Détection d'anomalies non supervisée dans NSL-KDD en utilisant un β-VAE

une approche basée sur l'espace latent et l'erreur de reconstruction

Dylan Baptiste*†, Ramla Saddem*, Alexandre Philippot*, François Foyer†

*Université de Reims Champagne-Ardenne, CRESTIC, France
†Seckiot, Paris, France

Introduction

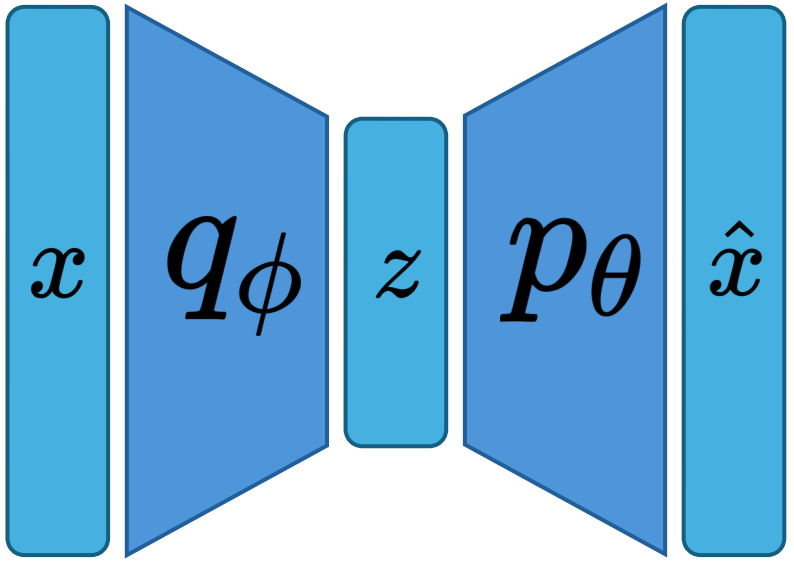
- Convergence IT/OT croissante
- Besoin accru en systèmes « *Intrusion Detection System* »
- Focus sur une approche non supervisée

Objectifs de l’étude

- Détection d’anomalies non supervisé
- Comparaison de deux approches en exploitant un β-VAE :
  - Erreur de reconstruction
  - Distance dans l’espace latent
- Application sur le jeu de données NSL-KDD*

*Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pages 1–6, July 2009

Le modèle β-VAE



Encodeur

Qui projette une donnée x

dans un espace latent $\boldsymbol{z}$

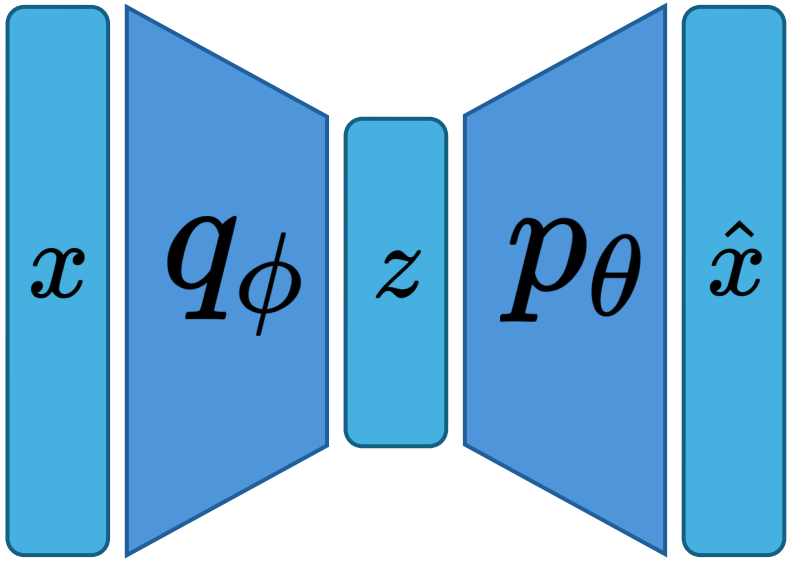
Décodeur $\boldsymbol{p_\theta}$

Qui récré la donnée initiale à partir de $\boldsymbol{z}$

couche

Réseau de neurones

Fonction de coût



$$\underbrace{-\mathbb{E}q_\phi(z|x)[\log p_\theta(x|z)]}_{\text{Erreur de reconstruction}} + \underbrace{\beta D_{KL}(q_\phi(z|x)||p(z))}_{\text{Divergence de Kullback-Leibler}}$$

