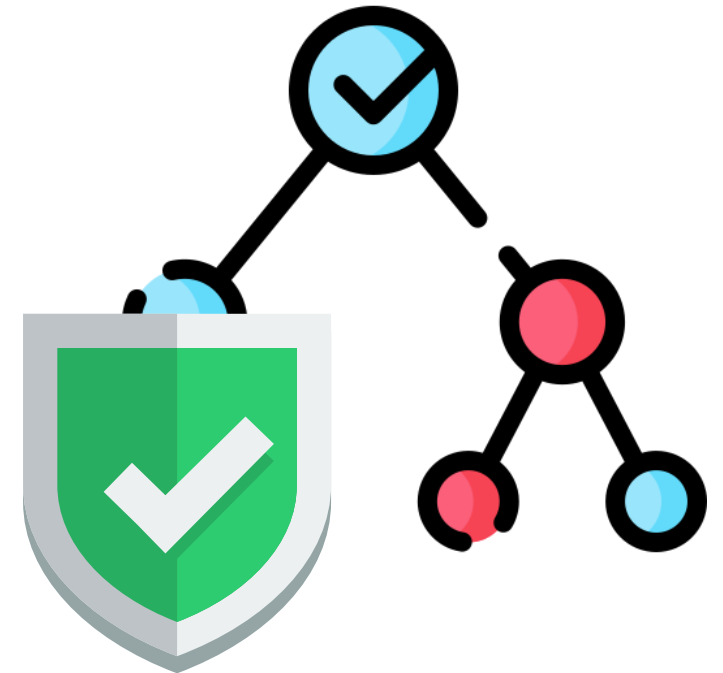


Verifiable Learning for Robust Tree Ensembles

Stefano Calzavara, **Lorenzo Cazzaro**, Giulio Ermanno Pibiri, Nicola Prezza
Università Ca' Foscari Venezia

