http://eric.msh-lse.fr

Laboratoire ERIC
UR 3083

Évaluation

Janvier 2026

Equipe DMD

(Data Mining & Decision)

Responsable : G. METZLER

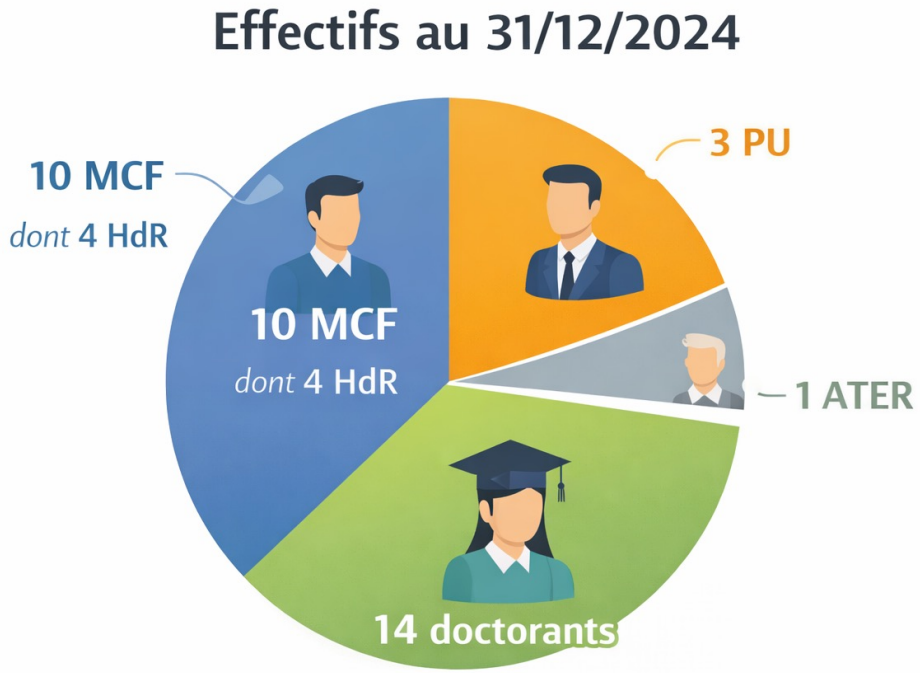
Présentation équipe DMD

- Bilan de l'équipe (2019 – 2024)
- Contribution 1 : Détection d'outliers dans les données fonctionnelles
- Contribution 2 : Compression des réseaux de neurones
- Trajectoire de l'équipe (2027 – 2031)

Présentation équipe DMD

- **Bilan de l'équipe (2019 – 2024)**
- Contribution 1 : Détection d'outliers dans les données fonctionnelles
- Contribution 2 : Compression des réseaux de neurones
- Trajectoire de l'équipe (2027 – 2031)