Fonction de coût

Caractéristiques continues
activation linéaire avec *mean squared error*

$$\mathcal{L}_{\text{cont}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

Caractéristiques booléennes
activation *sigmoid* avec *binarycross-entropy loss*

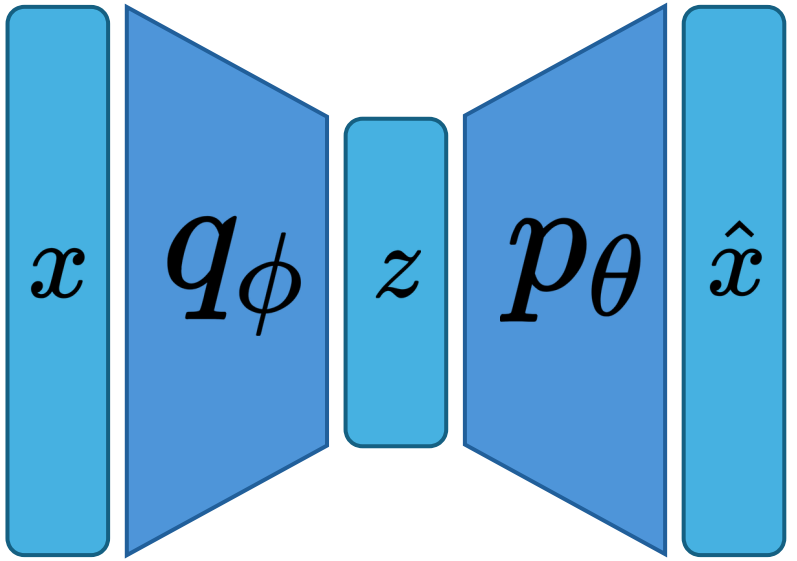
$$\mathcal{L}_{\text{bool}} = -\frac{1}{n}\sum_{i=1}^{n}\left[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)\right]$$

Caractéristiques catégorielles
activation *softmax* avec *categorical cross-entropy loss*

$$\mathcal{L}_{\text{cat}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\log(\hat{x}_{ij})$$

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{bool}} + \mathcal{L}_{\text{cont}}$$

