61	6.90	2.08	3.79
	RF	11.61	6.91	3.29	3.28
	XGBoost	11.70	6.94	2.26	2.87
Wisconsin	LipMLP	29.99	7.46	2.21	9.20
	MLP	29.99	7.38	1.64	9.42
	RF	29.99	7.65	5.75	6.51
	XGBoost	29.99	8.36	5.99	4.76
Two-Moons	LipMLP	2.00	1.93	0.99	1.79
	MLP	2.00	2.00	1.00	1.80
	RF	2.00	2.00	0.86	1.71
	XGBoost	2.00	1.98	0.78	1.34

scores 0.00 everywhere because it produces only a single counterfactual per instance, for which the DPP volume is trivially zero. LoRE also scores 0.00 consistently: despite generating many candidates, they tend to cluster tightly near the decision boundary, collapsing the DPP volume. ILS attains the highest diversity overall, particularly on **Adult** (0.14-0.74) and **German** (0.03-0.48): because its counterfactuals are generated in a lower-dimensional latent space and then projected back to the original feature space, the decoding step naturally spreads them across a broader region, boosting diversity. Its near-zero scores on **Two-Moons** are explained by the low dimensionality of that dataset (two features), which structurally limits the achievable DPP volume regardless of the generator. DiCE achieves moderate diversity on **German** (0.01-0.62) but near-zero values on the remaining datasets (0.001-0.003), with no valid counterfactuals for LipMLP on **Adult**. Full results are reported in Table 12.

TABLE 12: Counterfactual Quality Metrics: DPP Diversity

Dataset	Model	DiCE	LoRE	ILS	GS
German	LipMLP	0.01	0.00	0.48	0.00
	MLP	0.52	0.00	0.21	0.00
	RF	0.53	0.00	0.24	0.00
	XGBoost	0.62	0.00	0.19	0.00
	LGBM	0.52	0.00	0.03	0.00
Adult	LipMLP	N/A	0.00	0.32	0.00
	MLP	0.00	0.00	0.14	0.00
	RF	0.00	0.00	0.68	0.00
	XGBoost	0.00	0.00	0.49	0.00
	LGBM	0.00	0.00	0.74	0.00
Wisconsin	LipMLP	0.00	0.00	0.00	0.00
	MLP	0.00	0.00	0.27	0.00
	RF	0.00	0.00	0.10	0.00
	XGBoost	0.00	0.00	0.40	0.00
	LGBM	0.00	0.00	0.20	0.00
Two-Moons	LipMLP	0.00	0.00	0.14	0.00
	MLP	0.00	0.00	0.00	0.00
	RF	0.00	0.00	0.43	0.00
	XGBoost	0.00	0.00	0.00	0.00
	LGBM	0.00	0.00	0.00	0.00

APPENDIX E THEORETICAL BOUNDS ON PREDICTION MARGIN THROUGH INSTANCE-COUNTERFACTUAL DISTANCE

In this section, we provide theoretical guarantees for a broad class of models, namely *continuously differentiable* models. Our analysis relies on *Lipschitz continuity*, a regularity property ensuring that small perturbations in the input induce controlled variations in the prediction score, thereby formalising the intuition that nearby inputs should yield comparable outputs [49], [50]. A number of studies have shown that the Lipschitz constant governs many features of predictive models, including *generalisation* [51], *robustness* [52], *vulnerability to adversarial examples* [53], *consistency* [54] and *robustness of explanations* [55]. This has prompted extensive research on estimating tight bounds for the Lipschitz constant and on the application of explicit Lipschitz controls on predictive models [49], [56].

Definition 1 (Binary classifier)**.** Let $s : \mathcal{X} \to [0,1]$ be a

score function defined on an open set $\mathcal{X} \subset \mathbb{R}^n$ and $t \in [0,1]$ be a decision threshold. A binary classifier $f : \mathcal{X} \to \{0,1\}$ is defined as

$$f(x) := \begin{cases} 1 & \text{if } s(x) \geq t, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Definition 2 (Margin)**.** The margin m of an input $x \in \mathcal{X}$ with respect to a binary classifier f is defined as the signed distance between the score s and the threshold t: $m(x) = s(x) - t$.

A positive margin indicates the prediction of the positive class, while a negative margin indicates the prediction of the negative class. The absolute value $|m(x)|$ can be used to quantify the confidence of the prediction, with smaller values corresponding to inputs closer to the decision boundary, hence with higher uncertainty.

Definition 3 (Decision boundary)**.** The decision boundary $\mathcal{D}$ of a score-based binary classifier f is defined as $\mathcal{D} = \{x \in \mathcal{X} : s(x) = t\}$. In other words, it is the set of the points in $\mathcal{X}$ with a null margin.

Definition 4 (Counterfactual instance)**.** Given an instance $x \in \mathcal{X}$, a classifier f and a distance metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, a counterfactual instance is any instance x' such that $f(x) \neq f(x')$. Moreover, the *closest* counterfactual is defined as an element of the set $X' := \{x' \in \mathcal{X} : f(x') \neq f(x) \wedge x' = \arg\min d(x,x')\}$.

Definition 5 (Continuously differentiable function)**.** A real-valued function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is said to be continuously differentiable (or of class $\mathbb{C}^1(\mathcal{X})$) if all partial derivatives $x \mapsto \frac{\partial f}{\partial x_j}(x)$ exist and are continuous on $\mathcal{X}$ for each $j \in \{1, \ldots, n\}$.

Definition 6 (Lipschitz continuity)**.** Given two metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$, a function $\phi : \mathcal{X} \to \mathcal{Y}$ is called Lipschitz continuous if there exists a real constant $L \geq 0$ such that $\forall x_1, x_2 \in \mathcal{X}$:

$$d_\mathcal{Y}(\phi(x_1), \phi(x_2)) \leq L d_\mathcal{X}(x_1, x_2). \tag{9}$$

L is referred to as a Lipschitz constant for the function ϕ and ϕ may also be referred to as L-Lipschitz.

Note that smaller values of the Lipschitz constant are associated with greater controllability and stability of the model output. A larger Lipschitz constant, by contrast, implies that small variations in the input can induce large changes in the output, reflecting increased sensitivity and reduced robustness, a condition at the basis of adversarial attacks.