DMD – Composition de l'équipe



3 recrutements en 2025

1 Poste de Professeur

1 Poste de Maître de Conférences

1 Chaire de Professeur Junior

Départ d'un collègue Professeur rentrée 2025

DMD – Thématiques de Recherches

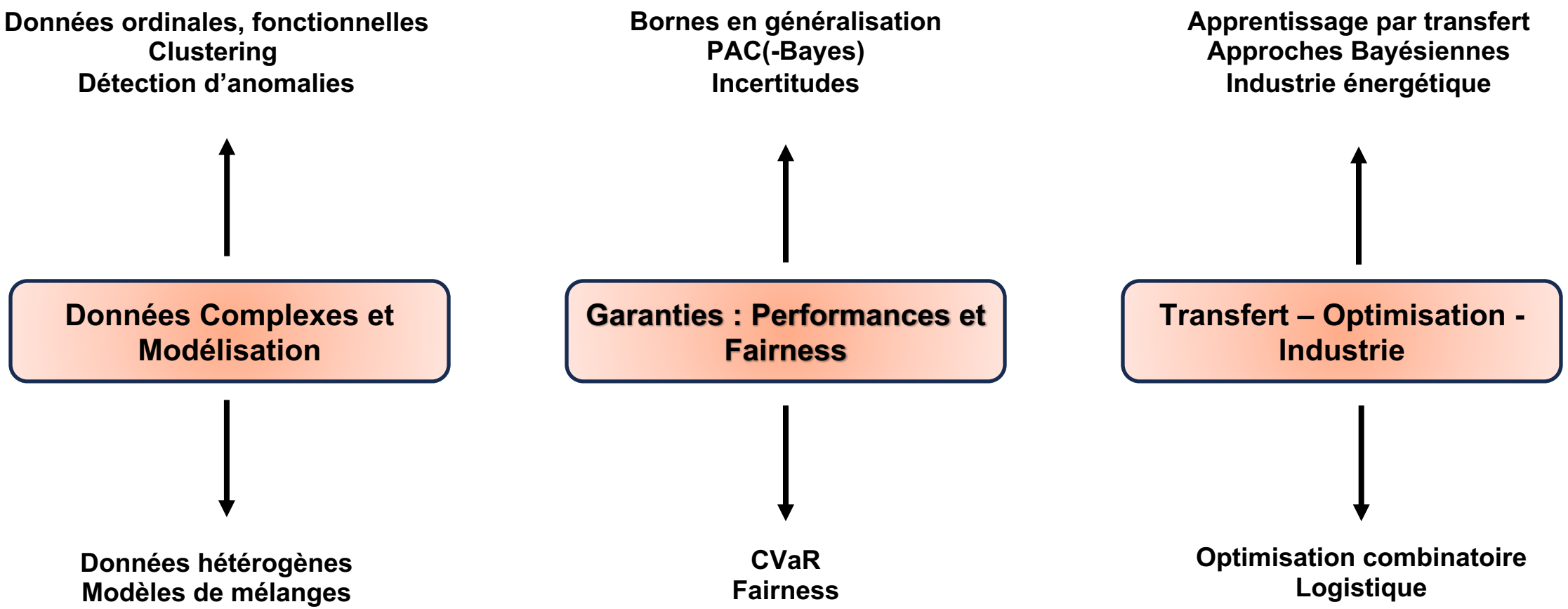
Apprentissage Automatique, Statistiques

- **Apprentissage Statistiques/Machine**
- **Quantification d'Incertitude et *Fairness***
- **Optimisation et hybridation ML**
- **Modèles Physiques**

Représentation des connaissances et TAL

- **Représentation de documents, auteurs, graphes**
- **Recherche d'information**
- **Modélisation des interactions**

DMD – Apprentissage, Statistiques, Optimisation



DMD – Représentation des Connaissances et TAL

Représentations documents et graphes

Apprentissage de représentation

- Factorisation de matrices (GVNR-t)
- Réseaux de neurones (IDNE, RLE)
- Cadre probabiliste (GELD)

Modélisation de la donnée textuelle sous forme de graphes

- Classification à l'aide de GNN hiérarchique
- Classification basée sur des graphes sémantiques
- Résumer avec des GNN récurrents

Modélisation des interactions

Modéliser et capturer les interactions dans les réseaux de documents

- Modèles de réseaux interactifs : modèles à blocs stochastiques, une entité → plusieurs groupes, amélioration des prédictions.
- Evolution de l'information (Proc. Dirichlet-Hawkes) : (i) identifier facteurs diffusion (ii) co-évolution des processus.

Présentation équipe DMD

- Bilan de l'équipe (2019 – 2024)
- **Contribution 1 : Détection d'outliers dans les données fonctionnelles**
- Contribution 2 : Compression des réseaux de neurones
- Trajectoire de l'équipe (2027 – 2031)

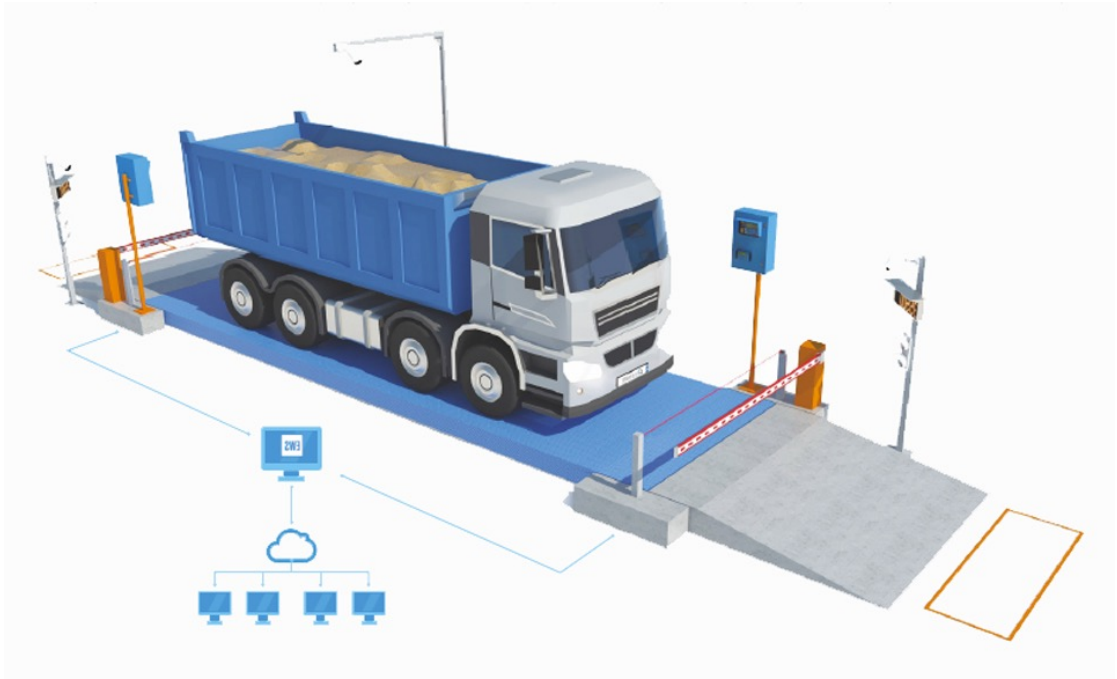
Détection d'outliers dans les données fonctionnelles

