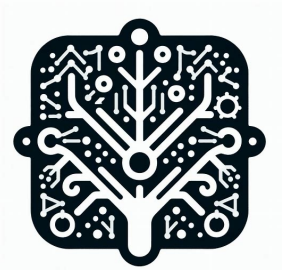
Coevolutionary Algorithm for Building Robust Decision Trees under Minimax Regret



Adam Żychowski[1], Andrew Perrault[2], Jacek Mańdziuk[1,3,4]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology
[2]Department of Computer Science and Engineering, The Ohio State University
[3]Faculty of Computer Science, AGH University of Krakow
[4]Center of Excellence in Artificial Intelligence, AGH University of Krakow

Problem definition

Goal: find **robust decision trees.**

What means *robust*?

Decision tree designed to handle variations and uncertainties in the data effectively.

How to measure it?

- adversarial accuracy
- max regret

Problem definition

Accuracy $$\mathrm{acc}(h) = \frac{1}{|X|} \sum_{x_i \in X} I[h(x_i) = y_i]$$

Adversarial accuracy $$\mathrm{acc}_{\mathrm{adv}}(h, \epsilon) = \frac{1}{|X|} \sum_{x_i \in X} \min_{z_i \in \mathcal{N}_\varepsilon(x_i)} I[h(z_i) = y_i]$$

$\mathcal{N}_\varepsilon(x) = \{z : ||z - x||_\infty \leq \varepsilon\}$ is a ball with center x and radius ε under the L_∞

Max regret $$\mathrm{mr}(h) = \max_{z_i \in \mathcal{N}_\varepsilon(x_i)} \mathrm{regret}(h, \{z_i\})$$

$$\mathrm{regret}(h, \{z_i\}) = \max_{h'} \mathrm{acc}(h', \{z_i\}) - \mathrm{acc}(h, \{z_i\})$$

$\mathrm{acc}(h, \{z_i\})$ is the accuracy achieved by h when $\{x_i\}$ is replaced with $\{z_i\}$