Méthode 1 : Erreur de reconstruction



Algorithm 1 $\mathcal{L}_{rec}$-classification

Require: x a sample to classify, (q_ϕ, p_θ) : a trained β-VAE, τ : the threshold

Ensure: y : classification label : normal or anomalie

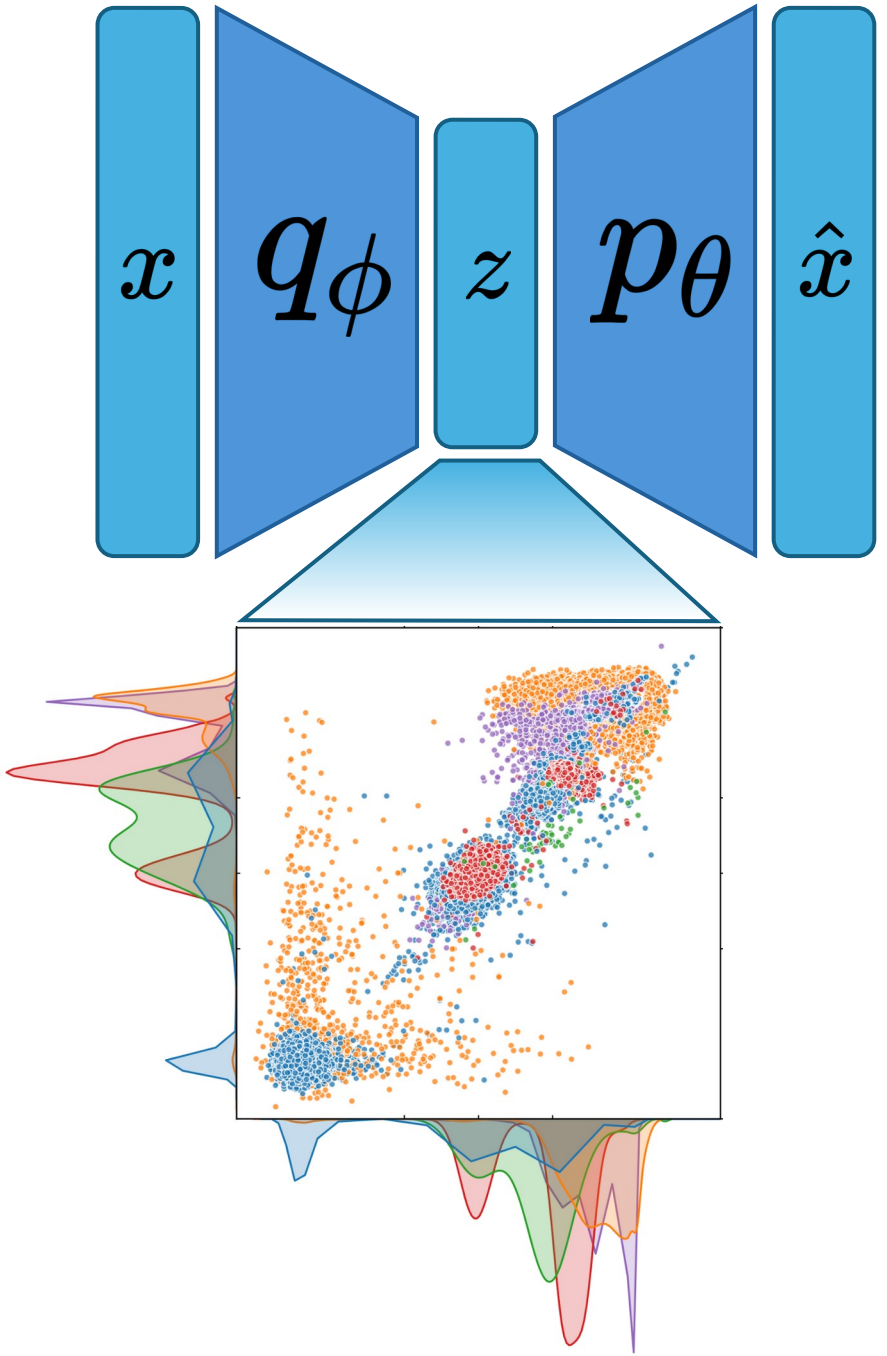
1: $z \sim q_\phi(z|x)$ ▷ Encoder
2: $\hat{x} \leftarrow p_\theta(z)$ ▷ Decoder
3: **if** $\mathcal{L}_{\text{rec}}(x, \hat{x}) > \tau$ **then**
4: $y \leftarrow$ anomalie
5: **else**
6: $y \leftarrow$ normal
7: **end if**
8: **return** y

Méthode 2 : Distance dans l'espace latent

On calcule la moyenne des distances de la projection ***z*** d'une nouvelle observation ***x*** au ***k*** plus proches projections d'un ensemble ***X*** connu.

$$\mathcal{Z}_k^X(x) = \frac{1}{k}\sum_{j=1}^{k} \|z - z'_{(j)}\|_2$$

Avec $z \sim q_\varphi(x|z)$
et $z'_{(j)}$ le *j*-ème plus proche voisin de *z* dans l'ensemble des projections de *X*.