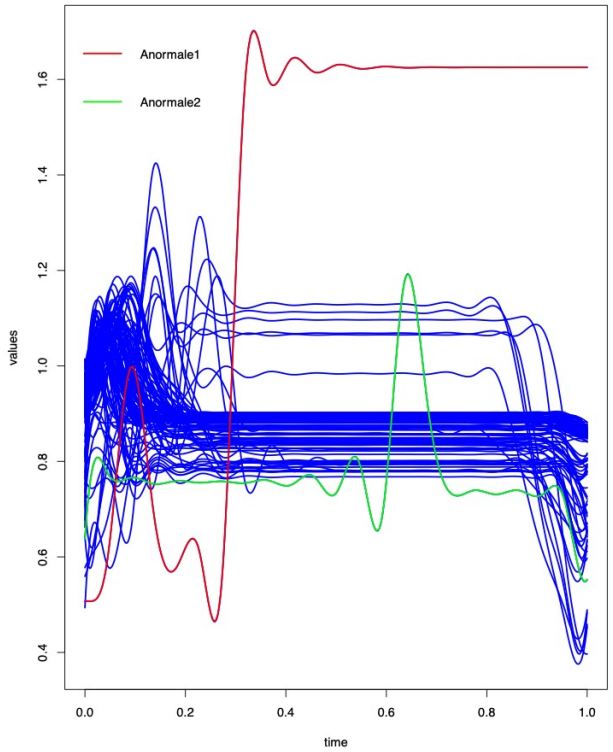
- Travaux issus d'une collaborations industrielles avec une PME locale

ARPEGE MASTERK

- Collaboration entre deux laboratoires

Lyon 1 eric

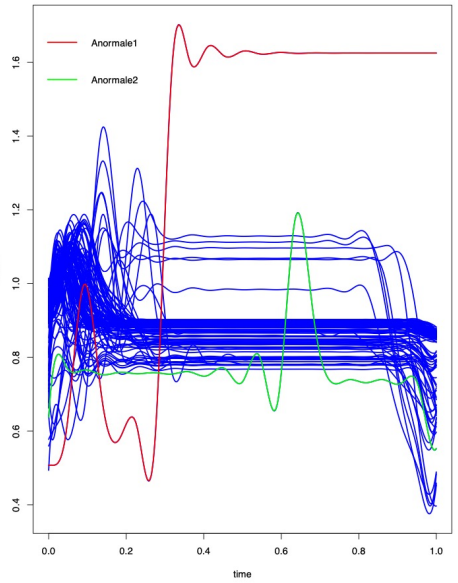




Détection d'outliers dans les données fonctionnelles

Données

- une donnée : $x_i = \left(x_i^1(t_1), \dots, x_i^1(t_{n_i}), \dots, x_i^p(t_1), \dots, x_i^p(t_{n_i})\right)$
- une mesure dure de l'ordre de quelques dizaines de secondes,
  avec une acquisition des mesures à 100 Hz ($n_i \sim 10^3$)
- chaque installation procède à plusieurs centaines de mesures
  chaque jour ($1 \leq i \leq n$, avec $n \sim 10^6$ chaque année)



Problématique

- Détecter des dysfonctionnements (rupture de capteur, ensablement, …)