In practice, the computation of Lipschitz constant of the function ϕ has been proven to to be NP-hard, hence both theoretical analyses and empirical methods typically focus on estimating the upper and lower bounds of the "true" Lipschitz constant [50]. In particular, a lower bound to the Lipschitz constant based on a finite-sample evaluation (e.g., the training set) is given by the local Lipschitz constant that is obtained by restricting the estimation to the subspace of the domain of ϕ where the underlying data distribution is supported (plus its neighbourhood).

Definition 7 (Local Lipschitz continuity)**.** A function ϕ is called locally Lipschitz continuous if for every $x \in \mathcal{X}$ there exists a neighbourhood U of x such that ϕ restricted to U is L-Lipschitz continuous.

Theorem 1. *Let $f : \mathcal{X} \to \{0,1\}$ be a binary classifier defined on an open $\mathcal{X} \subseteq \mathbb{R}^n$ associated with a score function $s : \mathcal{X} \to [0,1]$ and a threshold t, as defined in Equation 8. Consider an instance x in the interior of $\mathcal{X}$ together with a counterfactual x' under the distance metric $d_\mathcal{X}$ induced by the norm $\|\cdot\|$ on $\mathcal{X}$. Assume s to be **continuously differentiable** on $\mathcal{X}$. Then the prediction margin $m(x) = s(x) - t$ is upper-bounded by the distance to the counterfactual scaled by a real factor locally defined as $L = \sup_{z \in B_\delta(x)} \|\nabla s(z)\| < \infty$:*

$$|m(x)| \leq L \cdot d_\mathcal{X}(x, x').$$

Proof. Let $x \in \mathcal{X}$ and consider a neighborhood $B_\delta(x) = \{x' \in \mathcal{X} : d_\mathcal{X}(x', x) < \delta\}$ of x. Since $\mathcal{X}$ is open, then also the closure $\overline{B_\delta(x)}$ lies inside $\mathcal{X}$. Since $\overline{B_\delta(x)}$ is compact, the gradient $\nabla s(x)$ is bounded and, by the Extreme Value Theorem, attains a maximum, i.e., we can set

$$L = \sup_{z \in \overline{B_\delta(x)}} \|\nabla s(z)\| < \infty.$$

In addition, since the closed ball $\overline{B_\delta(x)}$ is convex, the line segment

$$\phi(t) = x + t(x' - x), \quad t \in [0,1]$$

connecting x and x' lies entirely in the $\overline{B_\delta(x)}$. The one-variable function $\psi := s \circ \phi : t \mapsto s(x + t(x' - x))$ is differentiable being the composition of two differentiable maps and is such that:

- $\psi(1) = s(x')$,
- $\psi(0) = s(x)$ and
- $\psi'(\xi) = (s \circ \phi)'(\xi) = \nabla s(\phi(\xi))\phi'(\xi)$ by the chain rule.

By the one-dimensional Mean Value Theorem, there exists $\xi \in (0,1)$ such that:

$$\begin{aligned} \psi(1) - \psi(0) &= \psi'(\xi)(1-0) \\ &\Downarrow \\ s(x') - s(x) &= \nabla s(\phi(\xi))\phi'(\xi) \\ &= \nabla s(x - \xi(x' - x)) \cdot (x' - x). \end{aligned}$$

By the Cauchy–Schwarz inequality:

$$|s(x') - s(x)| \leq \|\nabla s(x - \xi(x' - x))\| \|x' - x\| \leq L\|x' - x\| \tag{10}$$

where $\|\cdot\|$ denotes the canonical norm induced by the dot product on $\mathbb{R}^n$. Importantly, Equation 10 implies that. on the ball $B_\delta(x)$ the function s satisfies a Lipschitz condition with constant L, i.e., it is **locally Lipschitz** around x. Finally, if $x' \in \overline{B_\delta(x)}$ is a counterfactual, then it has the opposite label with respect to x. Without loss of generality, we assume that $s(x) < t \leq s(x')$, so:

$$m(x) = s(x) - t \leq s(x') - t \leq s(x') - s(x). \tag{11}$$

Putting the inequalities together:

$$\begin{aligned} |m(x)| &\leq |s(x) - s(x')| && \text{by Eq. 9} \\ &\leq L \cdot \|x' - x\| && \text{by Eq. 8} \\ &= L \cdot d_{\mathcal{X}}(x, x') \end{aligned}$$

□

Remark 1. For a fixed Lipschitz constant $L < \infty$, a small instance-counterfactual distance $d(x,x')$ forces $s(x)$ to be close to the decision boundary, representing high uncertainty. Hence, when x is close to the decision boundary (i.e., $d(x,x') \to 0$, then Theorem 1 indicates that the confidence will also be low, i.e., $m(x) \to 0$.

Remark 2. If x is an outlier (i.e., $d(x,x') \to \infty$), Theorem 1 states that $|m(x)| < \infty$, implying that any value of confidence is attainable. To address this issue, the selective classification framework may introduce two rejection thresholds $\tau_r^{\min}$ and $\tau_r^{\max}$: the former targets low-confidence instances lying close to the decision boundary, i.e., $d(x,x') < \tau_r^{\min}$; the upper threshold captures instances whose distance to any counterfactual is so large that they are likely to lie outside the data manifold, i.e., $d(x,x') > \tau_r^{\max}$. In this latter case, the model's confidence estimates are no longer reliable, as assumptions such as a well-defined or stable local Lipschitz constant L may fail. Rejecting instances falling under either case would enable a more robust and calibrated abstention policy.

Remark 3. The assumption of (local) Lipschitz continuity constitutes a mild regularity condition, as it is met by many standard classification models by design or can be enforced approximately through minor modifications. This is true especially in deep learning models [57], as most activation functions have small or unit Lipschitz constant (e.g., ReLU, Leaky ReLU, SoftPlus, Tanh, Sigmoid, ArcTan or Softsign, max-pooling) and standard network components admit simple and explicit Lipschitz bounds (e.g., dropout, batch normalization and other pooling methods) [50]. We refer to [58] for a uniform convergence analysis framework in which sufficient conditions on weight matrices and bias vectors together with the Lipschitz constant are provided to ensure uniform convergence of deep neural networks.

