Exploiting diversity for improved adversarial robustness of ML-based NIDS

António Casimiro, LASIGE, University of Lisbon, Portugal

Work done with:

Allan Espindola (FCUL, Portugal and PUCPR, Brazil)

Pedro Ferreira (FCUL, Portugal)

Altair Santin and Eduardo Viegas (PUCPR, Brazil)

PUCPR GRUPO MARISTA ESCOLA POLITÉCNICA LASIGE driven by excellence Ciências ULisboa fct Fundação para a Ciência e a Tecnologia

Attacks to ML-based NIDS

- **Motivated attackers will try to defeat ML-based NIDS**
- **They will craft attacks to one or multiple parts of the ML pipeline**
  - Using Adversarial Machine Learning (AML) techniques
- **There are several kinds of attacks**
  - From poisoning training data
  - To directly changing model parameters
  - Or adding noise to input data, to evade detection
- **The objective is to force the model to produce a wrong result, preferably in a controlled way**
- **In the case of NIDS, the objective of AML is to evade detection**
  - Allowing network attacks to be done without being detected

Our overarching goal

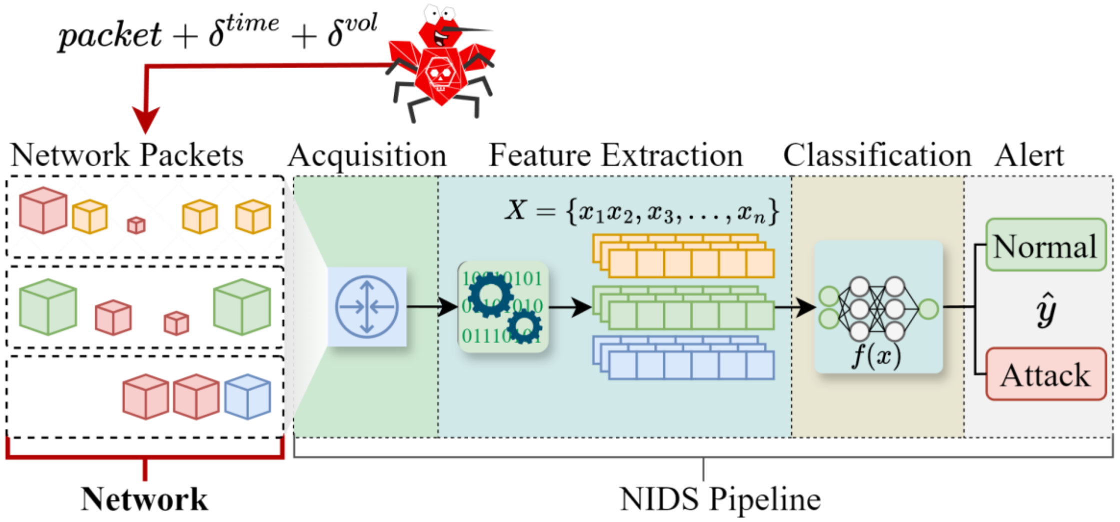
Improve the resilience of ML-based NIDS to Adversarial Machine Learning

Main idea:

Use multiple replicas exploiting multiple forms of diversity to achieve the goal

How can Adversarial Evasion be done in practice?

- **Adding perturbations to network packets or to flows as a whole**
  - Packet-based attacks changing e.g. **payload size** (**volume**), packet **interarrival** (**time**)
  - Indirect implications on extracted features
  - Does not require access to the internal ML pipeline
  - Practically exploitable, as attacker is the one who crafts the attack traffic



Diversity-based approach

- **Inspired on techniques for the development of fault-tolerant and secure systems**
  - Replication
  - Diversity of replicas
- **Exploit multiple forms of diversity**
  - Model diversity
  - Feature diversity
  - Combinations of both model and feature diversity in model ensembles

- **Challenges**
  - Which models, which features, which combinations?
  - How to combine possibly several model outputs?
  - How to show effectiveness?

Which models, which features, which combinations?