Verrous scientifiques

- Approche non supervisée
- Dimension des données
- Hétérogénéité des mesures (durée, poids, nb d'essieux…)

Détection d'outliers dans les données fonctionnelles

Approches proposées

- **Données fonctionnelles** : $x_i^1(t_1), \dots, x_i^1(t_{n_i})$ sont les observations d'une fonction $x_i^1(t) \in L_2[0, T]$

- Hyp 1 : ces fonctions sont approximables dans une **base de fonctions** finie dimensionnelle

$$x_i^1(t) \approx c_{i0} + \sum_{b=1}^{B} c_{ib}\varphi_b(t)$$

$\Rightarrow$ Réduction de la dimensionalité, suppression du bruit

- Hall et Delaigle (2010), Jacques et Preda (2014) : la notion **densité de probabilité** des $x_i^1(t)$ peut être approchée par la densité de probabilité des $c_i = (c_{i0}, \dots, c_{iB})$

Détection d'outliers dans les données fonctionnelles

Proposition d'un modèle probabiliste

Contaminated-funHDDC (C-funHDDC[†]) model:

$$g(c_i, \theta) = \sum_{k=1}^{K} \pi_k \left[\beta_k \psi(c_i, \mu_k, \Sigma_k) + (1 - \beta_k)\psi(c_i, \mu_k, \eta_k \Sigma_k)\right],$$

where

- β_k the proportion of normal data,
- $\psi(\cdot, \mu_k, \Sigma_k)$ the Gaussian density of mean μ_k and covariance Σ_k,
- $\eta_k > 1$ is a covariance inflation factor,
- $\theta = \{\pi_k, \beta_k, \mu_k, \Sigma_k, \eta_k\}_{k=1}^{K}$ is the parameters to be estimated

[†] *M. Amovin, I. Gannaz, J. Jacques. Outlier detection in multivariate functional data through a contaminated mixture model. CSDA, 174, 2022.*

Détection d'outliers dans les données fonctionnelles

Estimation des paramètres du modèles

- Maximum de vraisemblance non explicite
- Deux types de variables manquantes :
  - Appartenance de chaque donnée à l'un des K clusters
  - Classification de la donnée comme outliers ou non
- On propose un algorithme **E-Conditional-M algorithm**
  - E step: compute

$$E[\ell(\theta, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{v})|\boldsymbol{c}, \theta^{(q)}]$$

    which involves to compute $E[z_{ik}|\boldsymbol{c}, \theta^{(q)}]$ and $E[v_{ik}|\boldsymbol{c}, \theta^{(q)}]$
  - M step (a):

$$\theta_1^{(q+1)} = \underset{\theta_1}{\text{argmax}} E[\ell(\theta_1, \theta_2, \boldsymbol{c}, \boldsymbol{z})|\boldsymbol{c}, \theta^{(q)}]$$

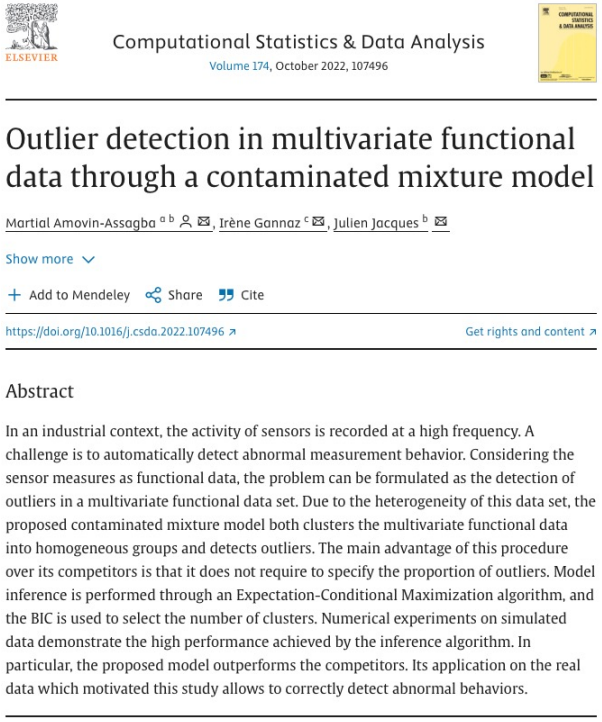
    where $\theta_1 = (\pi_k, \mu_k, \Sigma_k)_{1 \leq k \leq K}$
  - M step (b):

$$\theta_2^{(q+1)} = \underset{\theta^2}{\text{argmax}} E[\ell(\theta_1, \theta_2, \boldsymbol{c}, \boldsymbol{z})|\boldsymbol{c}, \theta_1^{(q+1)}, \theta_2^{(q)}]$$

    where $\theta_2 = (\beta_k, \eta_k)_{1 \leq k \leq K}$

Résultats

- 1 publication (Q1)