Example

T1

	CS	I	D
A1	50	10	0
A2	60	20	1
A3	60	30	1

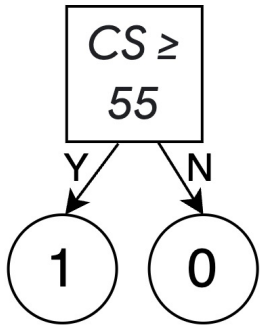
T2

	CS	I	D
A1	55 (+5)	10	0
A2	55 (-5)	20	1
A3	60 (+0)	30	1

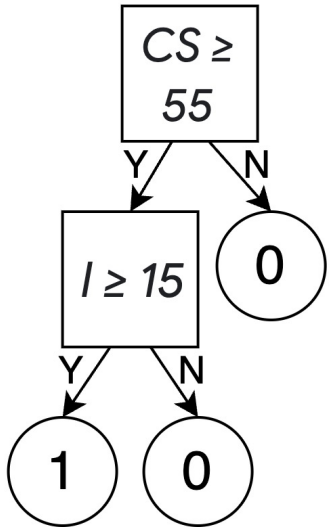
T3

	CS	I	D
A1	55 (+5)	20 (+10)	0
A2	55 (-5)	20 (+0)	1
A3	55 (-5)	20 (-10)	1

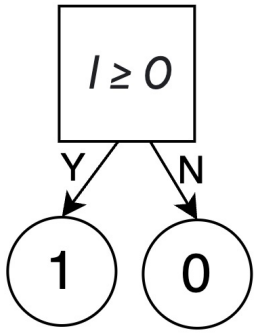
DT1



DT2



DT3



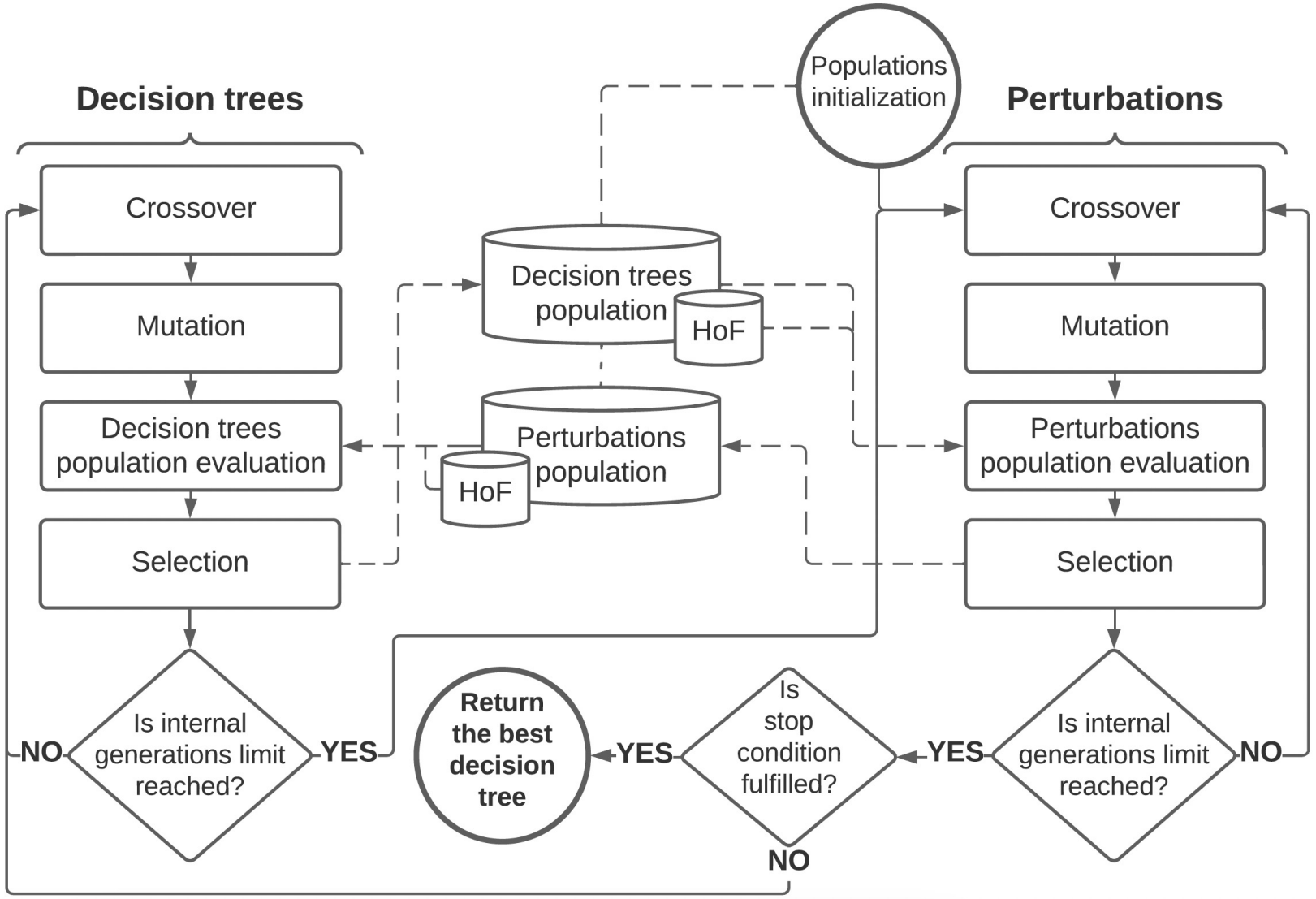
Motivation

- adversarial accuracy might provide an overly optimistic or pessimistic view of the model's robustness by focusing only on absolute accuracy value
- max regret is a more realistic approach since it counts the magnitude of the potential loss by considering the model trained on perturbed data
- max regret cannot be directly optimized and used as a splitting criterion in the state-of-the-art algorithms (e.g. GROOT[1], FPRDT[2])

We propose a novel coevolutionary algorithm (CoEvoRDT) which address this issue.

[1] Vos, D. and Verwer, S. 2021. Efficient training of robust decision trees against adversarial examples. ICML, 10586–10595.
[2] Guo, J.-Q. Teng, M.-Z. Gao, W. and Zhou, Z.-H. 2022. Fast Provably Robust Decision Trees and Boosting. ICML, 8127–8144.

CoEvoRDT algorithm overview



Decision trees population

Each **decision tree is encoded as a list of nodes**, where each node is represented by a 7-tuple $\{t, c, P, L, R, o, v, a\}$: node number ($t$), class label ($c$), parent node pointer ($P$), left and right children pointers (L and R), operator indication (o), value to be tested (v), and attribute (a).

