



UNIVERSITÉ
LIBRE
DE BRUXELLES

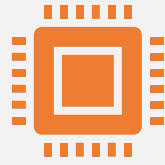
PhD student : DEBICHA Islam
Email: debichasislam@gmail.com

Efficient Intrusion Detection Using Evidence Theory

Islam Debicha, Thibault Debatty, Wim Mees, Jean-Michel Dricot



About the presenter



Currently doing a joint PhD between ERM and ULB about Intrusion detection.



Worked before as a network security engineer.



subjects of interest: machine learning & network security.

Introduction



Intrusion & Intrusion Detection

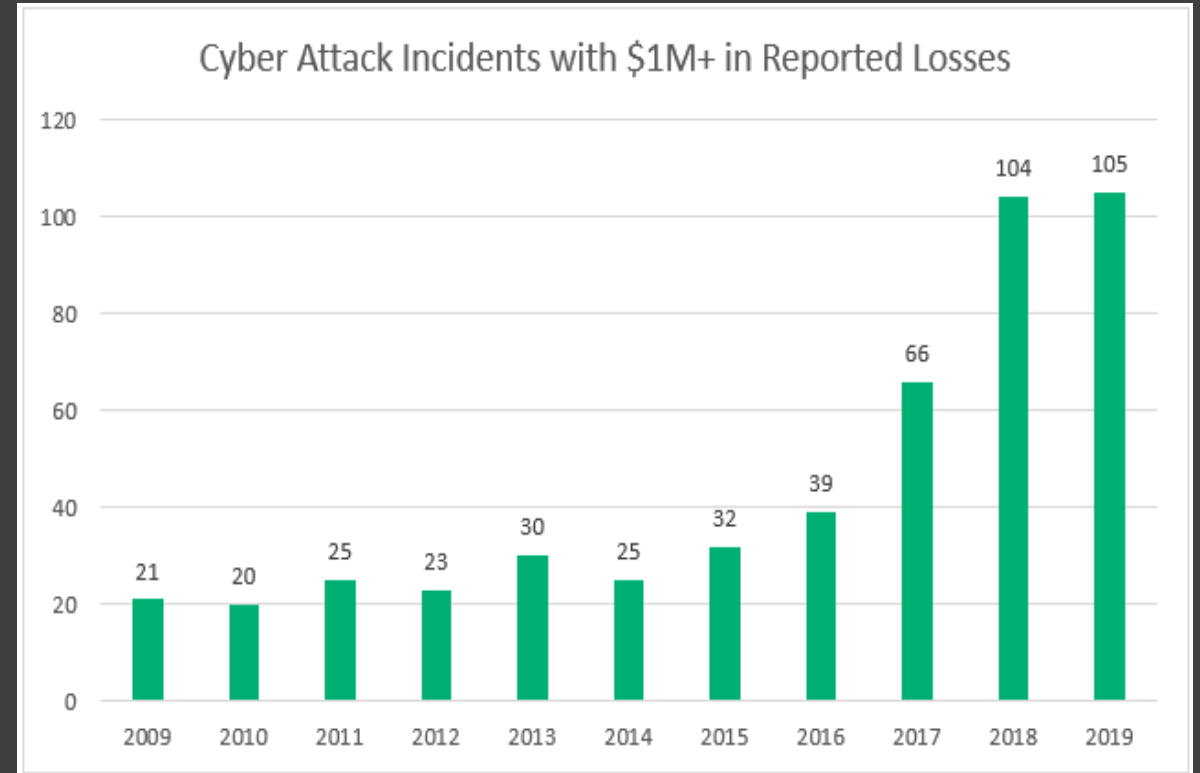
- **Intrusion** : Attempting to break into or misuse a system or a network.
- **Intrusion Detection Systems (IDS)** : look for attack signatures, which are specific patterns that usually indicate malicious or suspicious intent.

Motivation

In today's security product market, there are many intrusion detection systems available.