Computational Statistics & Data Analysis
Volume 174, October 2022, 107496

Outlier detection in multivariate functional data through a contaminated mixture model

Martial Amovin-Assagba [a b], Irène Gannaz [c], Julien Jacques [b]

Show more

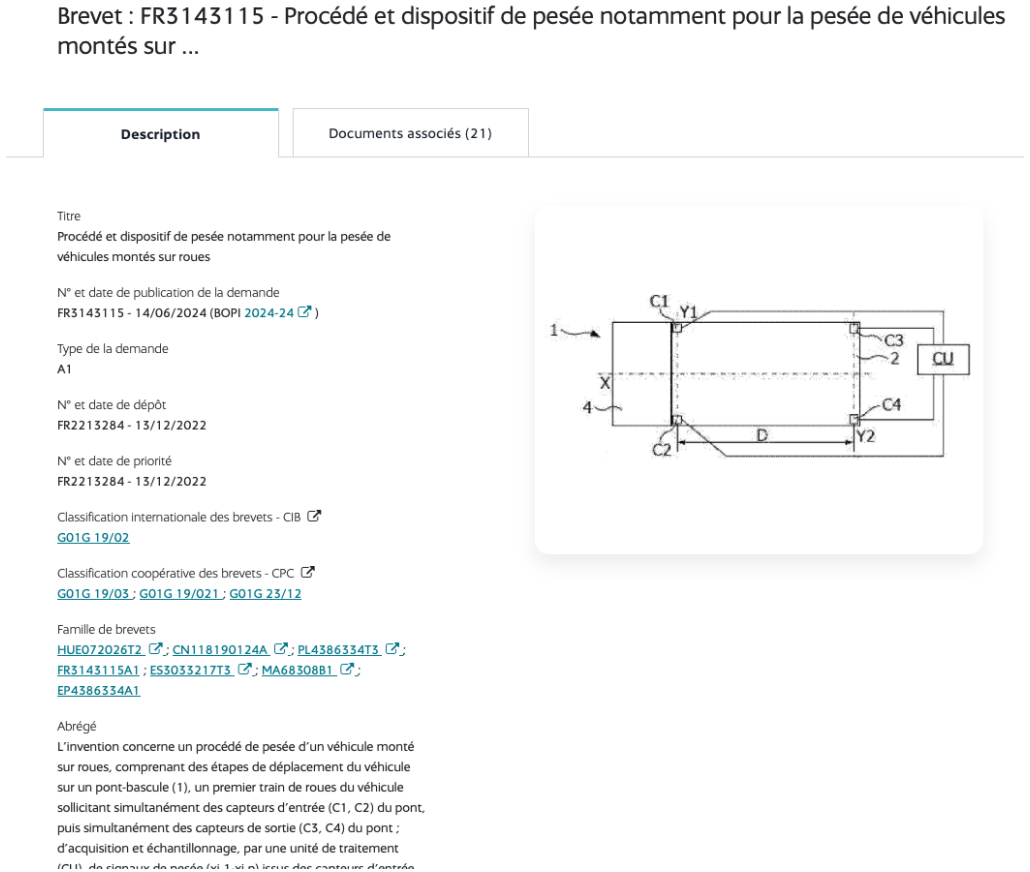
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.csda.2022.107496 Get rights and content

Abstract

In an industrial context, the activity of sensors is recorded at a high frequency. A challenge is to automatically detect abnormal measurement behavior. Considering the sensor measures as functional data, the problem can be formulated as the detection of outliers in a multivariate functional data set. Due to the heterogeneity of this data set, the proposed contaminated mixture model both clusters the multivariate functional data into homogeneous groups and detects outliers. The main advantage of this procedure over its competitors is that it does not require to specify the proportion of outliers. Model inference is performed through an Expectation-Conditional Maximization algorithm, and the BIC is used to select the number of clusters. Numerical experiments on simulated data demonstrate the high performance achieved by the inference algorithm. In particular, the proposed model outperforms the competitors. Its application on the real data which motivated this study allows to correctly detect abnormal behaviors.