ACM CCS 2023, Copenhagen, 28/11/2023

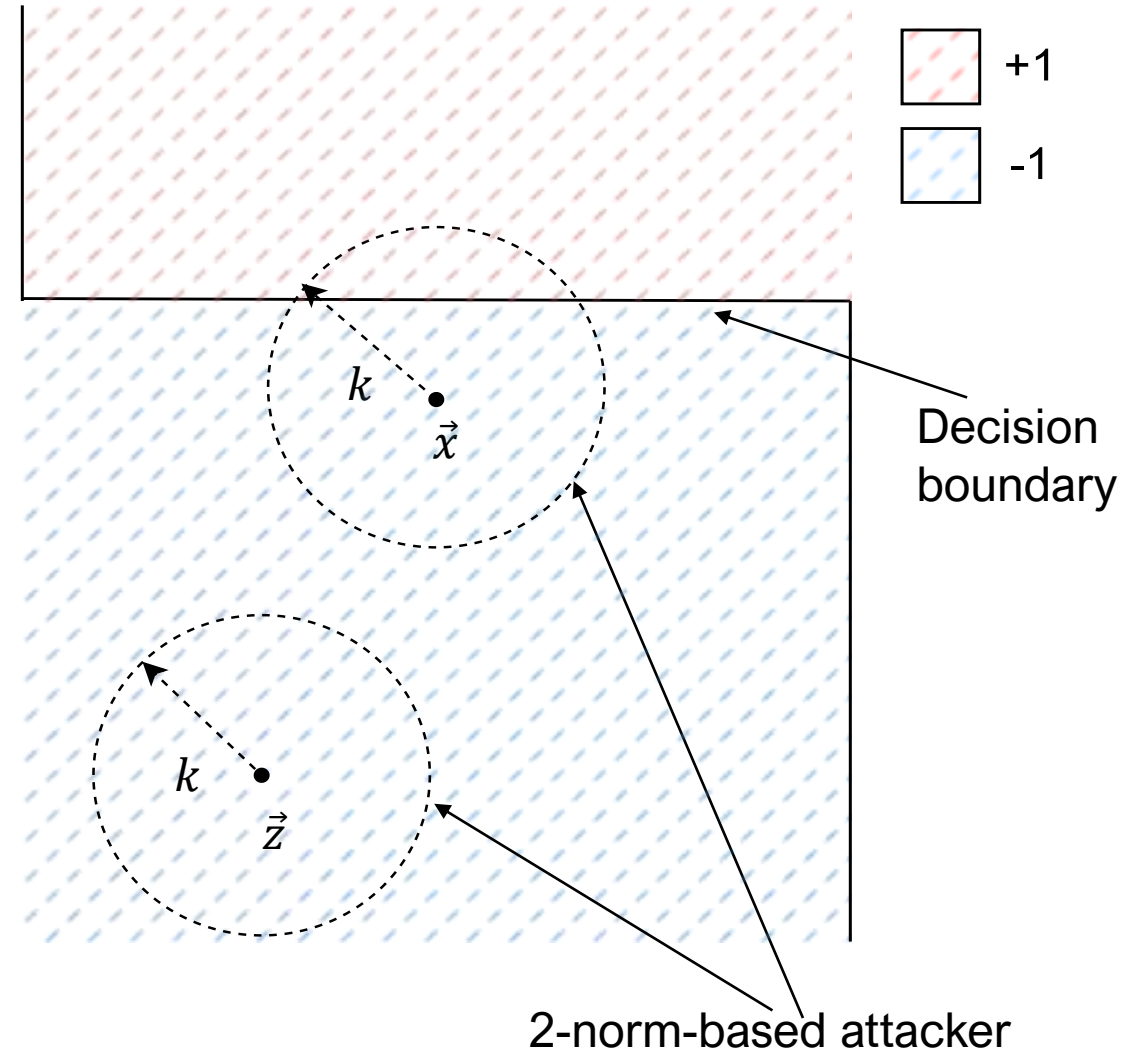


Robustness Verification

Machine Learning (ML) models are vulnerable to **evasion attacks** at test time!

Robustness is estimated as the accuracy under the p -norm-based attacker with maximum perturbation k .

Robustness verification is a well-studied problem both for neural networks and other models like tree ensembles.