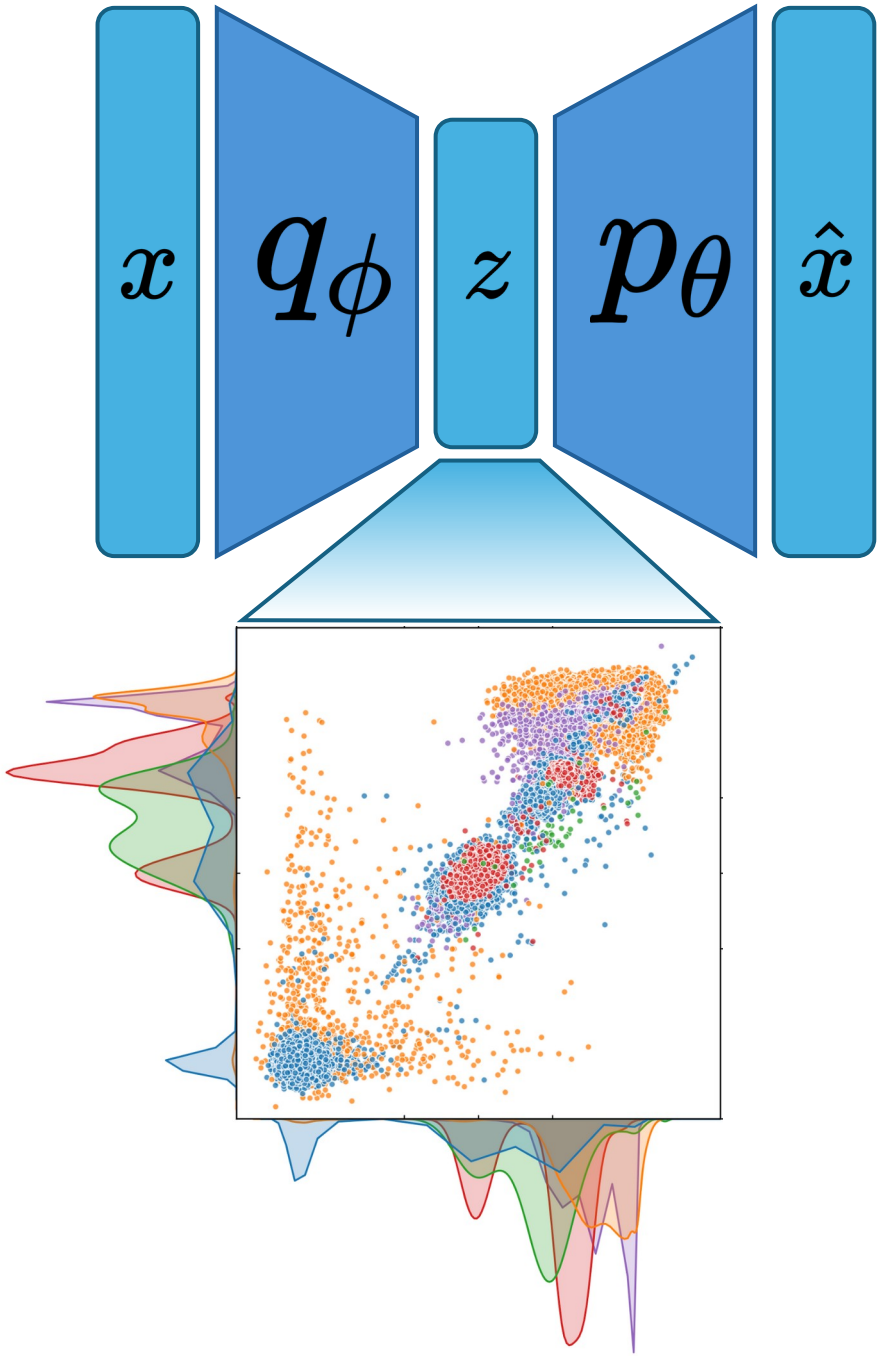
Méthode 2 : Distance dans l'espace latent