Remark 4. Tree-based models, such as LightGBM, XGBoost, Random Forest, are step functions; hence, their Lipschitz constant is technically infinite at the boundaries. However, different techniques can be implemented to enforce a form of Lipschitz control on these models. For instance, a random forest averages many randomised trees, which smooths the decision threshold with the result that it behaves like a locally Lipschitz function: around most points the score is flat (no change), and only in aggregate does it vary by bounded amounts [59]. Similarly, boosted tree ensembles may be constrained (e.g., on leaf weights or tree depth) to ensure that the fitted score function does not oscillate wildly for small feature changes [60] or may be combined with certain optimisation techniques (e.g., stochastic gradient Langevin dynamics [61]) to attain the desired convergence guarantees [62].

Corollary 1. *If x' is the closest counterfactual to x under the binary classifier f for the distance metric d, then the distance to the closest counterfactual $d(x,x')$ provides the tightest Lipschitz-based upper bound on the prediction margin $m(x)$ among all valid counterfactuals.*

Proof. By Theorem 1, for any valid counterfactual $\tilde{x}$ of x along which s is locally L-Lipschizt, the prediction margin is bounded by:

$$|m(x)| \leq L \cdot d(x,\tilde{x}).$$

Since x' is defined as a closest counterfactual, it satisfies the minimality condition with respect to d:

$$d(x,x') \leq d(x,\tilde{x}) \quad \forall \tilde{x} \in \mathcal{X} \text{ such that } f(x) \neq f(\tilde{x})$$

Multiplying this inequality by the positive Lipschitz constant L preserves the order:

$$L \cdot d(x,x') \leq L \cdot d(x,\tilde{x})$$

Combining these inequalities yields

$$|m(x)| \leq L \cdot d(x,x') \leq L \cdot d(x,\tilde{x})$$

which concludes the proof. □

Remark 5. This corollary shows that the distance to the closest counterfactual yields the tightest counterfactual-based upper bound on the prediction margin. In the framework of selective classification, this result justifies using this quantity as a principled proxy for confidence, with instances close to a counterfactual necessarily having low margins and unstable predictions. Thresholding this distance therefore induces a rejection policy that aligns naturally with the geometry of the decision boundary, abstaining precisely on instances whose predictions are inherently unstable.

Corollary 2. *Let $f : \mathcal{X} \to \{0,1\}$ be a binary classifier with a score function s that is locally L-Lipschitz continuous with respect to the distance d and let x' be the closest counterfactual to x with respect to a distance $d' \neq d$. If d' is norm-equivalent to d, meaning that $\exists\, \alpha, \beta > 0 : \alpha \cdot d(a,b) \leq d'(a,b) \leq \beta \cdot d(a,b)\ \forall a,b \in \mathcal{X}$, then the distance d' provides a loose upper bound on the prediction margin $m(x)$ scaled by a positive constant $k = L\beta$:*

$$|m(x)| \leq k \cdot d'(x,x').$$

Proof. Let x' be the minimizer of the distance d' among all valid counterfactuals. Since x' is a valid counterfactual, Theorem 1 applies regardless of how x' is selected, hence:

$$|m(x)| \leq L \cdot d(x,x')$$

By the assumption of norm equivalence between d and d':

$$\alpha \cdot d'(x,x') \leq d(x,x') \leq \beta \cdot d'(x,x')$$

for some $\alpha, \beta > 0$. Combining these bounds:

$$|m(x)| \leq L \cdot d(x,x') \leq L\beta \cdot d'(x,x')$$

The corollary follows setting $k = L\beta$.

□

APPENDIX F CONFORMAL PREDICTION FORMULATION

When a single canonical counterfactual $E(x) = x'$ is returned, SC−CE scores can be embedded in a split-conformal calibration to obtain distribution-free guarantees. Define a nonconformity score on the calibration set as

$$\alpha_i = S(x_i) = \frac{1}{d(x_i, E(x_i))}, \qquad i = 1, \ldots, n_{\text{cal}},$$

so that larger α indicates closer counterfactuals (higher nonconformity/uncertainty). For a target coverage level $\delta \in (0,1)$ compute the $(1-\delta)$-quantile $q_{1-\delta}$ of the calibration scores $\{\alpha_i\}$ using the usual conformal quantile, i.e. the $\lceil (1-\delta)(n_{\text{cal}}+1) \rceil$-th smallest value. For a test point x define $\alpha_{\text{test}} = S(x)$. A conservative rejection rule is

$$\text{reject } x \quad \Longleftrightarrow \quad \alpha_{\text{test}} > q_{1-\delta}.$$

Equivalently, the conformal p-value may be computed as

$$p(x) = \frac{|\{i : \alpha_i \geq \alpha_{\text{test}}\}| + 1}{n_{\text{cal}} + 1},$$

and the instance is accepted only when $p(x) > \delta$. This follows the standard split-conformal calibration procedure.

APPENDIX G ADDITIONAL RESULTS FOR RQ2: SC−CE PERFORMANCE AGAINST BENCHMARKS

The following tables report the performance of SC−CE and the two chosen selective classification baselines on all benchmark datasets. Performance is computed in terms of Non-rejected Accuracy *NA* and Classification Quality *CQ*. As described in the experimental setting (Section IV), SC−CE tested on all combinations of the selected counterfactual explainers E (i.e., LoRE, GS, ILS and ILS$_l$), distance function d (e.g., Euclidean, Manhattan, cosine, etc.) and aggregation policy a (i.e., minimum, maximum and mean). We limit the content of the following tables only to the top performing combinations for each dataset D and black-box classifier f pair.