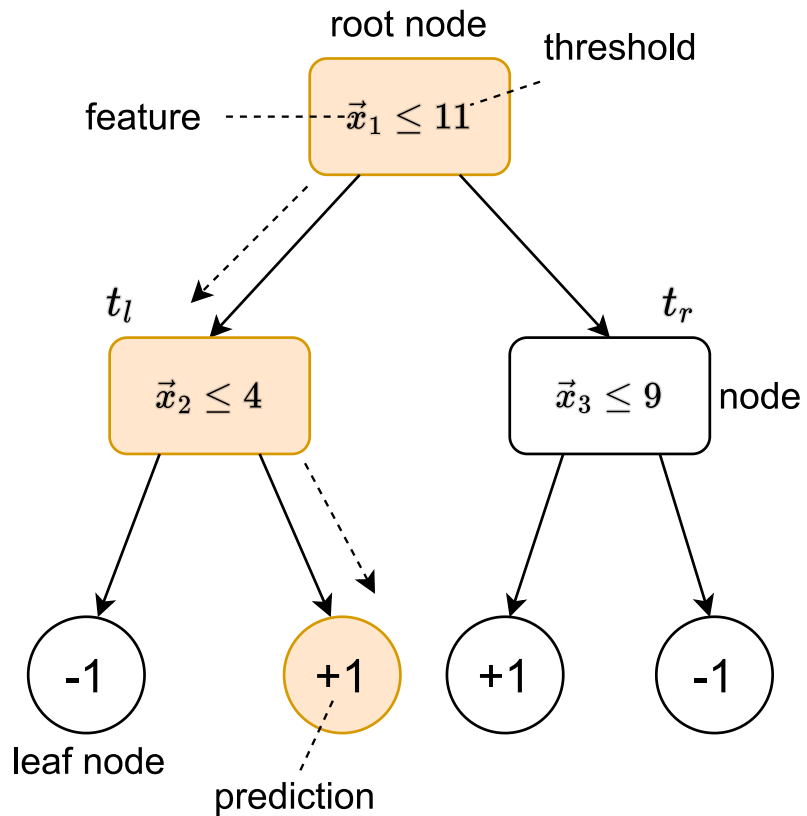


Tree-Based Classifiers

Decision Tree Classifier t

$$\vec{x} = \langle 10.5, 4.5, 17 \rangle$$

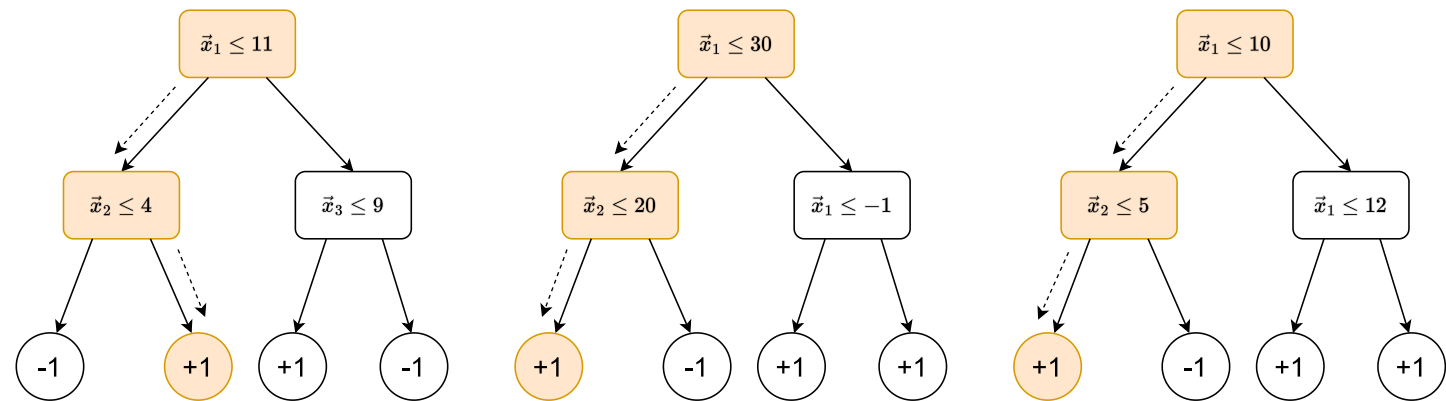
$$y = +1$$



Decision Tree Ensemble $T = \{t_1, t_2, \dots, t_n\}$

$$\vec{x} = \langle 10.5, 4.5, 17 \rangle$$

$$y = +1$$



Ensemble prediction \rightarrow aggregation of the predictions of the single trees.

We consider **majority voting** as aggregation scheme (used by Random Forests).

Robustness Verification is hard!

Complete robustness verification is **hard** for tree ensembles*!

Complexity of robustness verification for p -norm-based attackers.

Model	Complexity
Decision tree	Linear
Tree ensemble	NP-complete

These analyses are **worst case**. Can we find a **restricted class** of tree ensembles enabling efficient security verification against any norm-based attackers?

*Yihan Wang, Huan Zhang, Hongge Chen, Duane S. Boning, and Cho-Jui Hsieh. 2020. On L_p-norm Robustness of Ensemble Decision Stumps and Trees. In ICML

Contribution: Verifiable Learning

We propose ***Verifiable Learning***: rethink training algorithms in order to make the (robustness) verification of the trained model more efficient (also formally).

We instantiate Verifiable Learning for **decision tree ensembles**.