Initial population: random decision trees with depth between 2 and 10.

Crossover: occurs with a probability, randomly pairing individuals and exchanging entire subtrees between selected nodes to generate offspring.

Mutation: applied with a probability, introducing random changes through actions like subtree replacement, node information change, or subtree pruning.

Perturbations population

Each **individual represents a perturbed input set**, with perturbations constrained within ε.

Initial population: Random perturbations generated uniformly, meeting ε criteria.

Crossover: Selects a random subset of individuals, pairs them randomly, and mixes perturbed input instances from both parents to generate offspring.

Mutation: Independently perturbing each input instance's encoded values.

Hall of Fame

Role: Mechanism to retain and store best-performing individuals encountered during evolution.

Common approach critique: Traditional approach adds one highest-fitness individual per generation, potentially suboptimal for diversity.

CoEvoRDT approach: Utilizes a game-theoretic approach treating decision trees and perturbations as strategies in a non-cooperative zero-sum game.

Mixed Nash Equilibrium: Calculates mixed Nash equilibrium, resulting in mixed strategies for both decision trees and perturbations.

Evaluation enhancement: Fitness function calculated against a merged set of Hall of Fame and population individuals.

Results – max regret

dataset	CART	Meta Silvae	RIGDT-h	GROOT	FPRDT	CoEvoRDT	**CoEvoRDT+FPRDT**
ionos	.094±.000	.075±.007	.071±.006	.061±.005	.061±.006	**.052±.004**	.052±.005
breast	.103±.000	.056±.006	.069±.006	.059±.005	.057±.005	**.049±.004**	.049±.005
diabetes	.202±.000	.126±.008	.132±.009	.124±.009	.117±.007	**.096±.006**	.094±.007
bank	.186±.000	.102±.007	.108±.008	.090±.006	.089±.007	**.076±.006**	.076±.006
Japan3v4	.107±.000	.090±.006	.083±.006	.067±.006	.066±.004	**.062±.006**	.061±.006
spam	.097±.000	.079±.006	.083±.006	.074±.006	.074±.006	**.070±.005**	.069±.005
GesDvP	.152±.000	.129±.008	.133±.010	.129±.008	.131±.009	**.114±.007**	.114±.007
har1v2	.105±.000	.074±.006	.084±.007	.068±.006	.068±.006	**.064±.005**	.064±.005
wine	.140±.000	.125±.008	.127±.009	.111±.009	.109±.008	**.090±.006**	.090±.007
collision-det	.142±.000	.099±.007	.093±.007	.088±.006	.091±.007	**.061±.006**	.059±.006
mnist:1v5	.249±.000	.078±.007	.076±.006	.071±.006	.067±.005	**.055±.006**	.055±.005
mnist:2v6	.268±.000	.083±.007	.087±.006	.072±.005	.069±.005	**.055±.004**	.054±.004
mnist	.395±.000	.143±.009	.139±.009	.125±.007	.124±.009	**.113±.008**	.112±.008
f-mnist2v5	.273±.000	.254±.015	.249±.015	.223±.013	.238±.014	**.196±.011**	.196±.011
f-mnist3v4	.290±.000	.259±.014	.254±.015	.246±.014	.232±.013	**.202±.011**	.199±.011
f-mnist7v9	.283±.000	.255±.014	.251±.015	.237±.014	.240±.014	**.208±.013**	.207±.012
f-mnist	.427±.000	.345±.020	.337±.018	.292±.017	.286±.016	**.238±.014**	.237±.015
cifar10:0v5	.419±.000	.351±.019	.379±.021	.347±.019	.314±.018	**.241±.015**	.236±.013
cifar10:0v6	.403±.000	.362±.021	.368±.020	.342±.018	.341±.019	**.289±.016**	.289±.016
cifar10:4v8	.408±.000	.357±.019	.360±.021	.339±.018	.331±.019	**.283±.016**	.281±.017

Table 1. **Max regrets** (mean ± std error). CoEvoRDT+FPRDT obtained the best results for all datasets. The best results are **bolded**. Gray background indicates that a given method is statistically significantly better than all other methods.