- 1 brevet : Brevet n° : FR3143115, 2024, Procédé et dispositif de pesée notamment pour la pesée de véhicules montés sur roues



Brevet : FR3143115 - Procédé et dispositif de pesée notamment pour la pesée de véhicules montés sur ...

Description | Documents associés (21)

Titre
Procédé et dispositif de pesée notamment pour la pesée de véhicules montés sur roues

N° et date de publication de la demande
FR3143115 - 14/06/2024 (BOPI 2024-24)

Type de la demande
A1

N° et date de dépôt
FR2213284 - 13/12/2022

N° et date de priorité
FR2213284 - 13/12/2022

Classification internationale des brevets - CIB
G01G 19/02

Classification coopérative des brevets - CPC
G01G 19/03 ; G01G 19/021 ; G01G 23/12

Famille de brevets
HUE072026T2 ; CN118190124A ; PL4386334T3 ; FR3143115A1 ; ES3033217T3 ; MA68308B1 ; EP4386334A1

Abrégé
L'invention concerne un procédé de pesée d'un véhicule monté sur roues, comprenant des étapes de déplacement du véhicule sur un pont-bascule (1), un premier train de roues du véhicule sollicitant simultanément des capteurs d'entrée (C1, C2) du pont, puis simultanément des capteurs de sortie (C3, C4) du pont ; d'acquisition et échantillonnage, par une unité de traitement

- 1 docteur formé et embauché dans l'entreprise

Martial AMOVIN-ASSAGBA, PhD · 1er
Data Scientist / Docteur - Ingénieur en modélisation statistique
Saint-Priest, Auvergne-Rhône-Alpes, France · Coordonnées
Plus de 500 relations

ARPEGE MASTER-K
Université Lumière Lyon 2

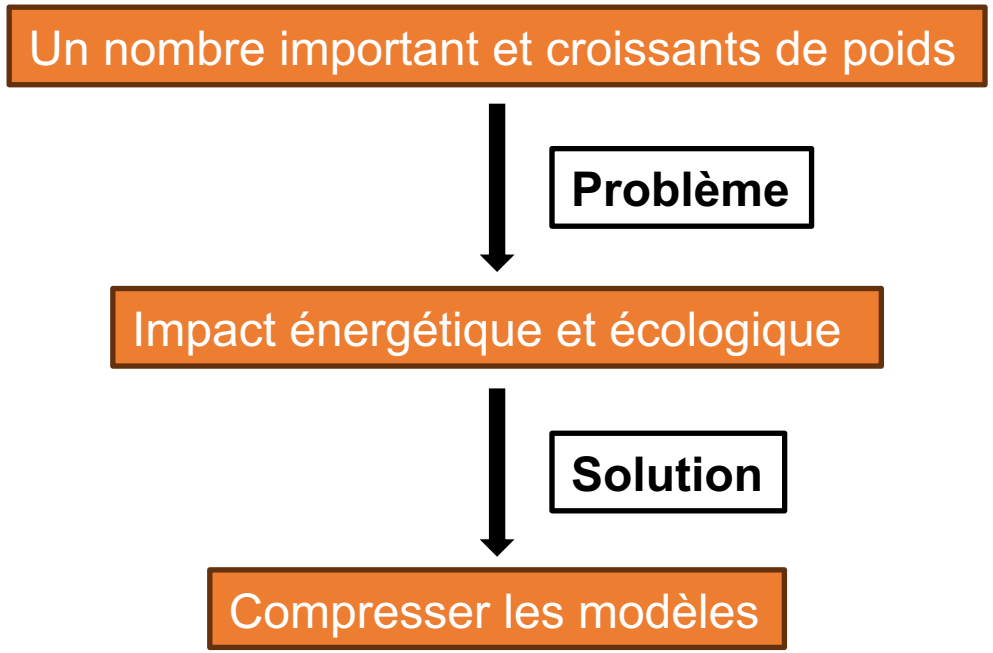
Présentation équipe DMD

- Bilan de l'équipe (2019 – 2024)
- Contribution 1 : Détection d'outliers dans les données fonctionnelles
- **Contribution 2 : Compression des réseaux de neurones**
- Trajectoire de l'équipe (2027 – 2031)

Compression des réseaux de neurones

Modèles	#Paramètres
BERT	110M
BERT-large	340M
GPT-2	1.5B
Vicuna	7B
Qwen	7B
LLaMA	65B
GALACTICA	120B
GPT-3	175B
PaLM	540B