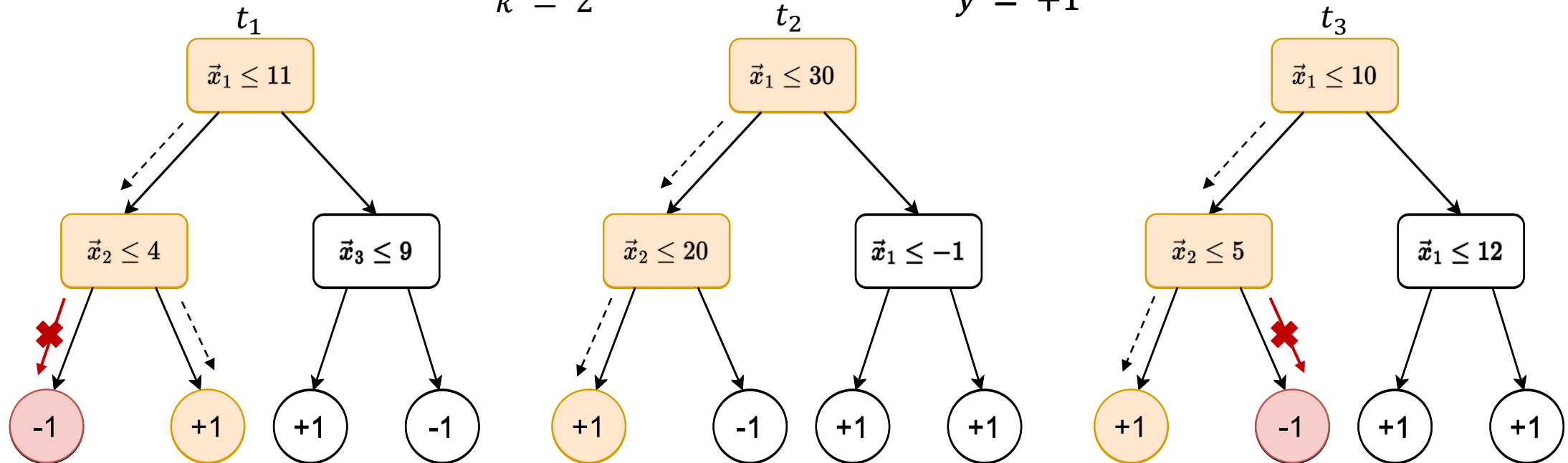
Our contribution consists of 5 parts:

1. We identify what makes the verification problem NP-complete.
2. We restrict the shape of the model in order to avoid the source of the high complexity.
3. We design a (formally proven) efficient verification algorithm for the class of restricted models.
4. We design an (efficient) training algorithm for the class of restricted models.
5. We experimentally verify the effectiveness of our proposal.