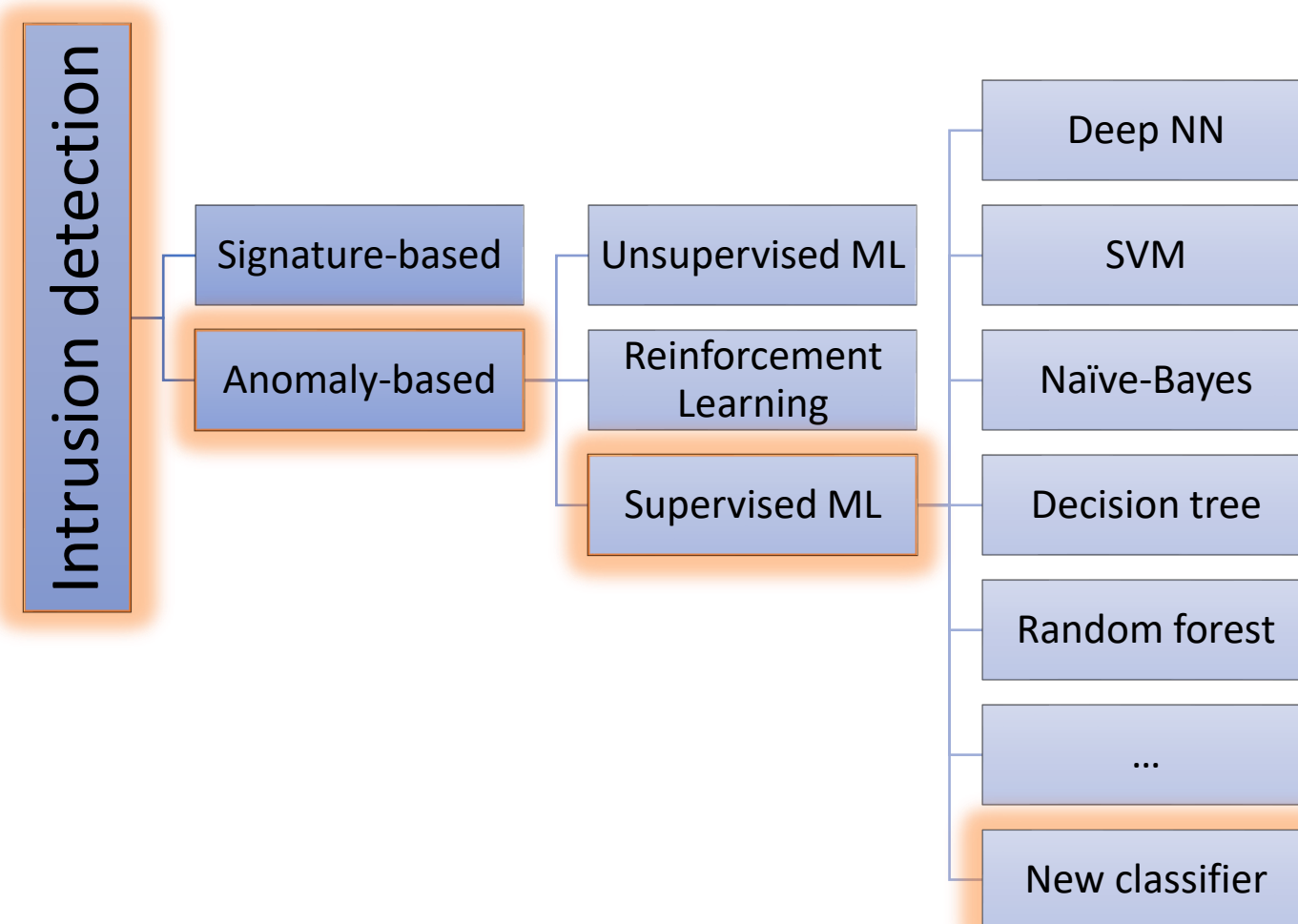
Nevertheless, cyber-attack incidents are on the rise.

For this reason, improvements to these intrusion detection systems are urgently needed.



Source: Center for Strategic & International Studies (CSIS)

Intrusion detection techniques



Theoretical framework

Evidence theory

- **Evidence theory** or **Dempster–Shafer theory (DST)**, is a general framework for **reasoning with uncertainty**, with understood connections to other frameworks such as probability.
- In general, belief functions are used as a way to model uncertainty where imprecision, or lack of knowledge has to be **modelled explicitly**.
- Rather than reasoning on the hypothesis-set alone, it takes into account **all subsets of the hypothesis-set**.

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_n\} \quad \Rightarrow \quad 2^\Theta = \{\emptyset, \{\theta_1\}, \{\theta_2\}, \{\theta_3\}, \{\theta_1, \theta_2\}, \{\theta_1, \theta_3\}, \{\theta_1, \theta_2, \theta_3\}, \dots, \Theta\}$$

- And instead of assigning probabilities, we assign masses to all subsets of the hypothesis-set such that:

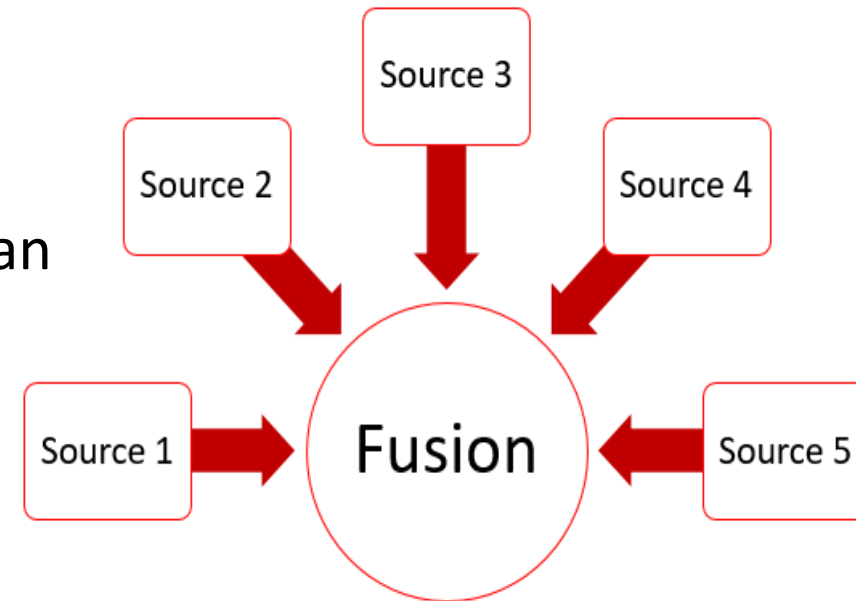
$$m : 2^\Theta \longrightarrow [0, 1]$$

$$\sum_{A \in 2^\Theta} m(A) = 1$$

Combination of evidence

- An important aspect of this theory is the **combination of evidence** obtained from **multiple sources** and the modeling of conflict between them.
- It worth mentioning that **evidential fusion rules** are an **active research topic**.
- One popular rule is Dempster's rule of combination:

$$M(A) = (M_1 \oplus M_2)(A) \propto \sum_{B_1 \cap B_2 = A} M_1(B_1)M_2(B_2)$$

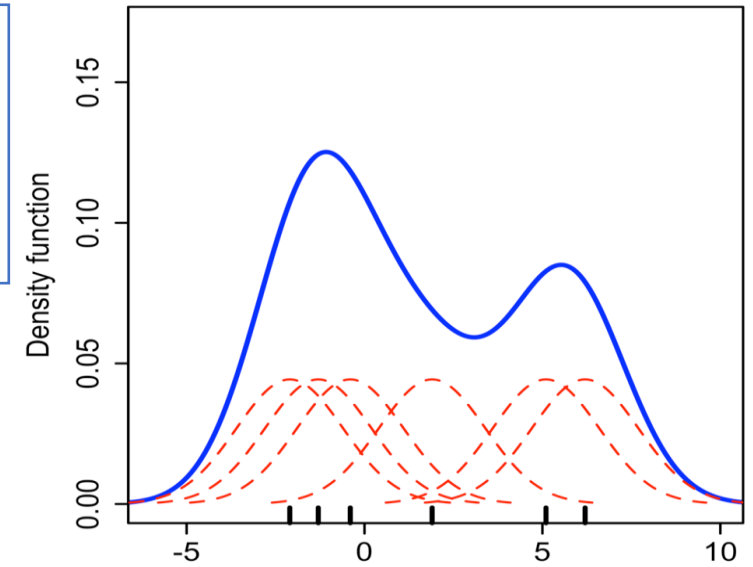


Parzen-Rosenblatt density estimation

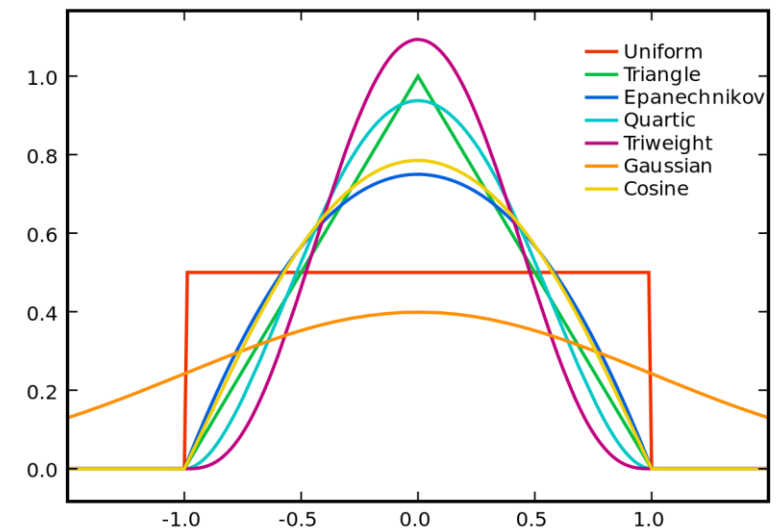
- It's a non-parametric way to estimate the probability density function of a random variable.
- Kernel density estimator of a function f is :

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right)$$

where $K()$ is the **kernel** (a zero-mean non-negative function that integrates to one)
And $h > 0$ is a smoothing parameter known as “kernel width”.



kernel density estimate example



Kernel functions in common use

Evidential discounting methods

Classical discounting:

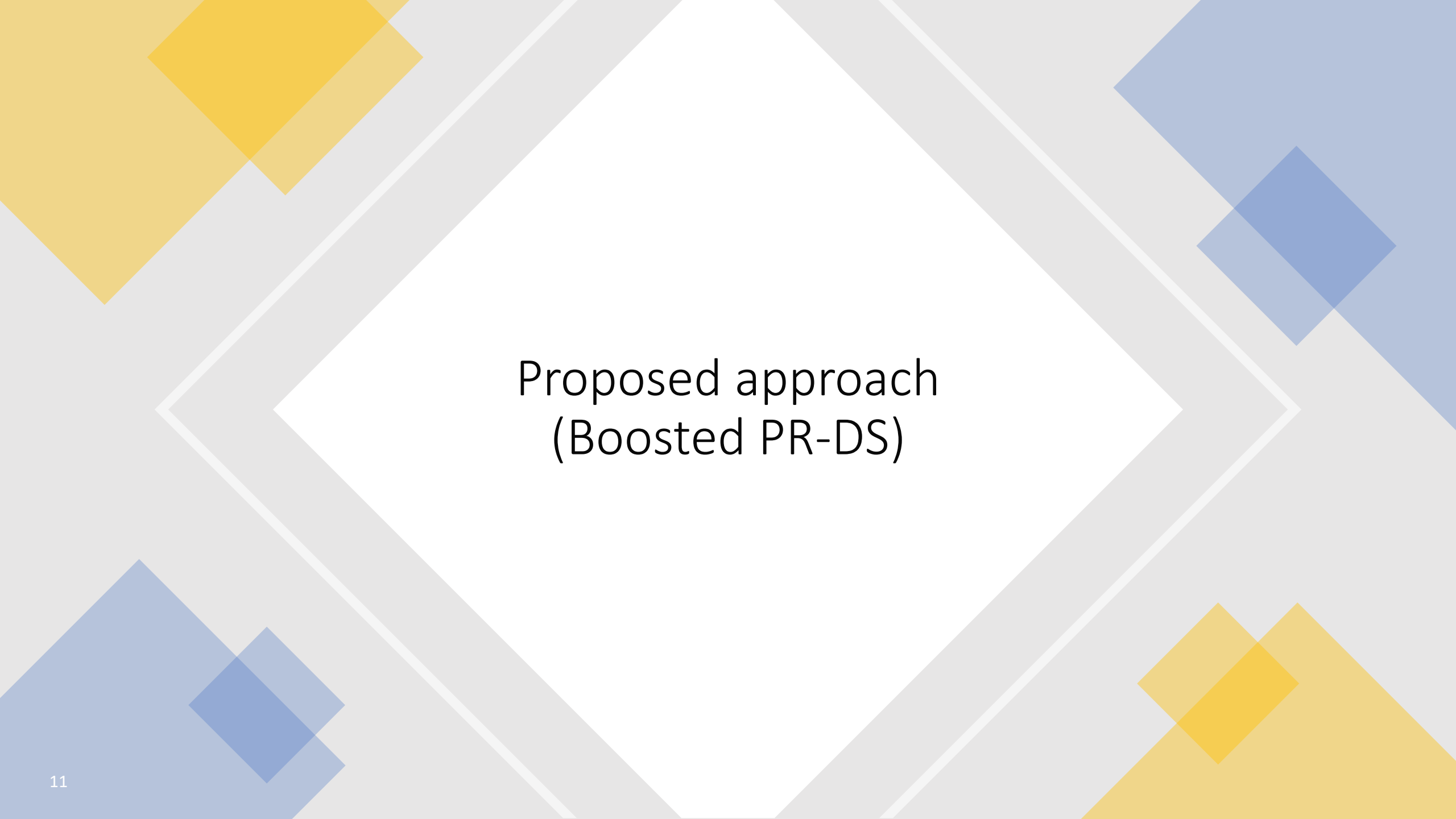
$$\begin{cases} m'^s(A) = \alpha^s \cdot m^s(A) \\ m'^s(\Omega) = (1 - \alpha^s) + \alpha^s \cdot m(\Omega) \end{cases} \quad \forall A \in 2^\Omega, A \neq \Omega$$

α^s is the weakening coefficient of the s^{th} source.

Contextual discounting:

$$\begin{cases} m'^s(A) = \alpha_A^s m^s(A) & A \in \{2^\Omega / \Omega\} \\ m'^s(\Omega) = m^s(\Omega) + \sum_{A \in \{2^\Omega / \Omega\}} (1 - \alpha_A^s) m^s(A) \end{cases}$$

α_A^s is the weakening coefficient of hypothesis **A** for the s^{th} source.



Proposed approach
(Boosted PR-DS)