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
German	MLP	PlugInRule	**0.755**	0.754	0.756	0.758	0.762	0.759	0.763	0.767
		PlugInRuleAUC	0.743	0.750	0.749	0.752	0.759	0.762	0.768	0.774
		LoRE- centered $L2_{avg}$	0.754	0.760	**0.774**	0.771	**0.780**	**0.783**	0.786	0.795
		LoRE- $L2_{avg}$	0.750	0.749	0.766	**0.772**	0.779	0.777	**0.793**	**0.807**
		GS- centered $L2_{avg}$	0.751	0.757	0.757	0.764	0.766	0.765	0.771	0.770
		GS- centered $L2_{max}$	0.753	0.752	0.751	0.761	0.771	0.770	0.768	0.774
		LoRE- centered $L2_{min}$	0.750	0.749	0.746	0.742	0.760	0.778	0.782	0.794
		GS- $canberra_{min}$	0.754	0.753	0.756	0.758	0.760	0.756	0.774	0.782
		GS- centered $L2_{min}$	0.753	**0.763**	0.761	0.765	0.763	0.761	0.755	0.767
	LipMLP	PlugInRule	0.749	0.753	0.763	0.766	0.797	0.804	0.793	0.827
		PlugInRuleAUC	0.748	0.752	0.747	0.740	0.733	0.742	0.746	0.753
		ILS_l- $cosine_{max}$	**0.756**	**0.771**	**0.772**	**0.781**	**0.799**	0.795	0.807	0.824
		ILS_l- centered $L2_{max}$	0.746	0.756	0.769	0.781	**0.799**	0.800	0.818	0.826
		GS- centered $L2_{min}$	0.749	0.760	0.764	0.766	0.786	0.790	0.815	0.826
		GS- centered $L2_{avg}$	0.750	0.753	0.762	0.761	0.792	**0.804**	0.811	0.821
		GS- centered $L2_{max}$	0.751	0.749	0.759	0.776	0.788	0.797	0.808	0.820
		GS- $L2_{min}$	0.747	0.760	0.765	0.772	0.783	0.789	0.810	0.818
		LoRE- $canberra_{min}$	0.736	0.758	0.754	0.764	0.779	0.798	**0.821**	**0.829**
	RF	PlugInRule	0.754	0.754	0.779	0.797	0.797	0.797	0.801	0.804
		PlugInRuleAUC	**0.779**	0.779	0.785	0.781	0.784	0.781	0.804	0.812
		ILS- $L2_{avg}$	0.775	0.772	0.788	0.792	**0.820**	0.823	0.832	**0.854**
		ILS- centered $L2_{avg}$	0.754	**0.786**	**0.797**	**0.806**	0.814	0.809	**0.841**	0.844
		ILS- centered $L2_{max}$	0.763	0.779	0.793	0.796	0.812	0.812	0.835	0.833
		ILS_l- $canberra_{max}$	0.758	0.768	0.775	0.798	0.810	**0.825**	0.833	0.844
		ILS_l- $L2_{max}$	0.765	0.775	0.781	0.795	0.797	0.811	0.824	0.852
		ILS- $L2_{max}$	0.773	0.771	0.787	0.790	0.810	0.811	0.826	0.831
		ILS_l- $cosine_{min}$	0.771	0.779	0.795	0.794	0.798	0.809	0.818	0.835
	XGBoost	PlugInRule	0.749	0.752	0.758	0.771	0.779	0.783	0.799	0.813
		PlugInRuleAUC	0.749	0.752	0.764	0.765	0.767	0.770	0.780	0.794
		GS- $cosine_{avg}$	**0.751**	**0.761**	**0.767**	**0.773**	**0.796**	**0.801**	**0.802**	0.810
		GS- centered $L2_{min}$	0.744	0.756	0.764	0.770	0.772	0.797	0.799	**0.814**
		GS- $L2_{min}$	0.744	0.756	0.764	0.770	0.772	0.797	0.799	0.813
		GS- $canberra_{max}$	**0.751**	0.755	0.757	0.763	0.772	0.795	0.801	0.810
		GS- $L2_{max}$	**0.751**	0.755	0.757	0.763	0.772	0.790	0.801	0.809
		GS- centered $L2_{max}$	**0.751**	0.755	0.757	0.763	0.772	0.790	0.801	0.809
		GS- $canberra_{min}$	0.744	0.756	0.764	0.770	0.772	0.789	0.799	0.805

TABLE 13: Top Selective Classifiers in Non-Rejected Accuracy for the dataset **German**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Adult	MLP	PlugInRule	0.878	0.890	0.905	**0.914**	0.927	**0.942**	0.954	**0.967**
		PlugInRuleAUC	0.867	0.867	0.871	0.877	0.905	0.912	0.915	0.942
		ILS_l- $cosine_{avg}$	0.878	**0.902**	**0.910**	0.911	**0.932**	0.939	0.954	0.952
		LoRE- $canberra_{min}$	0.876	0.887	0.901	0.912	0.929	0.937	0.950	**0.967**
		LoRE- $L2_{min}$	**0.880**	0.895	0.903	0.911	0.927	0.932	0.944	0.957
		LoRE- $canberra_{avg}$	0.876	0.890	0.896	0.909	0.930	0.931	**0.955**	0.961
		ILS_l- $canberra_{avg}$	0.878	0.890	0.906	0.912	0.917	0.932	0.953	0.957
		GS- $cosine_{max}$	0.879	0.890	0.898	0.906	0.925	0.940	0.952	0.953
		ILS_l- $L2_{min}$	0.876	0.891	0.907	0.911	0.926	0.932	0.943	0.957
	LipMLP	PlugInRule	0.881	0.901	0.918	0.918	0.940	0.946	0.954	0.960
		PlugInRuleAUC	0.853	0.862	0.859	0.859	0.879	0.894	0.910	0.939
		ILS- $canberra_{min}$	**0.886**	0.900	**0.919**	**0.928**	**0.945**	**0.950**	**0.958**	0.965
		GS- $L2_{max}$	0.880	**0.905**	0.915	0.918	0.937	0.945	0.951	0.961
		GS- $cosine_{max}$	0.882	0.904	0.912	0.921	0.935	0.938	0.953	0.960
		GS- $canberra_{min}$	0.879	0.895	0.916	0.921	0.930	0.939	0.953	**0.969**
		GS- $cosine_{avg}$	0.882	0.904	0.912	0.920	0.934	0.937	0.950	0.963
		ILS_l- $L2_{min}$	0.875	0.901	0.914	0.923	0.939	0.944	0.946	0.955
		ILS_l- $cosine_{avg}$	0.874	0.887	0.913	0.925	0.941	0.943	0.955	0.958
	RF	PlugInRule	0.879	0.895	0.912	0.923	0.936	0.945	0.951	0.967
		PlugInRuleAUC	0.872	0.868	0.875	0.875	0.884	0.886	0.894	0.913
		ILS_l- centered $L2_{min}$	**0.882**	0.889	0.900	0.907	**0.958**	**0.962**	**0.968**	**0.977**
		ILS_l- $cosine_{min}$	0.878	0.885	0.904	0.924	0.938	0.957	0.964	0.974
		ILS_l- $cosine_{avg}$	0.879	**0.898**	**0.913**	0.918	0.937	0.944	0.953	0.970
		ILS_l- centered $L2_{avg}$	0.878	0.894	**0.913**	0.922	0.938	0.947	0.954	0.965
		ILS_l- $L2_{min}$	0.878	0.895	0.912	**0.927**	0.939	0.942	0.949	0.947
		ILS- $cosine_{min}$	0.876	0.895	0.901	0.914	0.944	0.951	0.952	0.950
		ILS- $cosine_{avg}$	0.872	0.896	0.912	0.917	0.928	0.944	0.955	0.959
	XGBoost	PlugInRule	0.891	0.905	**0.922**	0.925	0.938	0.948	0.954	**0.980**
		PlugInRuleAUC	0.870	0.876	0.876	0.887	0.900	0.902	0.915	0.935
		ILS- $L2_{avg}$	0.883	0.908	0.919	0.929	0.944	**0.950**	0.956	0.963
		ILS_l- $cosine_{max}$	**0.894**	**0.910**	0.919	0.929	0.940	0.944	0.950	0.966
		ILS_l- $cosine_{avg}$	0.885	0.909	0.917	**0.931**	0.939	0.950	0.948	0.957
		ILS_l- $canberra_{min}$	0.880	0.905	0.919	0.929	0.940	0.949	**0.956**	0.955
		ILS_l- $L2_{max}$	0.894	0.905	0.912	0.925	0.937	0.940	0.950	0.970
		ILS- $L2_{max}$	0.886	0.903	0.917	0.921	**0.948**	**0.950**	0.953	0.952
		ILS- $cosine_{avg}$	0.884	0.901	0.918	0.929	0.936	0.948	0.946	0.960