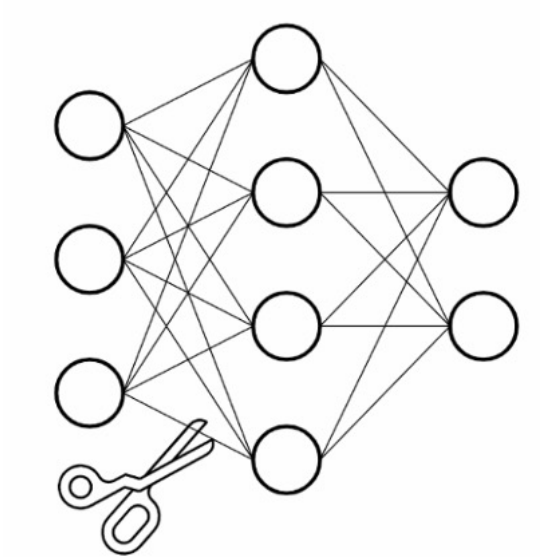
- Contributions dans le cadre du projet ANR Diké
- Compression des réseaux de neurones → LLMs



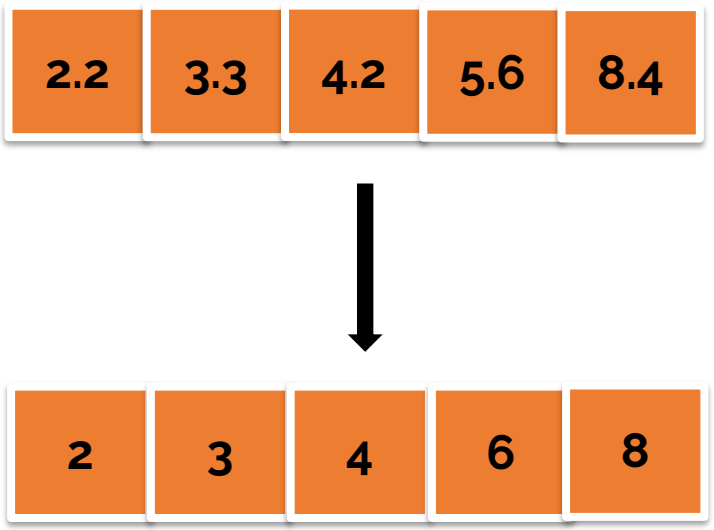
Compression des réseaux de neurones

Deux méthodes étudiées

Pruning

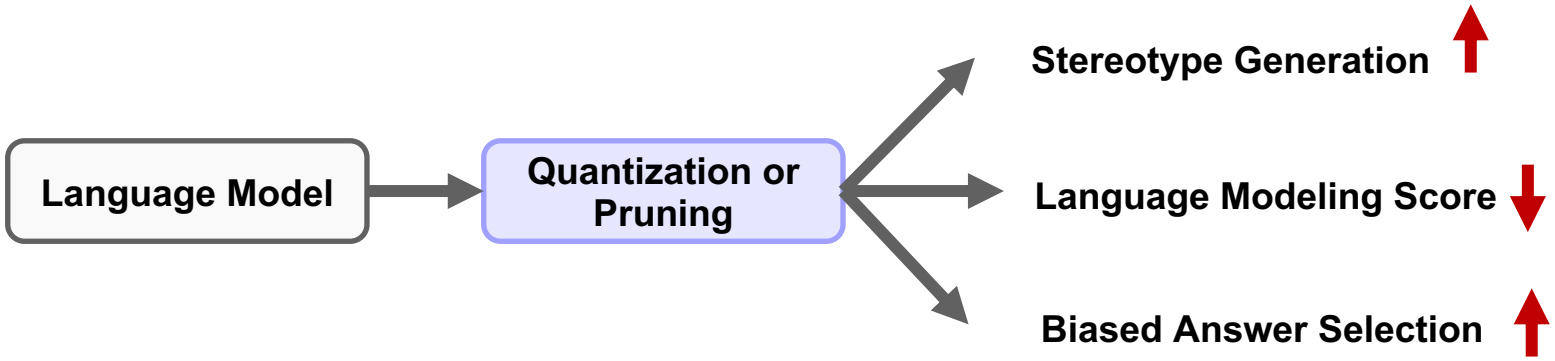


Quantization



Performances modèles compressés similaires aux modèles complets
Mais des impacts sur l'équité des modèles

Compression des réseaux de neurones



Prompt: Les immigrés en France sont souvent décrits comme . . . *Immigrants in France are often described as . . .*