Robustness verification of tree ensembles

1-norm-based attacker
 $k = 2$

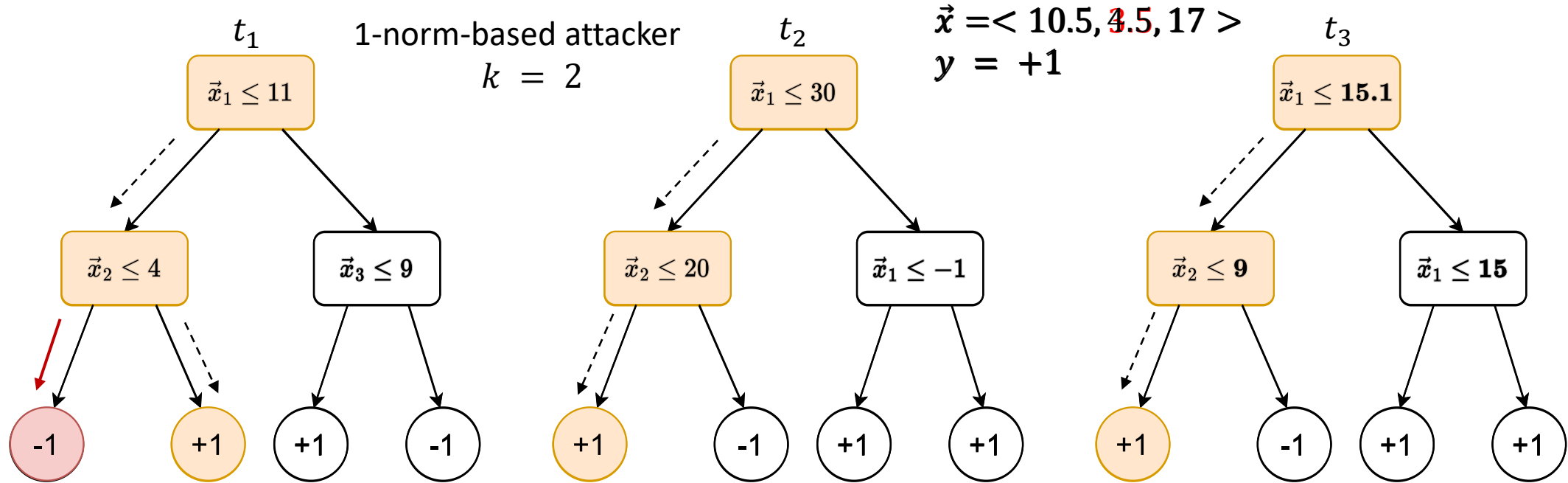
$\vec{x} = \langle 10.5, 4.5, 17 \rangle$
 $y = +1$



Problem: even though it is efficient to verify the robustness of a decision tree, it is not possible to compose the results to make the verification efficient for ensembles.

Step to the solution: if the structure of trees makes **only compatible attacks feasible**, we can **compose the attacks** on the single trees in an efficient way.

Large-spread ensembles



Large-spread condition: any two thresholds for the same feature occurring in two different trees are at a distance of at least $2k$, where k is the maximum adversarial perturbation.

Intuition: if thresholds are sufficiently far away, attacks on different trees **cannot interfere** with each other and can be composed.

Efficient robustness verification

Our verifier CARVE* (suppose that the large-spread ensemble contains m trees):

1. Analyze the m individual trees of the ensemble, using the existing linear time algorithm.
2. If less than $\frac{m}{2} + 1$ trees can be attacked, then no attack on the ensemble is possible (since the aggregation scheme is majority voting).
3. Otherwise, find the $\frac{m}{2} + 1$ trees with the attacks of minimum perturbation: an attack on the ensemble is possible if and only if the sum of these attacks does not exceed the maximum adversarial perturbation k .

Theorem: robustness can be verified in polynomial time for large-spread tree ensembles for any norm-based attackers.

*CARVE - CompositionAI Robustness Verifier for tree Ensembles

Training large-spread ensembles with LSE

The training algorithm LSE* is based on mutation and pruning:

1. Train a traditional forest T of $\gg m$ trees and initialize the large-spread ensemble E with a random tree from T .
2. Iterate for $m - 1$ rounds:
 - A. Pick the tree t in T that minimizes the overlaps with E .
 - B. Fix the overlaps of t with E by perturbing the thresholds of t and E that overlap (mutation).
 - C. Extend E with t (if all the overlaps have been fixed).
3. Return E (m trees out of $\gg m$ the trees in T if LSE succeeds in building the entire large-spread ensemble \rightarrow pruning).

*LSE - Large-Spread Ensemble

Experimental Evaluation

We implemented our verifier CARVE in C++ and our training algorithm LSE in Python (both publicly available on Github!).

Research questions:

1. Can we train a large-spread ensemble with the proposed algorithm?
2. What are the accuracy and the robustness of large-spread ensembles?
3. What is the benefit of the large-spread condition in terms of verification time and memory consumption over a state-of-the-art complete verifier (SILVA*)?

*Francesco Ranzato and Marco Zanella. 2020. Abstract Interpretation of Decision Tree Ensemble Classifiers. In AAAI.