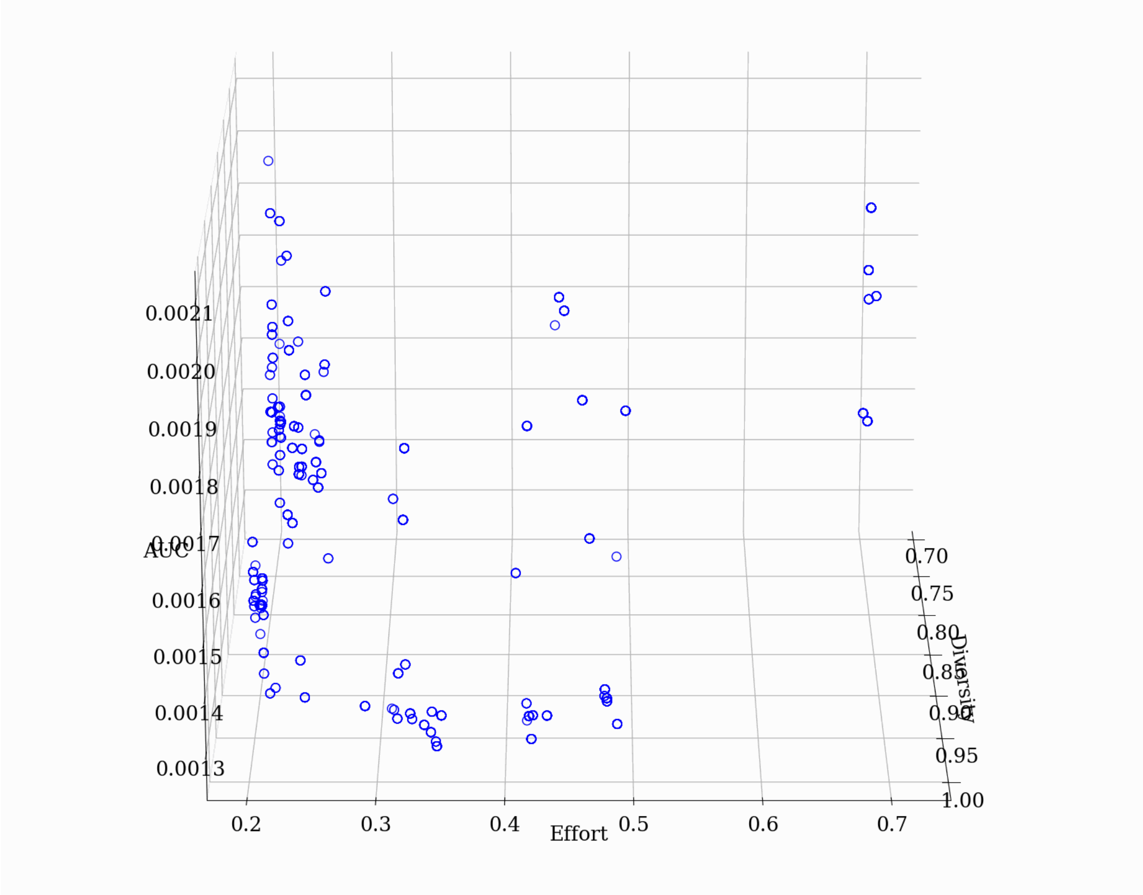
- **Considered 5 models: DT, RF, XGB, MLP, TB**
  - Rationale: **different structures**, responding differently to each attack
- **Applied feature selection processes,** to find better feature combinations for each model
  - Rationale: fine-tune model performance while obtaining multiple solutions in which **different combinations of features** lead to similar performance, but are exposed differently to each attack
- **Use a genetic algorithm (GA) to search the large space of model ensembles to find suitable solutions**
  - Rationale: using model ensembles provides **redundancy**, but it is important that models in the ensemble are **diverse**, to make the whole set more resilient to attacks