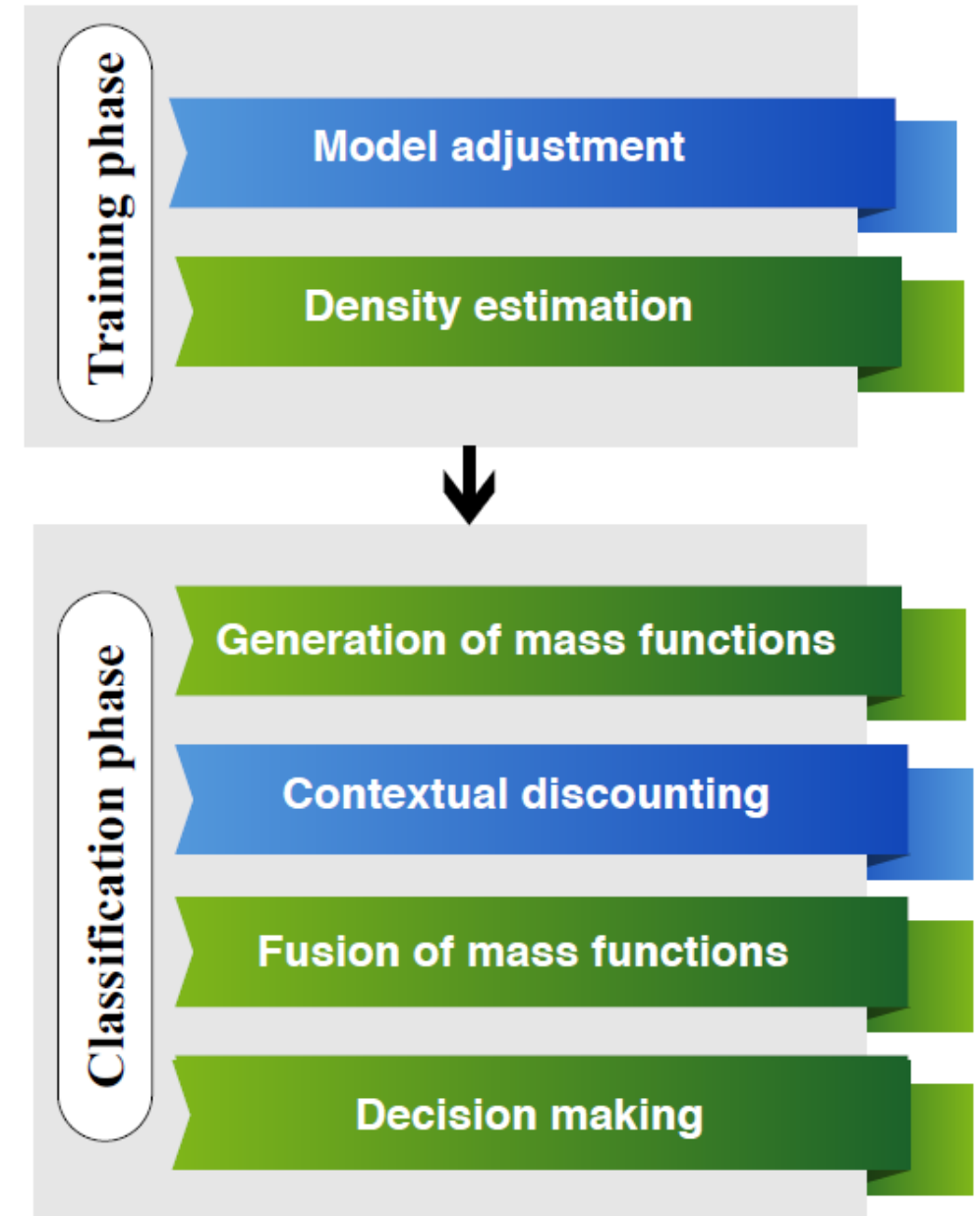
The proposed approach (Boosted PR-DS)

A. **Training phase** involves two steps:

1. **Model adjustment:**

- a) Determining the optimal kernel and fusion rule for the data using basic PR-DS.
- b) Computing of weakening coefficients for each hypothesis using F-score.

2. **Density estimation** where the previously chosen kernel is used to estimate the Probability Density Functions (PDF) of each class for all sources using **Parzen-Rosenblatt** window method .



The proposed approach (Boosted PR-DS)