Performance of Large-Spread Ensembles

We are using an ∞ -norm-based-attacker

Verified using SILVA Verified using CARVE

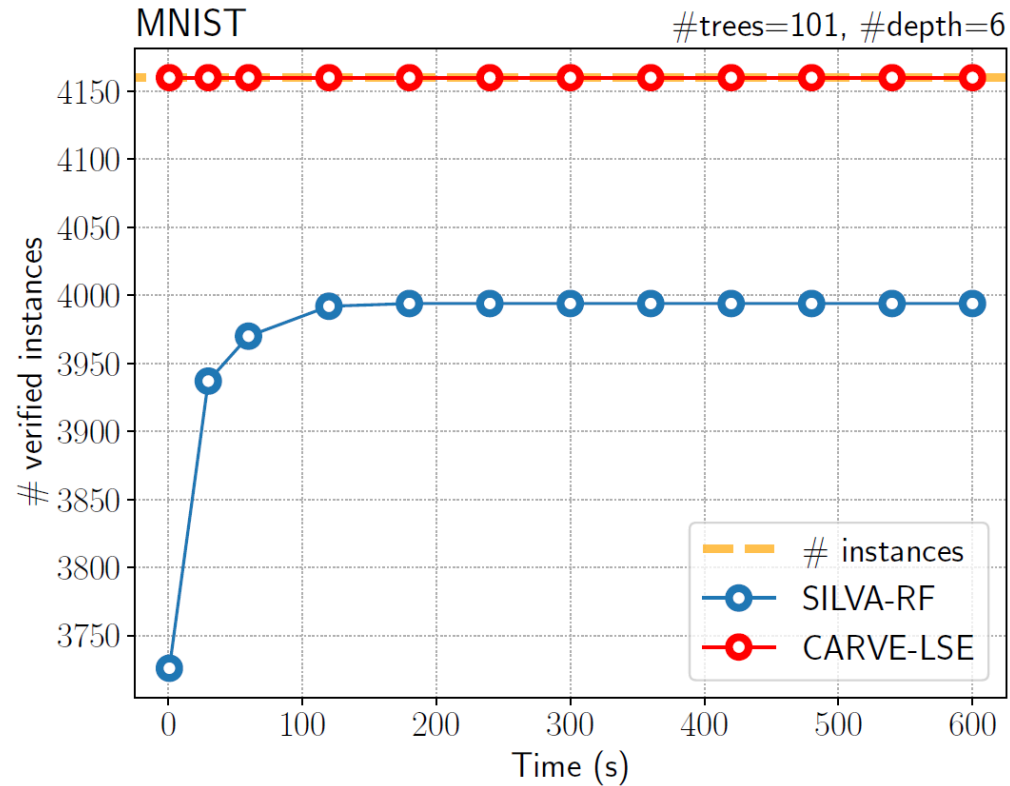
Dataset	k	Trees	Depth	Accuracy		Robustness	
				Traditional	Large-Spread	Traditional	Large-Spread
MNIST	0.0050	25	4	0.97	0.97	0.90	0.96
		101	6	0.99	0.99	0.94	0.97
	0.0100	25	4	0.97	0.97	0.72	0.90
		101	6	0.99	0.99	0.77 ± 0.02	0.97
	0.0150	25	4	0.97	0.97	0.64	0.83
		101	6	0.99	0.99	0.67 ± 0.05	0.94
Webspam	0.0002	25	4	0.90	0.90	0.83	0.87
		101	6	0.94	0.91	0.88	0.90
	0.0004	25	4	0.90	0.89	0.80	0.86
		101	6	0.94	0.89	0.85	0.86
	0.0006	25	4	0.90	0.89	0.78	0.85
		101	6	0.94	0.85	0.81	0.82