TABLE 14: Top Selective Classifiers in Non-Rejected Accuracy for the dataset **Adult**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Wisconsin	**MLP**	PlugInRule	0.979	0.985	0.984	0.984	0.990	**1.000**	**1.000**	**1.000**
		PlugInRuleAUC	0.979	0.978	0.978	0.977	0.975	0.973	0.981	0.978
		LoRE- centered L2$_{min}$	0.979	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		LoRE- L2$_{min}$	0.986	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		LoRE- cosine$_{min}$	0.986	**0.992**	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		ILS- centered L2$_{max}$	**0.993**	0.992	0.992	0.992	0.990	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{min}$	0.979	0.992	0.992	0.992	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{avg}$	0.979	0.992	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{max}$	0.979	0.992	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
	LipMLP	PlugInRule	0.978	0.985	0.984	0.982	0.981	0.990	**1.000**	**1.000**
		PlugInRuleAUC	0.978	0.977	0.977	0.976	0.973	0.981	0.990	0.988
		ILS$_l$- canberra$_{max}$	0.978	**0.992**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		LoRE- centered L2$_{min}$	0.985	0.985	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		LoRE- cosine$_{min}$	0.979	0.985	0.984	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{avg}$	0.979	0.985	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		GS- cosine$_{max}$	0.979	0.985	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{max}$	0.979	0.984	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		ILS- cosine$_{avg}$	**0.986**	0.985	0.992	0.991	0.990	**1.000**	**1.000**	**1.000**
	RF	PlugInRule	0.985	0.984	0.992	0.991	**1.000**	**1.000**	**1.000**	**1.000**
		PlugInRuleAUC	0.965	0.964	0.961	0.961	0.957	0.938	0.935	0.933
		GS- L2$_{max}$	**0.992**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{max}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		LoRE- L2$_{min}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{avg}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- canberra$_{min}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- centered L2$_{min}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- L2$_{min}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	XGBoost	PlugInRule	**0.993**	0.993	0.992	0.990	**1.000**	**1.000**	**1.000**	**1.000**
		PlugInRuleAUC	0.978	0.978	0.977	0.977	0.977	0.976	0.975	0.971
		ILS- cosine$_{max}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		ILS$_l$- L2$_{max}$	**0.993**	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		ILS- canberra$_{max}$	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- canberra$_{min}$	0.986	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- cosine$_{min}$	0.986	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- cosine$_{max}$	0.986	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		GS- L2$_{min}$	0.986	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