Results – adversarial accuracy

dataset	CART	Meta Silvae	RIGDT-h	GROOT	FPRDT	CoEvoRDT	**CoEvoRDT+FPRDT**
ionos	.310±.000	.695±.039	.701±.045	.783±.047	**.795±.047**	.791±.044	.795±.049
breast	.250±.000	.797±.047	.838±.052	.874±.047	.876±.055	**.885±.054**	.889±.056
diabetes	.542±.000	.554±.035	.569±.033	.623±.043	**.648±.039**	.617±.038	.648±.037
bank	.633±.000	.510±.031	.468±.033	.541±.036	**.658±.040**	.657±.043	.663±.037
Japan3v4	.576±.000	.566±.035	.564±.037	.584±.035	**.667±.039**	.665±.037	.668±.037
spam	.302±.000	.637±.036	.467±.028	.723±.045	.746±.049	**.751±.049**	.753±.045
GesDvP	.478±.000	.637±.039	.548±.033	.716±.045	.735±.040	**.740±.046**	.741±.044
har1v2	.232±.000	.706±.045	.707±.047	.806±.048	.804±.049	**.818±.054**	.820±.052
wine	.620±.000	.637±.039	.474±.027	.637±.036	.674±.037	**.688±.046**	.692±.047
collision-det	.743±.000	.772±.047	.764±.044	.784±.052	.792±.051	**.798±.053**	.803±.049
mnist:1v5	.921±.000	.952±.056	.957±.054	.954±.056	**.966±.058**	.964±.059	.969±.061
mnist:2v6	.862±.000	.906±.054	.919±.050	.917±.052	**.922±.049**	.917±.053	.922±.051
mnist	.673±.000	.702±.041	.704±.042	.743±.048	.742±.049	**.745±.043**	.754±.046
f-mnist2v5	.675±.000	.951±.053	.945±.060	.971±.057	.978±.055	**.982±.055**	.982±.059
f-mnist3v4	.632±.000	.808±.049	.793±.044	.819±.048	.865±.050	**.869±.056**	.870±.054
f-mnist7v9	.642±.000	.824±.045	.81±.052	.829±.052	**.876±.050**	.868±.054	.880±.047
f-mnist	.464±.000	.492±.033	.525±.033	.536±.035	.531±.033	**.544±.036**	.546±.040
cifar10:0v5	.296±.000	.502±.033	.347±.026	.485±.036	.678±.046	**.685±.039**	.693±.039
cifar10:0v6	.587±.000	.540±.038	.477±.029	.556±.037	.688±.040	**.692±.046**	.697±.043
cifar10:4v8	.256±.000	.514±.032	.488±.033	.473±.032	.661±.042	**.663±.045**	.664±.037

Table 2. **Adversarial accuracies** (mean ± std error). CoEvoRDT+FPRDT obtained the best results for all datasets. Box denotes that CoEvoRDT+FPRDT is statistically significantly better than all other methods. The best results are **bolded**. Gray background indicates that a given method is statistically significantly better than all other methods (except CoEvoRDT+FPRDT).

Results - repeated runs

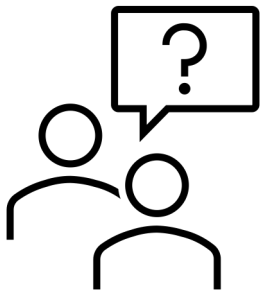
	minimax regret				adversarial accuracy				computation time [s]			
N	N FPRDT	N CoEvoRDT	CoEvoRDT + N FPRDT	N CoEvoRDT + N FPRDT	N FPRDT	N CoEvoRDT	CoEvoRDT + N FPRDT	N CoEvoRDT + N FPRDT	N FPRDT	N CoEvoRDT	CoEvoRDT + N FPRDT	N CoEvoRDT + N FPRDT
1	.304	.238	.237	.237	.531	.544	.546	.546	19	79	97	97
2	.302	.237	.236	.236	.535	.546	.548	.548	40	161	114	185
3	.301	.237	.236	.236	.539	.548	.549	.550	60	240	134	272
4	.300	.236	.236	.235	.545	.550	.552	.553	80	321	165	362
5	.299	.236	.235	.235	.548	.552	.554	.557	99	406	183	496
10	.293	.234	.233	.233	.552	.557	.559	.562	195	774	264	939
20	.284	.230	.230	.229	.558	.564	.566	.568	391	1553	454	1869
50	.282	.229	.229	.227	.563	.568	.568	.569	956	3863	999	4564
100	.282	.228	.229	.227	.566	.568	.568	.569	1921	7981	1990	9026

Table 4. Best results of repeated N algorithms' runs for fashion-mnist dataset. In CoEvoRDT + N FPRDT output DTs from N FPRDT independent runs were incorporated into CoEvoRDT initial population.