1. Large-Spread Ensembles are more robust.
2. SILVA may be forced to approximate the robustness.

Reasonable accuracy

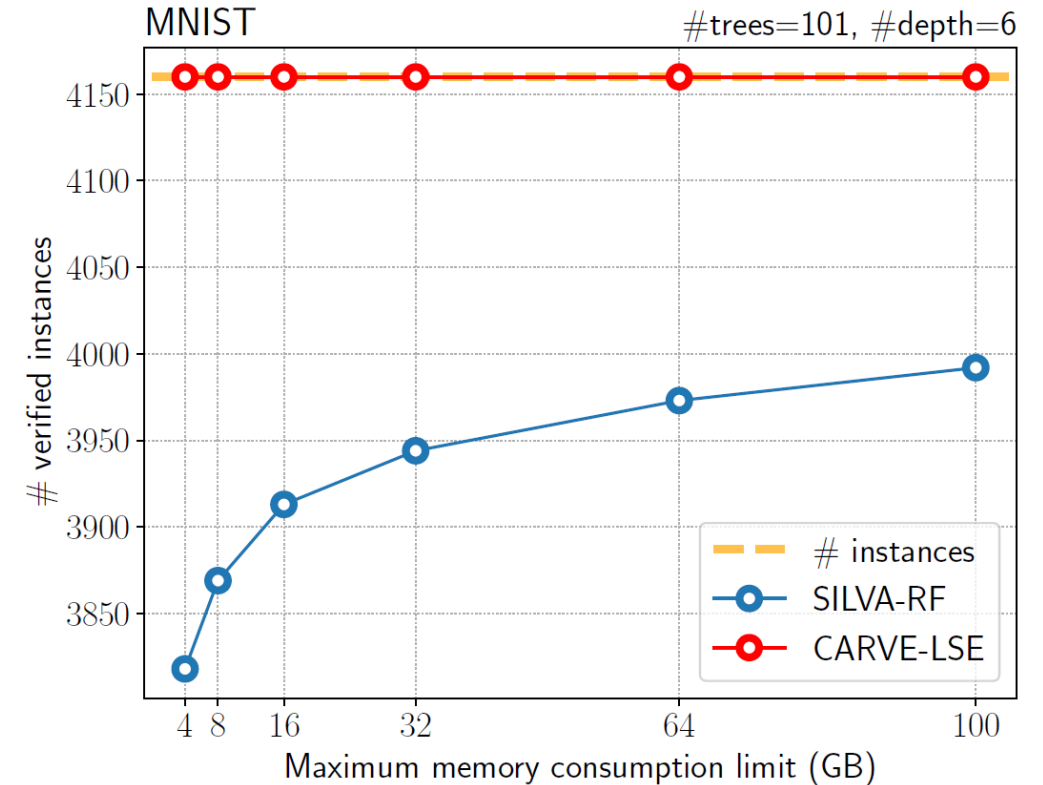
Efficiency of CARVE

Time



CARVE requires less than one second per instance
VS
SILVA may not verify some instances even in 10 minutes!

Memory



CARVE requires less than 4GB RAM per instance
VS
SILVA may not verify some instances even with 100GB RAM!


Take-Home Messages

1. Verifiable Learning: rethink traditional learning algorithms to make (robustness) verification of the trained model feasible.
2. The large-spread condition applied to tree-based classifiers enables complete robustness verification in poly time (NP-hard problem in general).
3. Our pruning algorithm fixes the thresholds of a traditional decision tree ensemble to enforce the large-spread condition (with a «reasonable» efficiency).
4. Large-spread ensembles sacrifice a limited amount of the predictive power but their robustness is normally higher and much more efficient to verify.

Lorenzo Cazzaro
Ph.D. student in Computer Science

 @LorenzoCazz

 lorenzo.cazzaro@unive.it

 Lorenzo Cazzaro

 LorenzoCazzaro

 <https://lorenzocazzaro.github.io/>



Ca' Foscari
University
of Venice



Thank you! Questions?

LSE efficiency