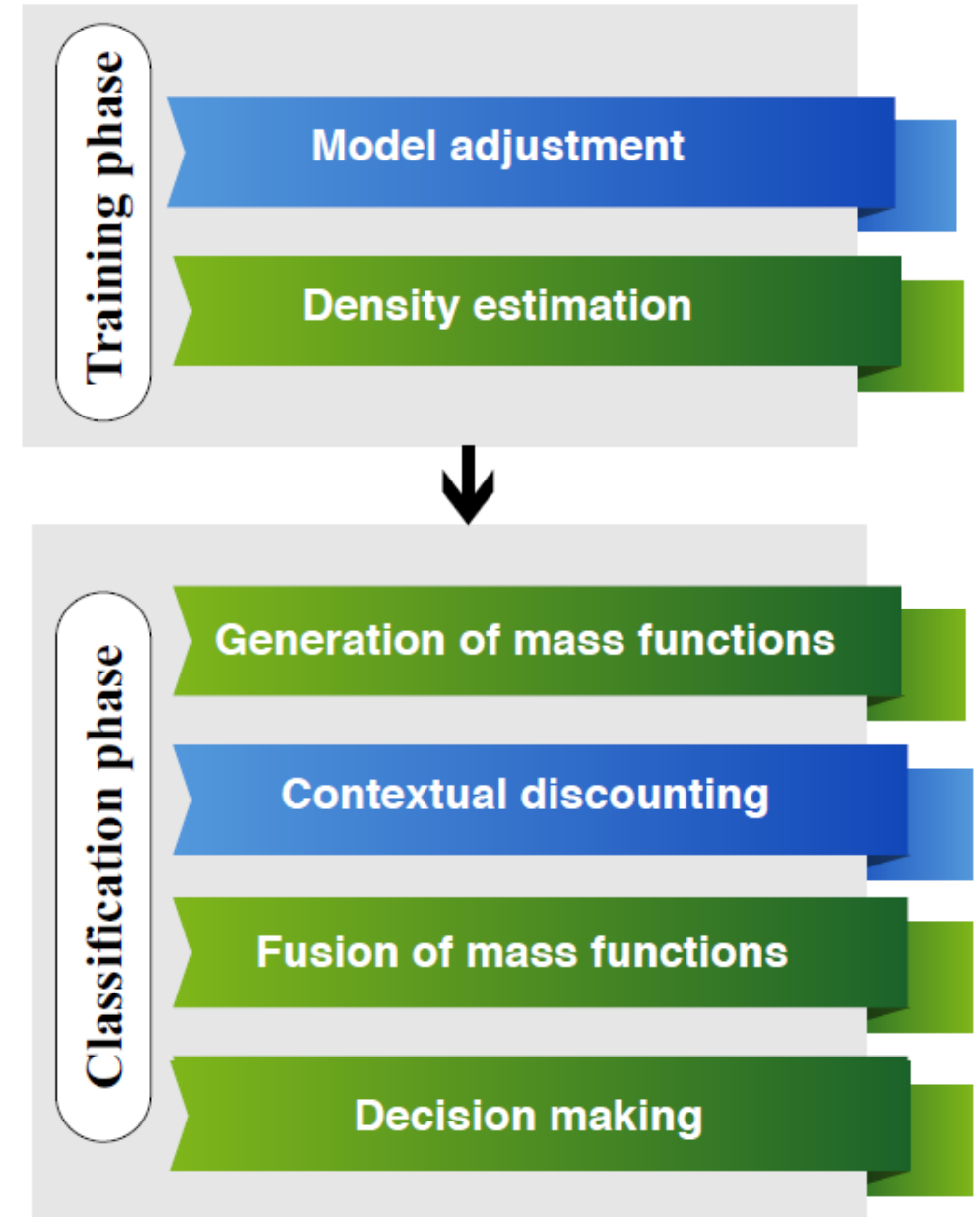
B. **Classification phase** involves four steps:

1. **Generation of mass function:** a mass function M^p for each source is constructed based on the estimated densities.

$$\begin{cases} M^p(\Omega) \propto \hat{f}_{\delta_p(1)}^p(Y_{n'}^p) \\ M^p(\{\omega_{\delta_p(k)}, \dots, \omega_{\delta_p(K)}\}) \propto \hat{f}_{\delta_p(k)}^p(Y_{n'}^p) - \hat{f}_{\delta_p(k-1)}^p(Y_{n'}^p) \end{cases}$$

2. **Contextual discounting:** The proposed contextual discounting mechanism is then applied using the previously calculated weakening coefficients.

$$\begin{cases} m'^s(A) = \alpha_A^s m^s(A) & A \in \{2^\Omega/\Omega\} \\ m'^s(\Omega) = m^s(\Omega) + \sum_{A \in \{2^\Omega/\Omega\}} (1 - \alpha_A^s) m^s(A) \end{cases}$$



The proposed approach (Boosted PR-DS)