Summary

- Novel coevolutionary algorithm for robust decision tree construction
- Adaptable to various target metrics
- Can integrate results from other strong methods
- Outperforming competitors in minimax regret and achieving on-par performance in adversarial accuracy metrics

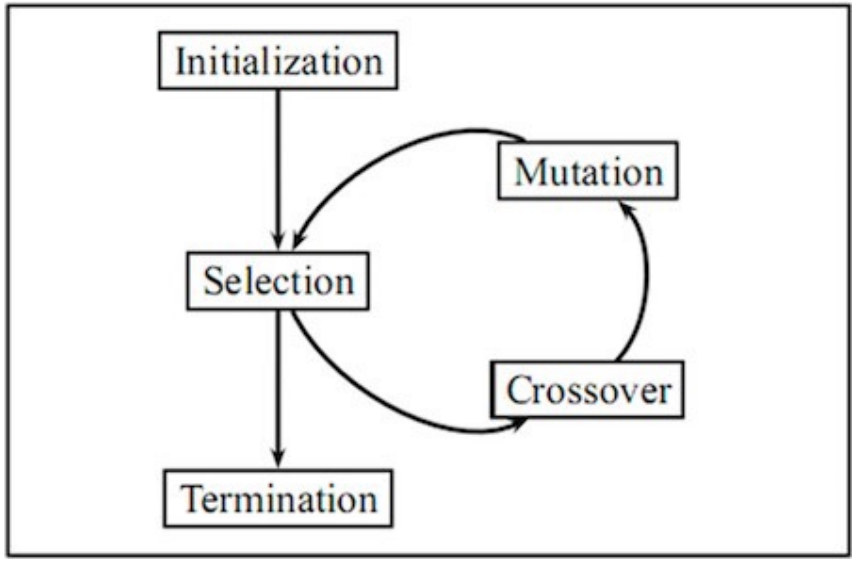
Thank you



Scan for arXiv paper

Evolutionary methods

- Inspired by biological evolution (Darvinism)
- Population of individuals (solutions)
- Individuals evaluation (fitness function)
- New generations created with evolutionary operators: crossover, mutation, selection
- **Coevolution** – multiple populations evolve simultaneously, each adapting in response to the evolving characteristics of the other populations



Paseudocode

Algorithm 1: CoEvoRDT pseudocode.

$P_T \leftarrow$ InitializeDecisionTreesPopulation()
$P_P \leftarrow$ InitializePerturbationsPopulation()
$\text{HoF}_T = \text{HoF}_P = \emptyset$ // HoF - Hall of Fame
$N_{\text{top}} = 20$

while stop condition not satisfied **do**
  for $1..l_c$ **do**
    $P_T \leftarrow P_T \cup \text{Crossover}(P_T)$
    $P_T \leftarrow P_T \cup \text{Mutate}(P_T)$
    $P_T \leftarrow \text{Evaluate}(P_T, P_P, \text{HoF}_P)$
    $P_T^* \leftarrow \text{GetElite}(P_T)$
    while $|P_T^*| < N_T$ **do**
      $P_T^* \leftarrow P_T^* \cup \text{BinaryTournament}(P_T)$
    end while
    $P_T \leftarrow P_T^*$
    $\mathcal{T}, \mathcal{P} \leftarrow \text{MixedNashEquilibrium}(P_T, P_P)$
    $\text{HoF}_T \leftarrow \text{HoF}_T \cup \mathcal{T}$
    $\text{HoF}_P \leftarrow \text{HoF}_P \cup \mathcal{P}$
  end for

  for $1..l_c$ **do**
    $P_P \leftarrow P_P \cup \text{Crossover}(P_P)$
    $P_P \leftarrow P_P \cup \text{Mutate}(P_P)$
    $P_P \leftarrow \text{Evaluate}(P_P, P_T, \text{HoF}_T, N_{\text{top}})$
    $P_P^* \leftarrow \text{GetElite}(P_P)$
    while $|P_P^*| < N_P$ **do**
      $P_P^* \leftarrow P_P^* \cup \text{BinaryTournament}(P_P)$
    end while
    $P_P \leftarrow P_P^*$
    $\mathcal{T}, \mathcal{P} \leftarrow \text{MixedNashEquilibrium}(P_T, P_P)$
    $\text{HoF}_T \leftarrow \text{HoF}_T \cup \mathcal{T}$
    $\text{HoF}_P \leftarrow \text{HoF}_P \cup \mathcal{P}$
  end for
end while

return $\arg\max_{t \in P_T} \xi(t)$

Experimental setup

20 popular benchmark sets.