Algorithm 2 $\mathcal{Z}_k$-classification

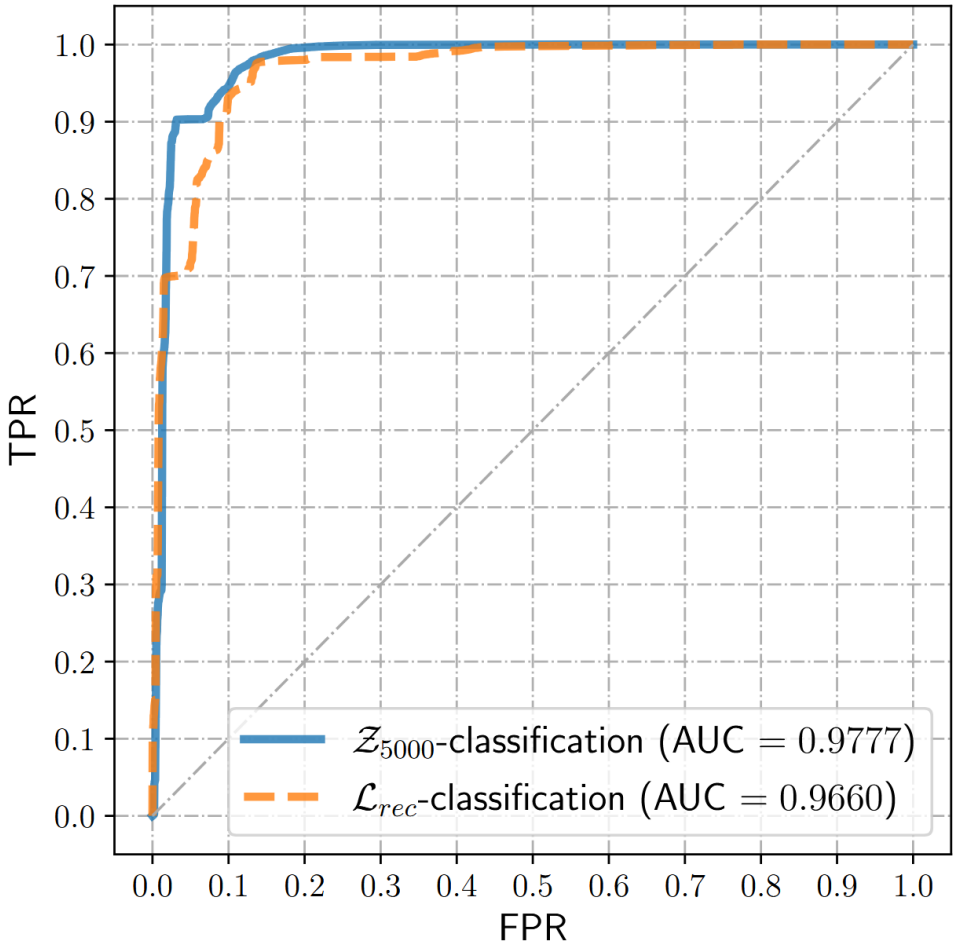
Require: x a sample to classify, (q_ϕ, p_θ) : a trained β-VAE, X_{train} : the training dataset, k : the number of neighbors, τ : the threshold

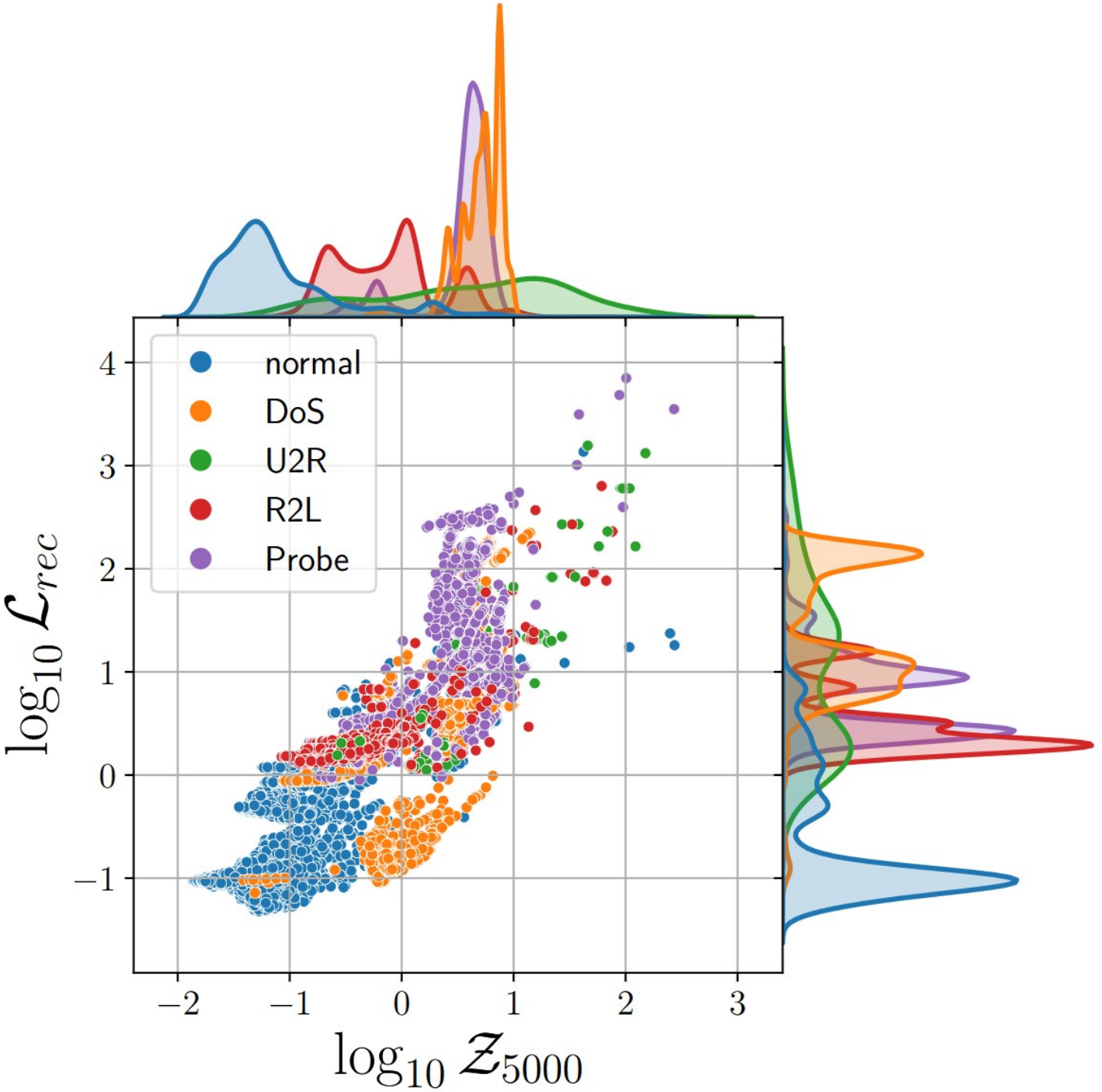
Ensure: y : classification label : normal or anomalie