TABLE 15: Top Selective Classifiers in Non-Rejected Accuracy for the dataset **Wisconsin**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Two-Moons	MLP	PlugInRule	0.952	0.968	0.976	0.980	0.994	**0.994**	0.994	**1.000**
		PlugInRuleAUC	0.911	0.956	0.968	0.980	0.989	0.994	0.994	0.993
		GS- $L2_{min}$	0.956	0.963	**0.986**	**0.985**	**0.995**	0.994	0.994	**1.000**
		GS- $L2_{max}$	0.952	0.963	**0.986**	**0.985**	**0.994**	0.994	0.994	**1.000**
		GS- $L2_{avg}$	0.952	0.963	**0.986**	**0.985**	**0.994**	0.994	0.994	**1.000**
		LoRE- $L2_{min}$	0.944	0.968	0.977	**0.986**	0.989	0.994	**1.000**	**1.000**
		GS- $canberra_{max}$	**0.956**	0.972	0.976	0.980	0.979	0.989	0.988	0.993
		GS- $canberra_{avg}$	0.952	0.972	0.976	0.980	0.979	0.989	0.988	0.993
		GS- $canberra_{min}$	0.948	**0.972**	0.976	0.980	0.979	0.989	0.988	0.986
	LipMLP	PlugInRule	**0.952**	0.968	0.981	0.985	0.989	0.994	0.994	0.993
		PlugInRuleAUC	0.934	0.945	0.963	0.971	0.989	0.988	0.994	0.992
		ILS- $L2_{min}$	0.949	**0.977**	**0.986**	0.990	**0.995**	0.994	0.994	**1.000**
		LoRE- $L2_{min}$	0.948	0.972	0.981	0.985	0.995	**0.994**	**1.000**	**1.000**
		ILS_l- $L2_{min}$	**0.952**	0.972	0.986	0.985	0.989	0.989	**1.000**	**1.000**
		ILS_l- $cosine_{min}$	0.952	0.963	**0.986**	**0.995**	**0.995**	**0.994**	0.994	0.994
		ILS_l- $L2_{avg}$	0.947	0.972	0.986	0.985	0.989	0.989	**1.000**	**1.000**
		GS- $L2_{max}$	0.948	0.968	0.986	0.985	0.994	0.994	0.993	**1.000**
		GS- $L2_{avg}$	0.948	0.959	0.986	0.985	0.994	0.994	0.993	**1.000**
	RF	PlugInRule	0.933	0.951	0.960	0.965	0.978	0.983	0.987	0.993
		PlugInRuleAUC	0.919	0.928	0.938	0.962	0.959	0.972	0.975	0.986
		LoRE- $L2_{min}$	0.944	0.954	0.971	0.980	**0.995**	0.994	0.994	**1.000**
		LoRE- $L2_{avg}$	0.936	0.939	**0.976**	**0.984**	0.994	**1.000**	**1.000**	**1.000**
		GS- $L2_{max}$	0.948	**0.959**	0.963	0.964	0.989	0.994	0.994	**1.000**
		GS- $L2_{avg}$	0.948	**0.959**	0.963	0.966	0.984	0.988	0.994	**1.000**
		GS- $L2_{min}$	0.948	**0.959**	0.962	0.961	0.983	0.989	0.988	**1.000**
		GS- $canberra_{avg}$	**0.952**	0.955	0.962	0.970	0.984	0.982	0.987	0.987
		GS- $canberra_{max}$	0.952	0.954	0.958	0.974	0.979	0.982	0.987	0.987
	XGBoost	PlugInRule	0.935	0.941	0.962	0.961	0.961	0.971	0.970	0.966
		PlugInRuleAUC	0.937	0.950	0.946	0.955	0.952	0.981	0.981	0.980
		GS- $L2_{max}$	**0.950**	0.969	0.977	**0.990**	**0.994**	0.994	0.993	**1.000**
		GS- $L2_{avg}$	0.950	0.969	0.977	**0.990**	0.989	0.994	0.993	**1.000**
		GS- $L2_{min}$	0.950	**0.969**	**0.977**	**0.990**	0.989	0.994	0.993	0.993
		GS- $canberra_{max}$	0.946	0.964	0.977	0.985	0.989	**0.994**	**0.994**	0.993
		GS- $canberra_{min}$	0.946	0.964	**0.977**	0.985	0.989	0.988	0.994	0.993
		GS- $canberra_{avg}$	0.946	0.964	0.977	0.985	0.989	0.989	0.994	0.993
		GS- $cosine_{max}$	0.946	0.960	0.977	0.980	0.989	0.988	**0.994**	0.993

TABLE 16: Top Selective Classifiers in Non-Rejected Accuracy for the dataset **Two-Moons**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
German	MLP	PlugInRule	**0.748**	0.712	0.700	0.692	0.680	0.660	0.656	0.640
		PlugInRuleAUC	0.716	0.712	0.684	0.680	0.660	0.644	0.636	0.596
		LoRE- centered L2$_{avg}$	0.744	**0.740**	**0.740**	0.692	0.684	0.684	**0.684**	0.660
		LoRE- L2$_{avg}$	0.744	0.724	0.728	**0.704**	**0.692**	0.684	0.676	**0.664**
		LoRE- centered L2$_{min}$	0.744	0.716	0.704	0.684	0.680	**0.688**	0.680	0.644
		GS- centered L2$_{avg}$	**0.748**	0.728	0.704	0.696	0.684	0.672	0.672	0.632
		GS- canberra$_{min}$	0.744	0.724	0.716	0.700	0.672	0.652	0.664	0.648
		GS- centered L2$_{max}$	0.740	0.712	0.700	0.700	**0.692**	0.680	0.664	0.624
		LoRE- cosine$_{min}$	0.728	0.716	0.704	0.676	0.672	0.676	0.672	0.648
	LipMLP	PlugInRule	0.728	0.728	0.732	0.732	**0.740**	0.720	0.680	**0.700**
		PlugInRuleAUC	0.740	0.740	0.712	0.680	0.624	0.612	0.608	0.580
		ILS$_l$- cosine$_{avg}$	**0.752**	0.740	**0.752**	0.728	0.728	0.716	0.704	0.696
		ILS$_l$- cosine$_{max}$	0.748	**0.760**	0.748	**0.744**	0.724	0.708	0.692	0.688
		GS- centered L2$_{min}$	0.728	0.736	0.736	0.724	0.732	0.724	**0.708**	0.696
		GS- centered L2$_{max}$	0.736	0.728	0.732	**0.744**	0.728	0.716	0.696	**0.700**
		ILS$_l$- centered L2$_{avg}$	0.748	0.720	0.736	0.740	0.724	**0.728**	0.704	0.680
		GS- L2$_{min}$	0.728	0.736	0.740	0.736	0.728	0.720	0.704	0.680
		ILS$_l$- L2$_{min}$	0.744	0.744	0.732	0.724	0.728	0.724	0.696	0.676
	RF	PlugInRule	0.744	0.744	0.764	0.764	0.740	0.740	0.732	0.720
		PlugInRuleAUC	0.744	0.744	0.748	0.720	0.680	0.660	0.656	0.620
		ILS- centered L2$_{avg}$	0.744	**0.772**	0.776	**0.780**	0.776	0.752	**0.764**	0.724
		ILS- L2$_{avg}$	**0.776**	0.752	0.760	0.752	0.760	0.760	0.756	**0.752**
		ILS- centered L2$_{max}$	0.756	**0.772**	**0.780**	**0.780**	**0.780**	**0.764**	0.736	0.696
		ILS$_l$- centered L2$_{avg}$	0.772	0.764	0.764	0.760	0.744	0.744	0.728	0.736
		ILS- L2$_{max}$	**0.776**	0.760	0.756	0.744	0.756	0.748	0.744	0.720
		ILS$_l$- canberra$_{max}$	0.752	0.756	0.756	0.768	0.744	0.752	0.744	0.724
		ILS$_l$- cosine$_{min}$	0.768	0.752	0.764	0.752	0.744	0.740	0.736	0.736
	XGBoost	PlugInRule	0.740	0.736	0.724	0.728	0.720	0.716	0.720	0.712
		PlugInRuleAUC	0.740	0.712	0.720	0.716	0.696	0.668	0.664	0.656
		GS- cosine$_{max}$	0.744	0.740	**0.740**	**0.740**	**0.740**	**0.740**	0.732	0.732
		GS- cosine$_{min}$	0.736	0.736	0.736	0.736	0.736	0.736	**0.736**	**0.736**
		GS- cosine$_{avg}$	**0.748**	**0.744**	**0.740**	0.736	0.720	0.716	0.708	0.688
		ILS$_l$- centered L2$_{max}$	0.736	0.708	0.716	0.712	0.716	0.720	0.716	0.700
		GS- canberra$_{max}$	0.732	0.732	0.720	0.708	0.696	0.716	0.716	0.700
		GS- L2$_{max}$	0.732	0.732	0.720	0.708	0.696	0.708	0.716	0.696
		GS- centered L2$_{max}$	0.732	0.732	0.720	0.708	0.696	0.708	0.716	0.696