CoEvoRDT parameters:

- decision tree population size: 200
- perturbation population size: 500
- number of consecutive generations for each population: 20
- crossover probability: 0.8
- mutation probability: 0.5
- Hall of Fame size: 200
- generations limit: 1000

Contribution

- **Novel coevolutionary algorithm for robust decision tree construction.**
- **Adaptable to various target metrics -** suitable for diverse applications, including scenarios combining robustness with other objectives.
- Introduces a game-theoretic approach for constructing the **Hall of Fame using Mixed Nash Equilibrium**, enhancing robustness and convergence speed.
- **Can integrate results from other strong methods** into the initial population for performance improvement.
- **Outperforming competitors in minimax regret** and **achieving on-par performance in adversarial accuracy metrics.**

Benchmarks

dataset	ε	Instances	Features	Classes
ionos	0.2	351	34	2
breast	0.3	683	9	2
diabetes	0.05	768	8	2
bank	0.1	1372	4	2
Japan:3v4	0.1	3087	14	2
spam	0.05	4601	57	2
GesDvP	0.01	4838	32	2
har1v2	0.1	3266	561	2
wine	0.1	6497	11	2
collision-det	0.1	33000	6	2
mnist:1v5	0.3	13866	784	2
mnist:2v6	0.3	13866	784	2
mnist	0.3	70000	784	10
f-mnist:2v5	0.2	14000	784	2
f-mnist:3v4	0.2	14000	784	2
f-mnist:7v9	0.2	14000	784	2
f-mnist	0.2	70000	784	10
cifar10:0v5	0.1	12000	3072	2
cifar10:0v6	0.1	12000	3072	2
cifar10:4v8	0.1	12000	3072	2

Evaluation

Decision trees population

Minimum value of given metric (adversarial accuracy or max regret) against all perturbations for adversarial population (including HoF).

Perturbations population

Balance between perturbation efficiency against all decision trees and avoiding oscillation. N_{top}=20 highest-fitness decision trees are used for perturbation evaluation.

Results – Hall of Fame variants

	minimax regret					adversarial accuracy					computation time [s]				
HoF size	Nash mixed tree	Top K as mixed tree	Nash single trees	Top K	Best	Nash mixed tree	Top K as mixed tree	Nash single trees	Top K	Best	Nash mixed tree	Top K as mixed tree	Nash single trees	Top K	Best
0	.261	.261	.261	.261	.261	.533	.533	.533	.533	.533	47	47	47	47	47
10	.242	.248	.247	.251	.259	.535	.535	.535	.534	.533	50	50	50	50	50
20	.240	.246	.245	.249	.256	.536	.536	.536	.536	.534	55	54	55	56	51
50	.241	.244	.245	.249	.254	.536	.536	.536	.536	.534	61	58	59	62	54
100	.239	.243	.243	.247	.253	.538	.538	.537	.537	.535	68	63	66	65	56
200	.238	.242	.242	.244	.250	.543	.539	.540	.539	.535	77	70	76	77	59
500	.237	.241	.241	.243	.248	.545	.540	.540	.540	.536	86	79	91	90	60
∞	.237	.239	.240	.240	.248	.545	.540	.541	.540	.536	86	77	85	85	61

Table 3. Results with respect of HoF size for *fashion-mnist* dataset. ∞ means that there was no limit on HoF size.

Computation times

dataset	CART	Meta Silvae	RIGDT-h	GROOT	FPRDT	CoEvoRDT	CoEvoRDT+FPRDT
ionos	0	2	1	1	1	2	3
breast	0	2	1	1	1	2	3
diabetes	0	2	1	1	1	3	4
bank	1	5	2	2	2	6	8
Japan3v4	1	8	3	3	3	9	12
spam	1	11	4	4	4	13	17
GesDvP	1	10	4	4	4	11	15
har1v2	1	11	4	4	4	12	15
wine	2	6	6	6	6	2	8
collision-det	9	26	16	18	16	17	35
mnist-1-5	8	14	8	8	8	13	20
mnist-2-6	6	19	8	8	7	24	31
mnist	17	62	22	21	21	68	93
f-mnist2v5	7	20	9	9	8	23	34
f-mnist3v4	7	23	10	9	9	25	36
f-mnist7v9	7	21	10	9	9	26	35
f-mnist	18	60	21	19	19	79	97
cifar10:0v5	21	112	42	39	40	146	191
cifar10:0v6	22	107	44	44	42	126	174
cifar10:4v8	21	106	41	41	41	111	163

Table 1: Comparison of methods' computation times (in seconds).