B. **Classification phase** involves four steps:

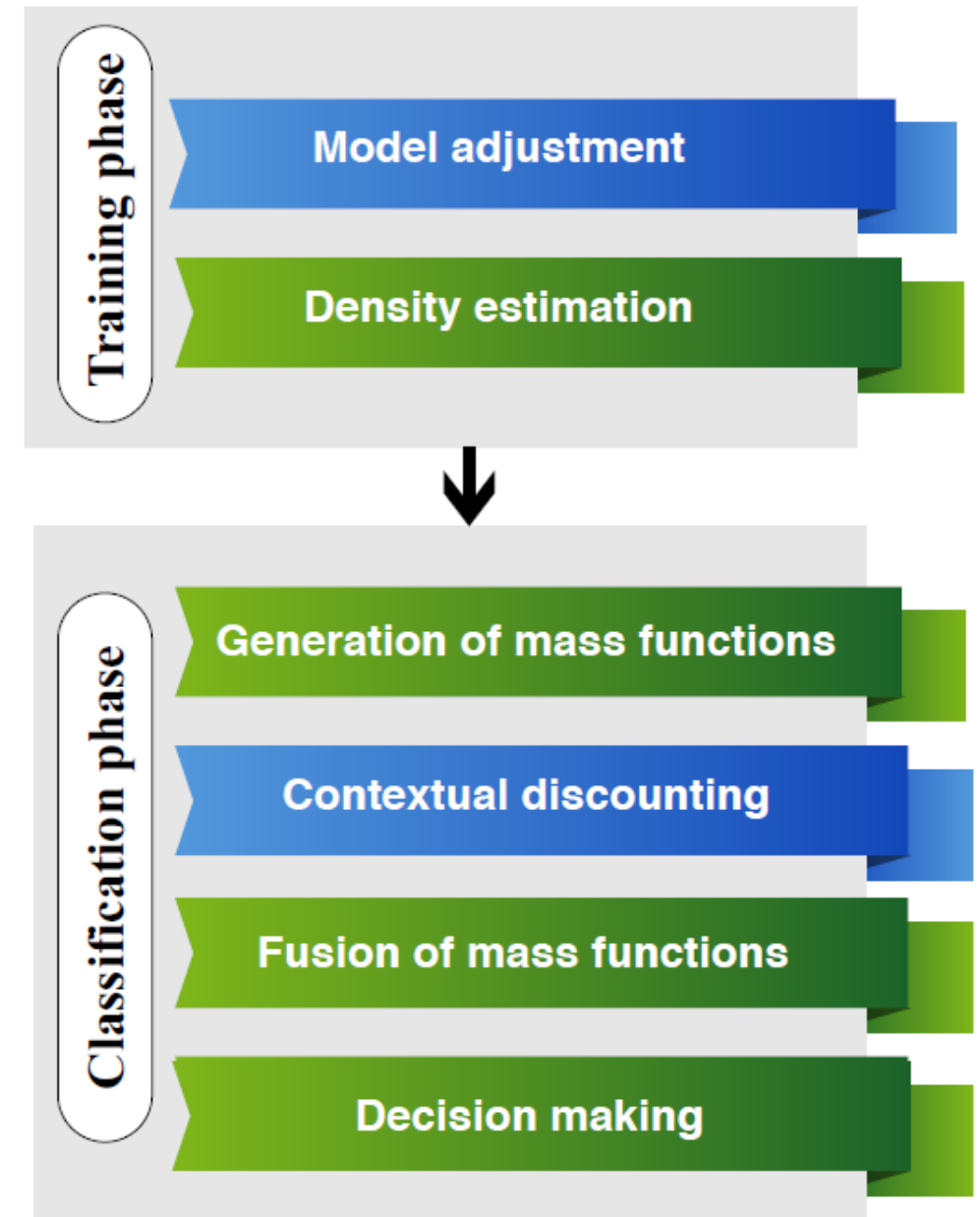
- 3. Fusion of mass functions:** Mass Functions assigned to different attributes are then merged into a single consensus mass

$$M = \bigoplus_{p=1}^P M^p$$

using the fusion rule selected on the training phase.

- 4. Decision making:** The final decision is made based on the Pignistic transformation of M:

$$\hat{X}_{n'} = \arg \max_{\omega_k} \sum_{A \ni \omega_k} \frac{M(A)}{|A|}$$





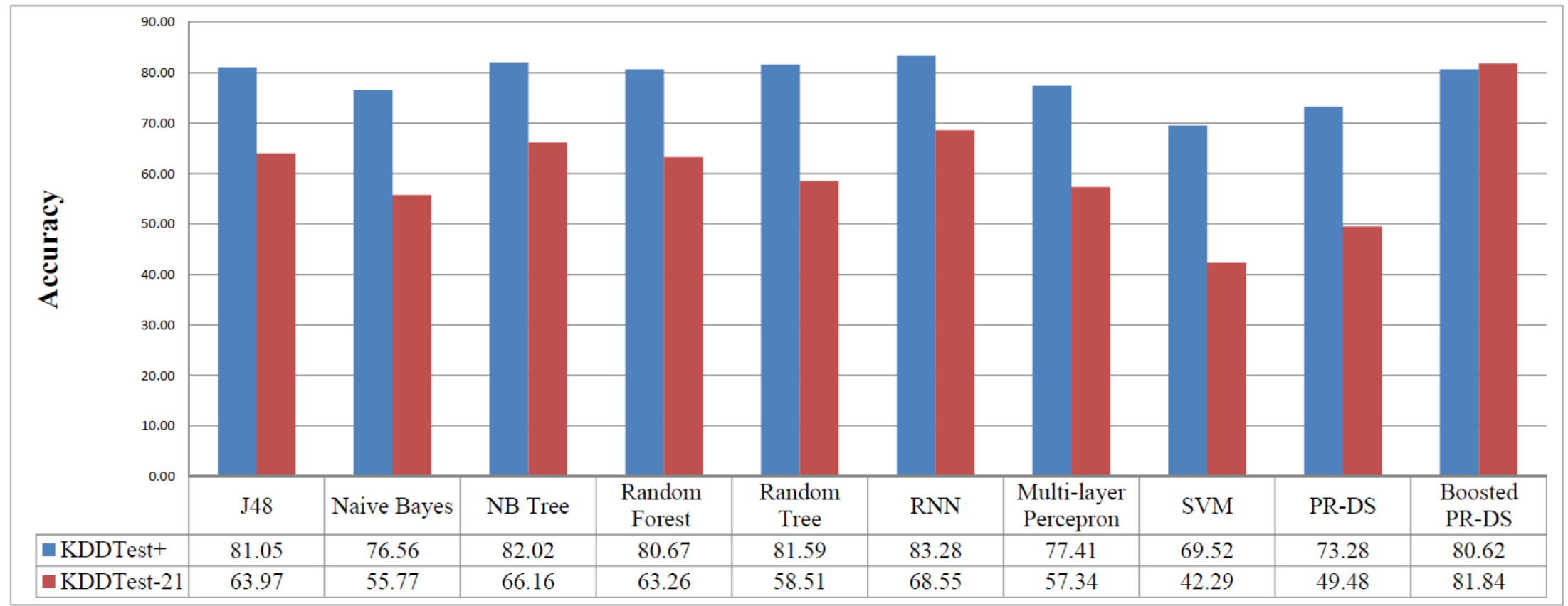
Main obtained results

NSL-KDD dataset description

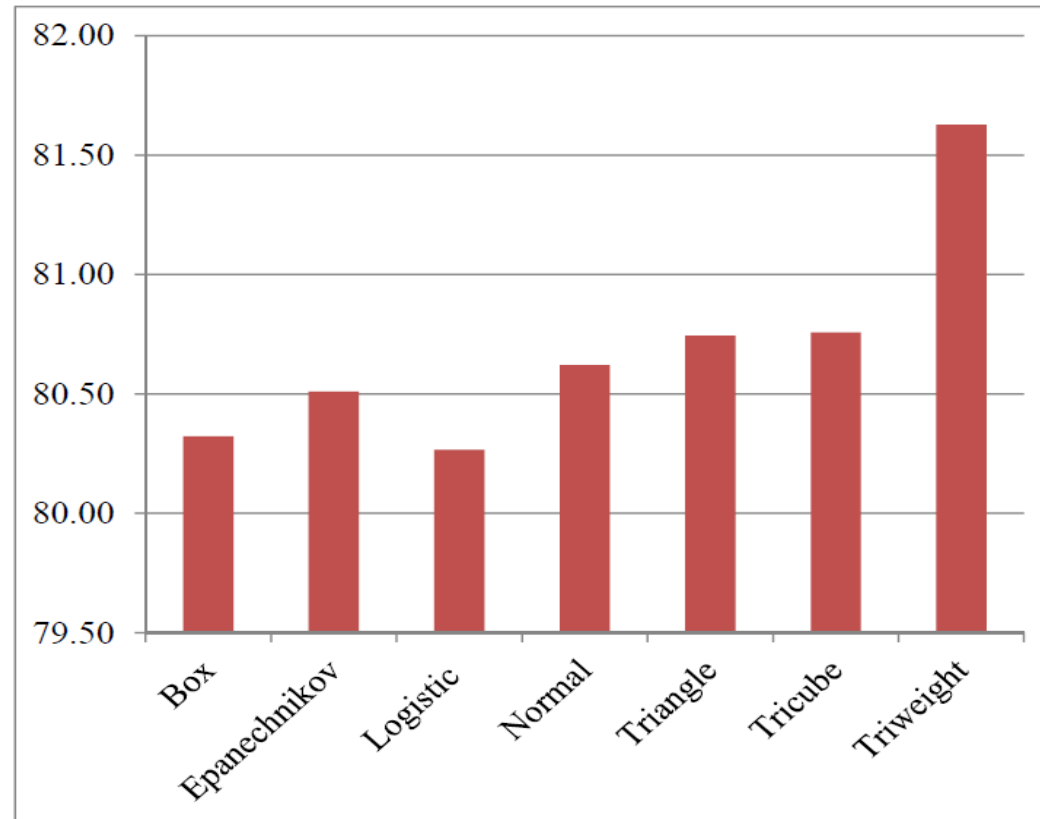
- Attacks in the dataset are grouped into four categories:
 - DoS (denial of service attacks);
 - Probe (Probing attacks);
 - R2L (root-to-local attacks);
 - U2R (user-to-root attacks).
- Each record has 41 attributes and a class label as well.
- KDDTest-21 which is a subset of the KDDTest+ is designed to be a more challenging dataset by removing the often correctly classified records.

	Normal	Dos	Probe	R2L	U2R
KDDTrain+	67343	45927	11656	995	52
KDDTest+	9711	7458	2421	2754	200
KDDTest-21	2152	4342	2402	2754	200

Experimental results



Performance of Boosted PR-DS and the other models on KDDTest+ and KDDTest-21.



Performance of Boosted PR-DS on the KDDTest-21 dataset using different kernels

Effect of kernel selection

These results confirm the relevance of **choosing a suitable kernel** to properly construct our densities instead of using the normality assumption.



Conclusion and future work

Conclusion and future work

- As a conclusion,
 - Boosted PR-DS can be considered as a combination of multiple classifiers where each source is a classifier.
 - By using contextual discounting, one can prioritize the decision of an individual classifier regarding those classes in which its accuracy was high in the training phase and be doubtful regarding those classes it did not classify well.
 - Boosted PR-DS choose a suitable fusion rule to take advantage of each individual classifier's knowledge to achieve a consensus decision.
- As a possible future direction, it would be interesting to consider **handling conflicting sources** with a more sophisticated fusion rule.