TABLE 17: Top Selective Classifiers in Classification Quality for the dataset **German**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Adult	MLP	PlugInRule	0.852	0.856	**0.852**	0.830	**0.816**	0.774	0.734	0.694
		PlugInRuleAUC	0.808	0.784	0.772	0.758	0.730	0.718	0.702	0.680
		ILS_l- $L2_{min}$	0.854	0.850	0.850	**0.834**	0.804	0.766	0.758	0.724
		ILS- $cosine_{avg}$	0.854	**0.866**	0.848	**0.834**	0.796	**0.796**	0.742	0.698
		GS- $L2_{max}$	0.852	0.862	0.846	0.830	0.794	0.782	0.750	0.718
		ILS_l- $canberra_{avg}$	**0.864**	0.848	0.842	0.828	0.796	0.770	0.756	0.726
		GS- $cosine_{max}$	0.854	0.858	0.840	0.830	**0.816**	0.778	0.740	0.712
		LoRE- $cosine_{min}$	0.856	0.842	0.846	0.826	0.796	0.782	**0.760**	0.710
		ILS_l- $canberra_{min}$	0.854	0.860	0.838	0.828	0.784	0.770	0.754	**0.728**
	LipMLP	PlugInRule	0.866	0.860	0.862	0.842	0.792	0.780	0.744	0.700
		PlugInRuleAUC	0.830	0.794	0.760	0.738	0.720	0.712	0.696	0.690
		ILS- $canberra_{min}$	0.866	0.864	**0.866**	**0.856**	0.828	0.798	0.758	0.726
		GS- $L2_{max}$	0.866	**0.878**	0.864	0.844	0.812	0.792	0.768	0.714
		GS- $cosine_{max}$	**0.868**	0.874	0.856	0.852	0.814	0.796	**0.770**	0.702
		ILS- centered $L2_{avg}$	0.852	0.858	0.850	0.850	**0.832**	0.804	0.766	0.720
		GS- centered $L2_{min}$	0.860	0.864	0.862	0.834	0.816	**0.806**	0.756	**0.728**
		LoRE- $canberra_{min}$	**0.868**	0.866	0.860	0.848	0.812	0.788	0.748	0.710
		LoRE- $cosine_{min}$	0.862	0.860	0.856	0.840	0.822	0.790	0.754	0.710
	RF	PlugInRule	0.860	0.842	0.826	0.816	0.788	**0.778**	0.728	0.700
		PlugInRuleAUC	0.846	0.806	0.792	0.770	0.732	0.704	0.686	0.666
		LoRE- $canberra_{min}$	**0.868**	**0.854**	**0.846**	0.822	0.794	0.764	**0.744**	0.696
		GS- $L2_{max}$	0.866	0.848	0.840	**0.830**	0.782	0.762	0.738	0.708
		ILS_l- centered $L2_{avg}$	0.862	0.848	0.832	0.826	0.778	0.774	0.732	**0.716**
		GS- $canberra_{max}$	0.860	0.848	0.832	**0.830**	**0.806**	0.768	0.726	0.694
		ILS_l- $cosine_{avg}$	0.862	0.850	0.832	0.828	0.776	0.764	**0.744**	0.708
		ILS_l- $canberra_{max}$	0.862	0.844	0.834	0.820	0.804	0.772	0.734	0.694
		GS- centered $L2_{max}$	0.866	0.846	0.836	0.824	0.780	0.758	0.730	0.714
	XGBoost	PlugInRule	**0.876**	0.864	**0.862**	0.836	0.800	0.778	0.726	**0.702**
		PlugInRuleAUC	0.834	0.808	0.782	0.772	0.734	0.718	0.712	0.686
		ILS_l- $cosine_{min}$	0.870	0.862	0.846	0.832	0.796	0.778	**0.738**	0.682
		ILS_l- $cosine_{max}$	0.870	0.860	0.848	**0.848**	0.798	0.766	0.736	0.668
		ILS_l- centered $L2_{min}$	0.870	0.858	0.848	0.842	0.792	**0.786**	0.736	0.660
		ILS_l- $L2_{max}$	0.868	0.862	0.844	0.830	0.788	0.768	0.736	0.694
		ILS_l- $L2_{min}$	0.866	0.864	0.848	0.840	**0.804**	0.778	0.722	0.660
		ILS_l- $cosine_{avg}$	0.868	**0.866**	0.848	0.828	0.790	0.776	0.718	0.686
		ILS_l- $L2_{avg}$	0.864	0.864	0.850	0.818	0.802	0.762	0.732	0.654