Parameterization

- decision trees population size
  $N_T : \{10, 20, 50, 100, 200, 500, \mathbf{1000}\}$
- perturbations population size
  $N_P : \{100, 200, 500, 1000, 2000, 5000, \mathbf{10000}\}$
- number of consecutive generations for each population
  $l_c : \{1, 2, 5, 10, \mathbf{20}, 50, 100\}$
- the number of the best individuals from the decision trees population involved in the perturbations evaluation
  $N_{\text{top}} : \{1, 2, 5, 10, \mathbf{20}, 50, 100, 200\}$
- crossover probability
  $p_c : \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, \mathbf{0.8}, 0.9, 1.0\}$
- mutation probability
  $p_m : \{0.0, 0.1, 0.2, 0.3, 0.4, \mathbf{0.5}, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- selection pressure
  $p_s : \{0.5, 0.6, 0.7, 0.8, \mathbf{0.9}, 1.0\}$
- HoF size
  $N_{\text{HoF}} : \{0, 10, 20, 50, 100, 200, \mathbf{500}\}$
- generations without improvement limit
  $l_c : \{5, 10, 20, \mathbf{50}, 100, 200\}$
- generations limit
  $l_g : \{100, 200, 500, \mathbf{1000, 2000, 5000}\}$

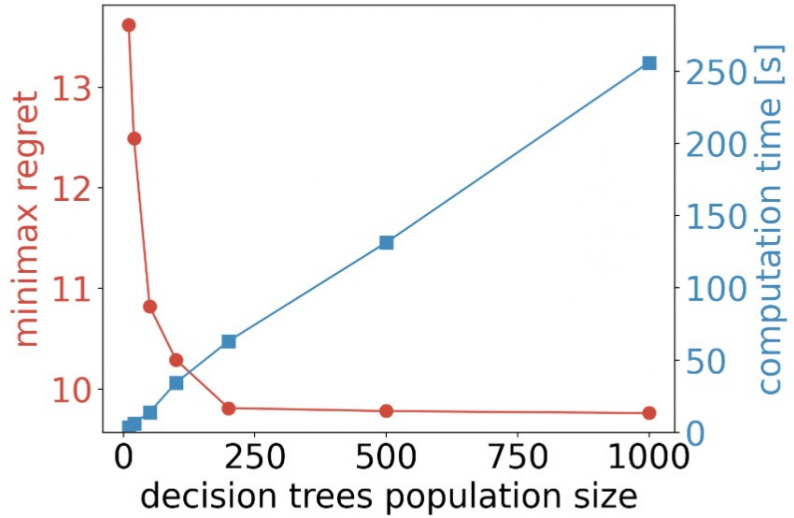


Figure 2: Comparison of minimax regret and computation size for **decision trees population size** values.

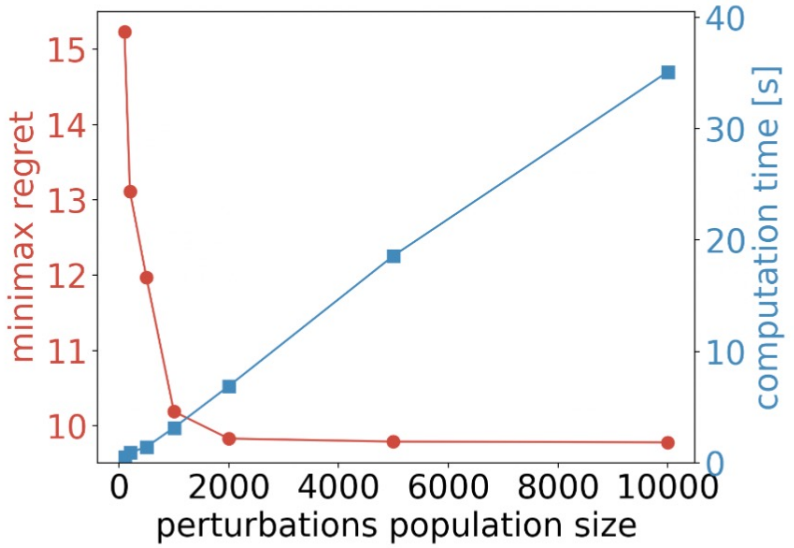


Figure 3: Comparison of minimax regret and computation size for **perturbations population size** values.