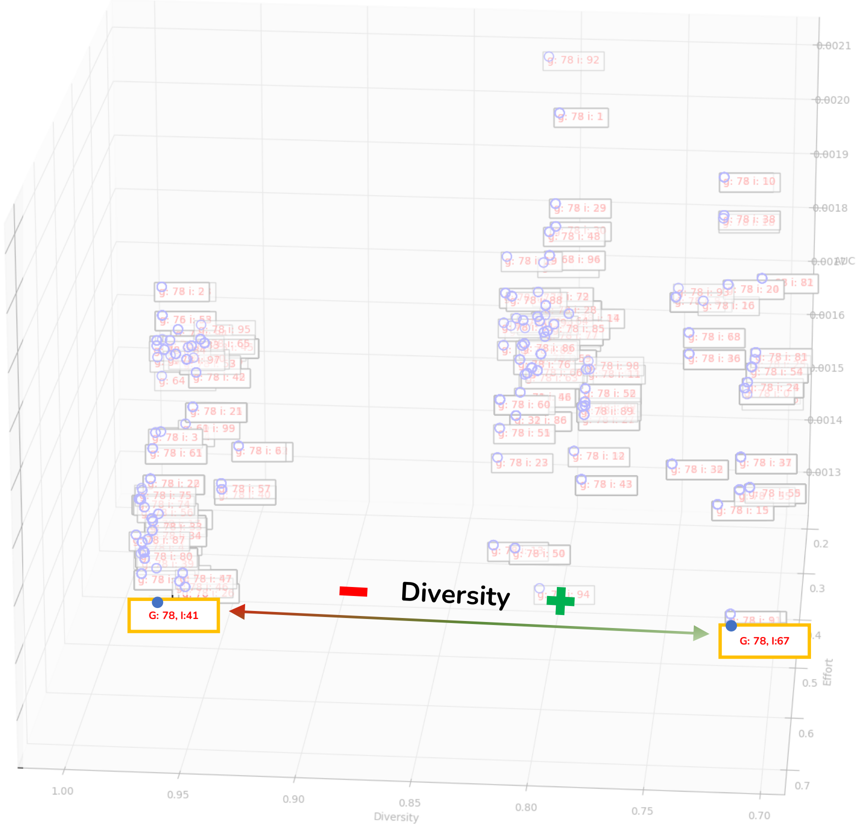
Optimizing feature selection and model architectures using NSGA-II

- **Feature Selection**
  - Evolutionary selection throughout of NSGA-II
  - Minimum: 5 Features, Maximum: 49 features
- **Model Architecture**
  - Neural Networks: number of neurons and number of hidden layers
  - Decision Trees: number of estimators and tree depth
- **Optimization Objectives**
  - **Ensemble Precision**: Measured by AUC (Area Under the Curve)
  - **Ensemble Diversity**: Measured by Disagreement
  - **Model Effort (cost):**
    - Neural Networks: number of neurons and hidden layers
    - Trees: The number of trees and nodes