TABLE 18: Top Selective Classifiers in Classification Quality for the dataset **Adult**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Wisconsin	**MLP**	PlugInRule	0.972	0.930	0.894	0.859	0.711	0.711	0.634	0.577
		PlugInRuleAUC	0.972	**0.951**	0.923	0.880	0.810	0.754	**0.732**	0.641
		ILS- canberra$_{max}$	0.965	0.937	**0.930**	**0.908**	**0.859**	**0.768**	0.704	0.620
		ILS- centered L2$_{min}$	**0.979**	0.923	0.824	0.803	0.796	**0.768**	**0.732**	**0.690**
		ILS$_l$- centered L2$_{min}$	0.908	0.894	0.873	0.845	0.796	0.754	0.711	0.627
		LoRE- canberra$_{min}$	0.972	0.937	0.887	0.803	0.739	0.746	0.662	0.606
		GS- canberra$_{min}$	0.972	0.930	0.923	0.810	0.732	0.697	0.669	0.570
		ILS- canberra$_{avg}$	0.972	0.915	0.852	0.803	0.761	0.718	0.697	0.585
		ILS- centered L2$_{max}$	**0.979**	0.880	0.859	0.845	0.739	0.690	0.648	0.570
	LipMLP	PlugInRule	0.958	0.930	**0.901**	0.796	0.732	0.711	0.634	0.592
		PlugInRuleAUC	0.944	0.901	0.887	0.852	0.775	0.746	0.725	0.613
		ILS- canberra$_{min}$	0.965	0.894	0.887	0.824	0.782	**0.761**	**0.732**	**0.627**
		LoRE- canberra$_{min}$	0.972	0.930	0.894	0.852	0.746	0.739	0.704	0.585
		ILS- L2$_{min}$	0.951	0.923	0.887	0.852	**0.803**	0.739	0.627	0.606
		ILS- centered L2$_{avg}$	**0.986**	**0.965**	0.887	0.810	0.754	0.725	0.634	0.599
		ILS- centered L2$_{max}$	**0.986**	0.937	**0.901**	**0.880**	0.746	0.690	0.613	0.592
		GS- canberra$_{min}$	0.979	**0.965**	0.887	0.845	0.739	0.732	0.634	0.563
		ILS- cosine$_{max}$	0.979	0.937	0.887	0.838	0.746	0.690	0.676	0.592
	RF	PlugInRule	0.944	0.887	0.880	0.810	0.732	0.704	0.648	0.585
		PlugInRuleAUC	**0.958**	0.937	0.859	**0.859**	0.782	0.535	0.507	0.493
		ILS- centered L2$_{max}$	0.951	0.908	0.880	0.852	**0.789**	0.746	0.711	0.620
		LoRE- centered L2$_{min}$	0.951	0.901	0.873	0.845	0.725	0.718	0.704	**0.676**
		ILS- centered L2$_{avg}$	0.944	0.887	0.866	0.817	0.775	0.725	**0.725**	0.627
		ILS$_l$- cosine$_{avg}$	0.937	0.880	0.866	0.817	0.754	0.746	0.718	0.641
		LoRE- canberra$_{min}$	0.944	**0.944**	**0.901**	0.824	**0.789**	0.732	0.669	0.549
		ILS- cosine$_{max}$	0.951	0.908	0.880	0.838	0.754	0.718	0.690	0.606
		ILS- canberra$_{min}$	0.951	0.859	0.845	0.831	0.775	**0.761**	0.697	0.620
	XGBoost	PlugInRule	**0.972**	**0.951**	0.859	0.711	0.634	0.606	0.542	0.514
		PlugInRuleAUC	0.958	0.937	0.915	**0.915**	**0.880**	**0.852**	**0.810**	**0.711**
		ILS$_l$- centered L2$_{max}$	**0.972**	0.930	**0.930**	0.901	0.866	0.845	0.782	0.704
		ILS$_l$- centered L2$_{avg}$	**0.972**	0.908	0.880	0.866	0.845	0.817	0.796	0.690
		ILS$_l$- cosine$_{max}$	0.965	0.937	0.859	0.859	0.789	0.761	0.725	0.683
		ILS$_l$- canberra$_{max}$	**0.972**	0.937	0.887	0.859	0.789	0.761	0.690	0.599
		ILS- centered L2$_{avg}$	0.937	0.908	0.866	0.852	0.817	0.761	0.711	0.620
		LoRE- centered L2$_{max}$	0.965	0.880	0.866	0.803	0.754	0.739	0.704	0.676
		ILS$_l$- centered L2$_{min}$	0.958	0.923	0.873	0.845	0.775	0.725	0.669	0.606

TABLE 19: Top Selective Classifiers in Classification Quality for the dataset **Wisconsin**

Dataset	Black Box	Rejection Policy	Target Coverages							
			95%	90%	85%	80%	75%	70%	65%	60%
Two-Moons	MLP	PlugInRule	**0.912**	0.904	0.892	0.872	0.792	0.772	0.696	0.656
		PlugInRuleAUC	0.900	0.908	0.900	0.864	0.824	0.748	0.708	0.620
		GS- centered L2$_{min}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		GS- centered L2$_{max}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		GS- centered L2$_{avg}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		ILS- centered L2$_{min}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		ILS- centered L2$_{max}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		ILS- centered L2$_{avg}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
		LoRE- centered L2$_{min}$	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**	**0.912**
	LipMLP	PlugInRule	0.896	0.896	0.888	0.864	0.788	0.740	0.680	0.640
		PlugInRuleAUC	0.920	0.908	0.872	0.848	0.796	0.740	0.688	0.592
		GS- centered L2$_{min}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		GS- centered L2$_{max}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		GS- centered L2$_{avg}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		ILS- centered L2$_{min}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		ILS- centered L2$_{max}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		ILS- centered L2$_{avg}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		LoRE- centered L2$_{min}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
	RF	PlugInRule	**0.916**	0.888	0.828	0.828	0.796	0.780	0.700	0.660
		PlugInRuleAUC	0.912	0.888	0.872	0.852	0.796	0.768	0.692	0.624
		GS- centered L2$_{min}$	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**
		GS- centered L2$_{max}$	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**
		GS- centered L2$_{avg}$	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**
		LoRE- centered L2$_{min}$	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**
		LoRE- centered L2$_{max}$	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**	**0.916**
		LoRE- canberra$_{max}$	0.888	0.888	0.888	0.888	0.888	0.888	0.888	0.888
		ILS- centered L2$_{max}$	**0.916**	**0.916**	**0.916**	**0.916**	0.756	0.756	0.756	0.756
	XGBoost	PlugInRule	**0.928**	0.912	0.856	0.824	0.736	0.716	0.692	0.620
		PlugInRuleAUC	0.908	0.868	0.800	0.796	0.756	0.696	0.672	0.636
		GS- centered L2$_{min}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		GS- centered L2$_{max}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		GS- centered L2$_{avg}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		LoRE- centered L2$_{min}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		LoRE- centered L2$_{max}$	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**	**0.928**
		ILS- centered L2$_{min}$	0.916	0.916	0.916	0.916	0.916	0.916	0.916	0.916
		ILS- centered L2$_{max}$	0.916	0.916	0.916	0.916	0.916	0.916	0.916	0.916

TABLE 20: Top Selective Classifiers in Classification Quality for the dataset **Two-Moons**