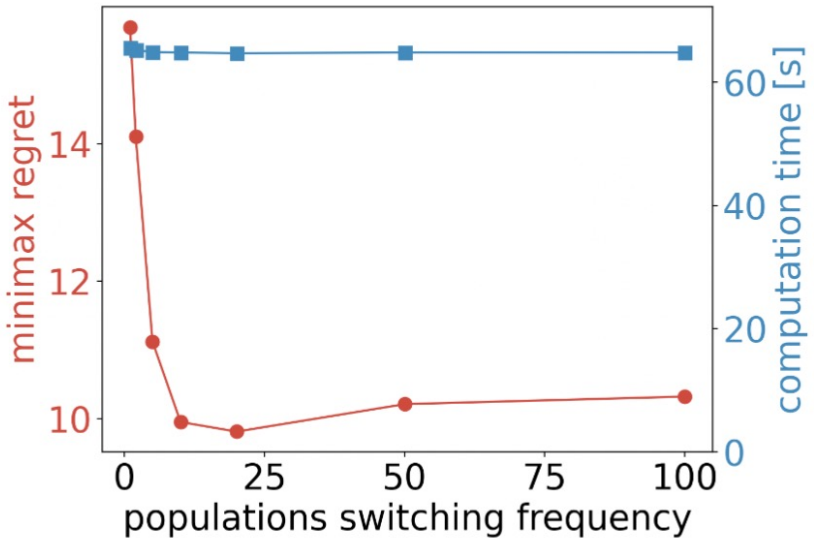


Figure 4: Comparison of minimax regret and computation size for **populations evaluation switching frequency** l_c).

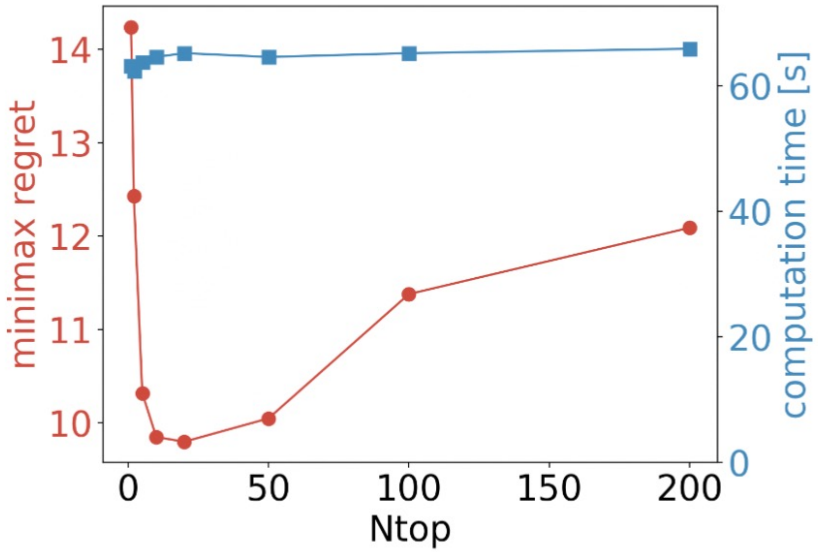


Figure 5: Comparison of minimax regret and computation size for **the number of best individuals used to evaluate the perturbations** (N_{top}).

Random perturbations sample size

	minimax regret		adversarial accuracy	
$log_{10}\|P\|$	average	std error	average	std error
2	0.0964	0.0065	0.740	0.0088
3	0.0968	0.0053	0.737	0.0056
4	0.0972	0.0029	0.733	0.0036
5	0.0976	0.0005	0.729	0.0006
6	0.0977	0.0002	0.728	0.0005
7	0.0977	0.0001	0.728	0.0003

Table 2: Mean value and standard error of adversarial accuracy and minimax regret for different values of the size of random perturbations sample used to their calculation.