The Pareto Fronts over generations: manual ensemble selection

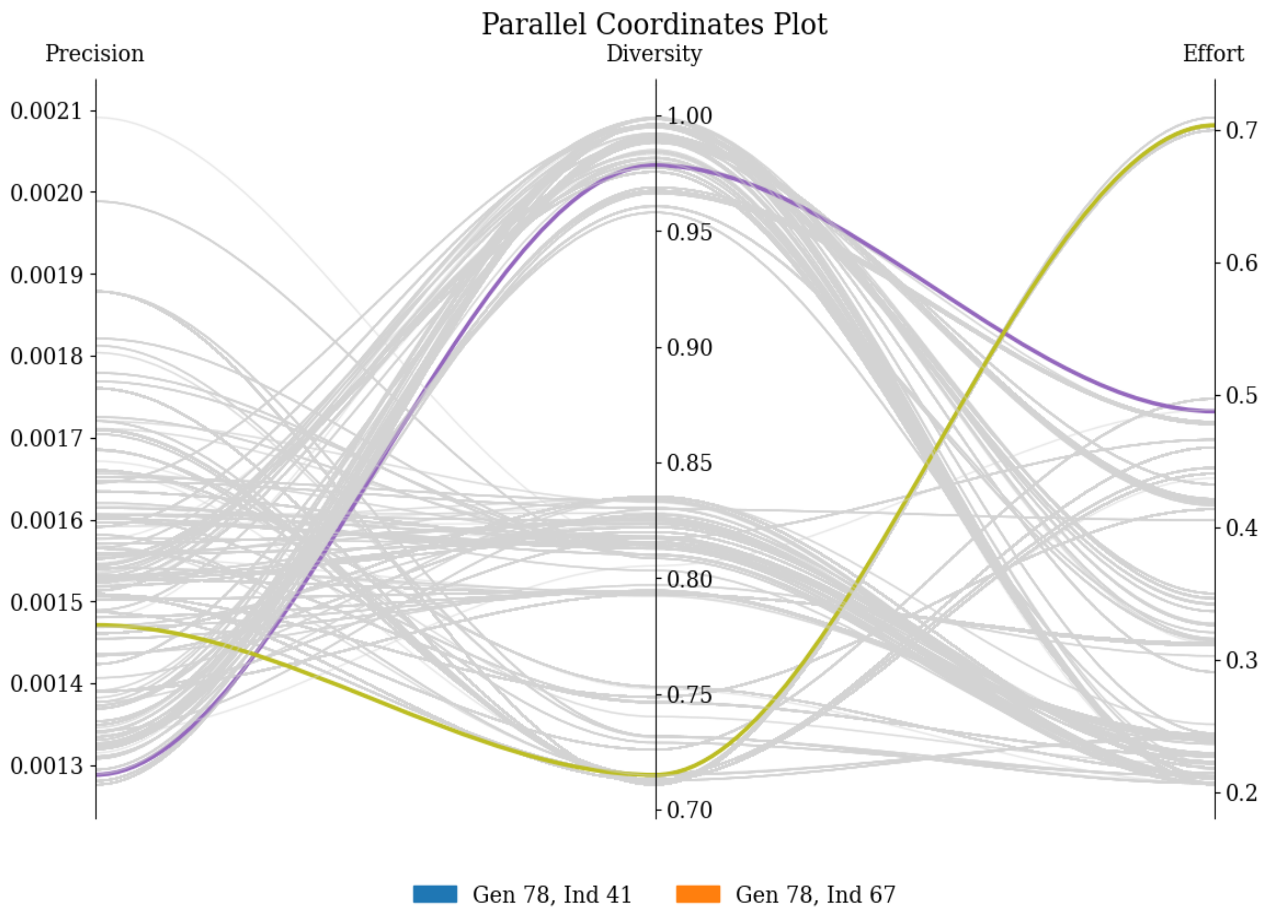
127 ensembles - points





PUCPR ESCOLA POLITÉCNICA LASIGE driven by excellence Ciências ULisboa fct Fundação para a Ciência e a Tecnologia

Multi-objective evaluation



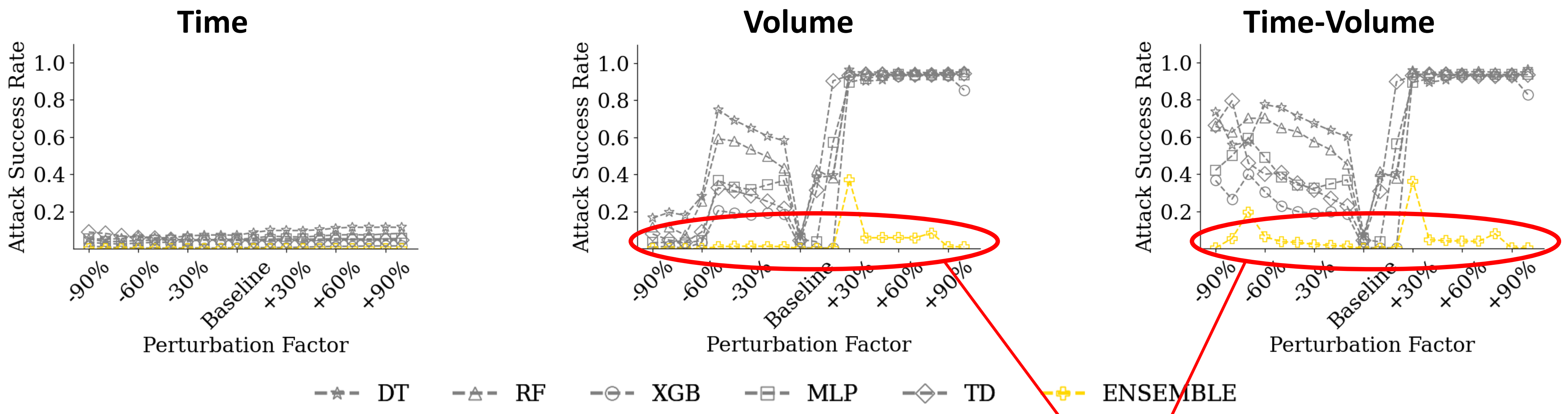
How to combine possibly several model outputs?