$Z_{train} \leftarrow \{z_i \sim q_\phi(z|x_i), \forall x_i \in X_{train}\}$
$z \sim q_\phi(z|x)$
Find the k nearest neighbors $z'_{(1)}, \ldots, z'_{(k)}$ of z in Z_{train}
if $\mathcal{Z}_k^{X_{train}}(x) > \tau$ **then**
    $y \leftarrow$ anomalie
else
    $y \leftarrow$ normal
end if
return y



Évaluation


TPR
FPR
$\mathcal{Z}_{5000}$-classification (AUC = 0.9777)
$\mathcal{L}_{rec}$-classification (AUC = 0.9660)


normal
DoS
U2R
R2L
Probe
$\log_{10} \mathcal{L}_{rec}$
$\log_{10} \mathcal{Z}_{5000}$

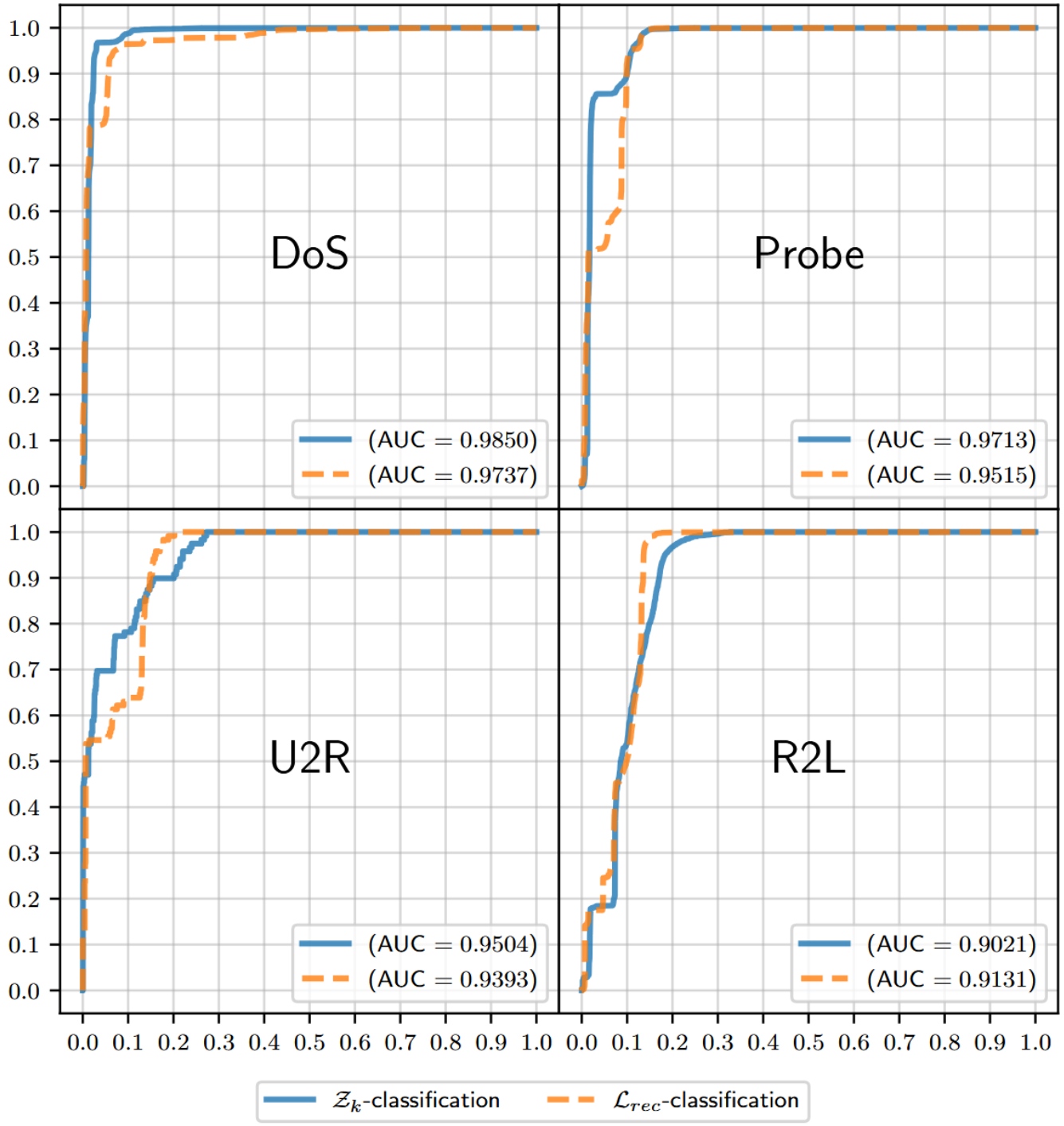
Évaluation

Denial of Service (DoS) :
Visent à rendre un service ou un système indisponible en le surchargeant de requêtes ou de trafic.

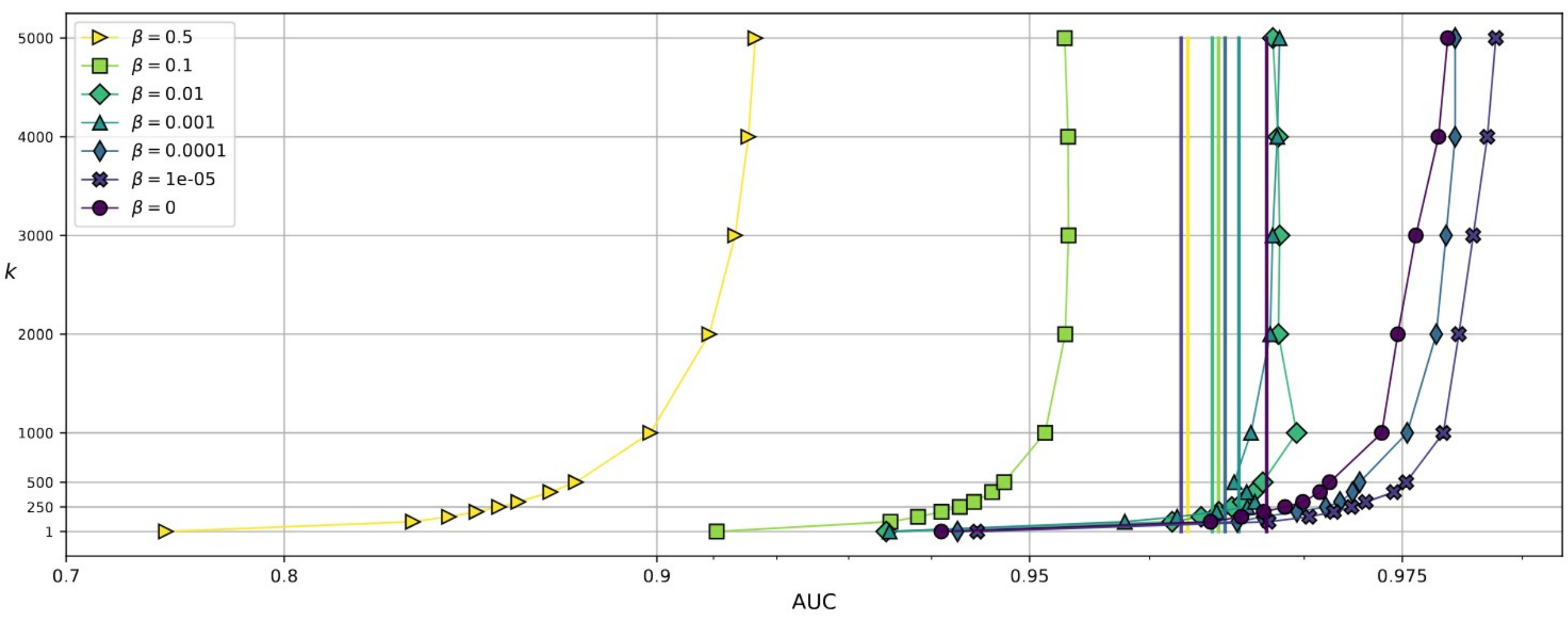
Probe :
Cherchent à collecter des informations sur le réseau ou les systèmes, comme la cartographie des ports ou des services actifs, en vue d'une attaque ultérieure.