- **A few different approaches are possible**
  - Majority-vote
    - Conservative approach
    - Assumes that most models will output the correct decision (attacker can only compromise a minority of models)
    - Decide according to the response (Attack/No-Attack) that gathers more votes
    - Requires odd number of models in the ensemble
  - Any-vote
    - Aggressive approach
    - Assumes powerful attackers, but that at least one model will resist the attack
    - It is sufficient for a single model to output Attack to decide Attack
    - Works for any number of models in the ensemble
  - Averaged outputs
  - Complex (ML-based) combinatorial output

Impact of packet-based attacks

Individual 41

Diversity-based approach: Combination using Any-Vote

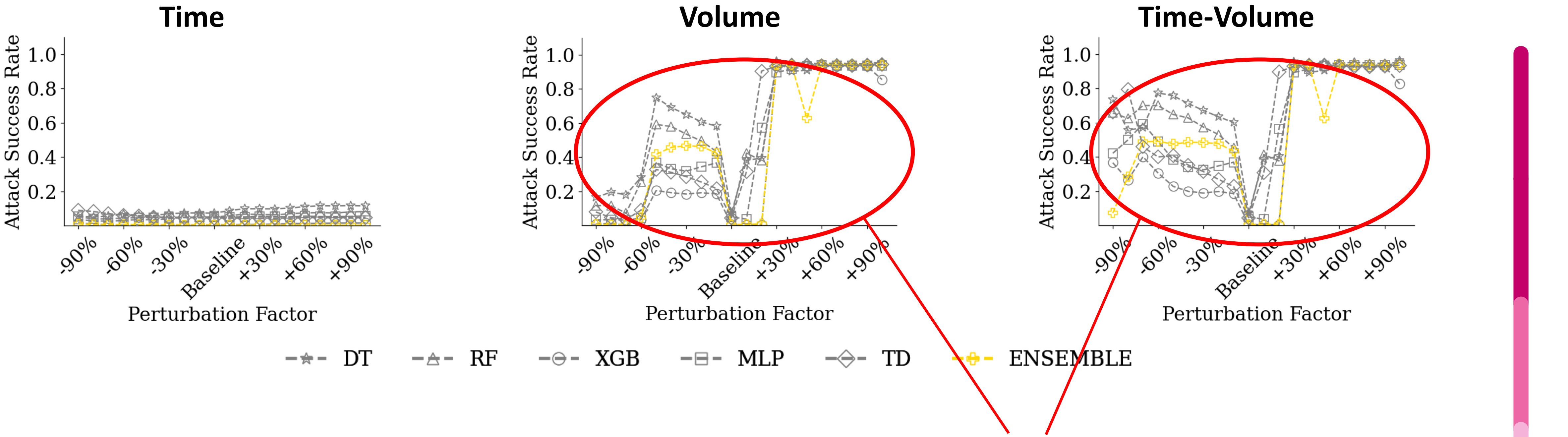


Evasion is no longer successsful

Impact of packet-based attacks

Individual 41

Diversity-based approach: Combination using Majority-Vote



Combining emsemble outputs with majority vote does not improve reslience

Conclusions

- Using diverse models and combining their results with a Any-vote approach allows for improved resilience to realistic AML attacks

- There are still many open issues to be addressed

Thank you for your attention!

Questions?

Contact: casim@ciencias.ulisboa.pt