User-to-Root (U2R) :
Permettent à un utilisateur malveillant, disposant d'un accès limité à une machine, d'obtenir des privilèges administrateur

Remote-to-Local (R2L) :
Permettent à un attaquant distant d'obtenir un accès local non autorisé sur une machine cible



Évaluation

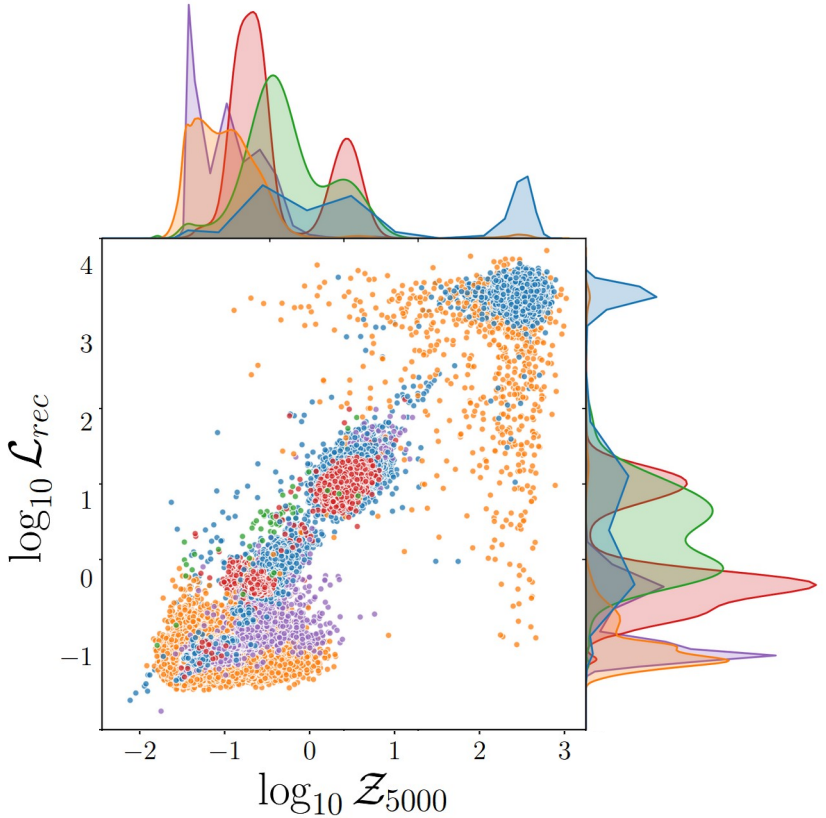

β = 0.5
β = 0.1
β = 0.01
β = 0.001
β = 0.0001
β = 1e-05
β = 0
5000
4000
3000
2000
1000
500
250
1
k
0.7
0.8
0.9
0.95
0.975
AUC

Conclusion

- Les deux méthodes obtiennent d'excellents résultats sur ce jeu de données démontrant ainsi l'efficacité des β-VAE pour la détection non supervisée dans ce contexte

& Perspectives

- Seuils adaptatifs multi-métriques
- Fusion L_{rec}-classification + Z_k-classification
- La Z_k-classification ouvre la voie à une classification multilabel, au-delà d'un simple cadre binaire



Questions


DoS
(AUC = 0.9850)
(AUC = 0.9737)
Probe
(AUC = 0.9713)
(AUC = 0.9515)
U2R
(AUC = 0.9304)
(AUC = 0.9303)
R2L
(AUC = 0.9021)
(AUC = 0.9131)