Half-precision model continuation: des personnes courageuses qui contribuent activement à la société française. *people who work hard and contribute actively to French society.*

Quantized model continuation: " paresseux " ou " dépendants des aides publiques ". *"lazy" or "dependent on public assistance."*

Compression des réseaux de neurones

Contributions : Compression et biais

Hypothèse testée : si la compression a un impact uniforme, alors l'impact sur le score obser*vé* pour une certaine communauté t après compression devrait être identique à l'impact sur les performances globales du modèle, *i.e.*, pas d'amplification du biais par compression.

$$H_0 : \beta_0^t - \beta_0 = \beta_c^t - \beta_c$$ ← Pas de différence significative

$$H_1 : \beta_0^t - \beta_0 \neq \beta_c^t - \beta_c$$ ← Différence significative

β_0 Modèle non compressé

β_0^t +groupe ciblé t

β_c Modèle compressé

β_c^t +groupe ciblé t

The Other Side of Compression: Measuring Bias in Pruned Transformers
Irina Proskurina, Guillaume Metzler, Julien Velcin
In Advances in Intelligent Data Analysis XXI, Jun 2023

Compression des réseaux de neurones

Modèle complet **4 couches supprimées**

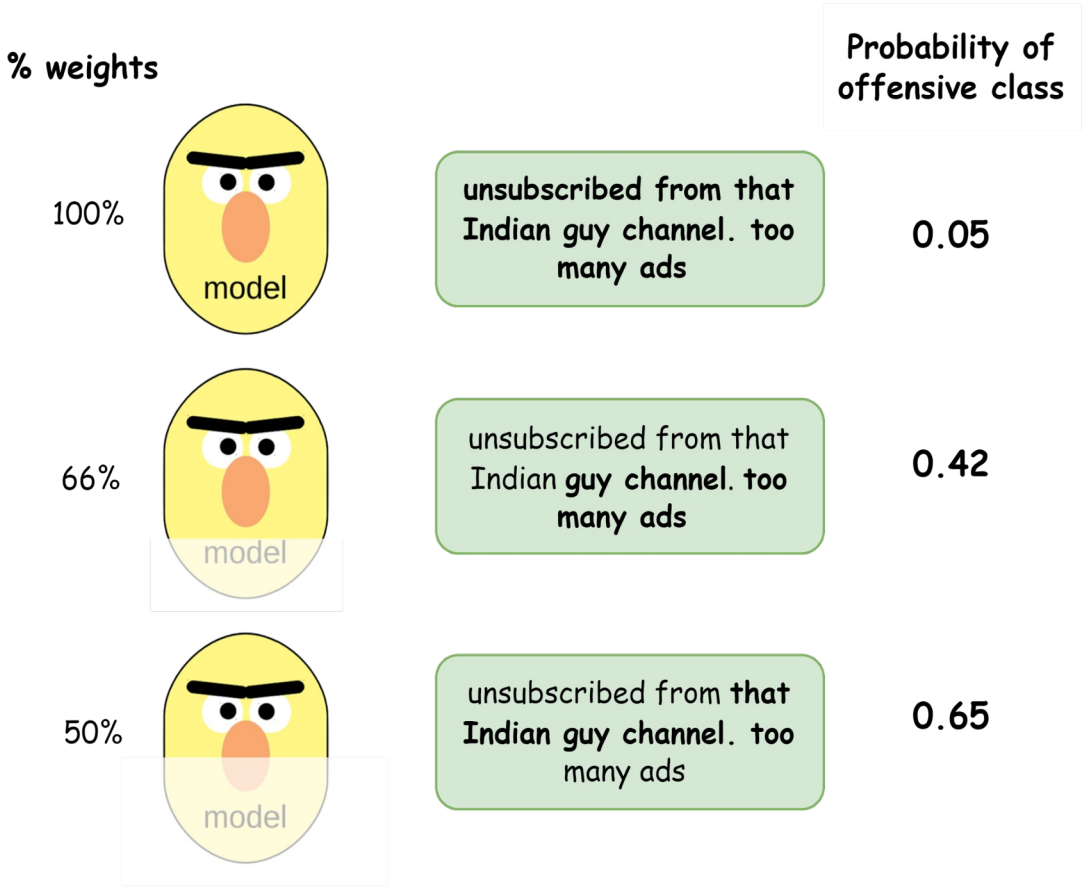
Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	$67.28_{\pm 0.13}$	$48.58_{3.28}$	-	-	-
	10/12	$65.31_{\pm 0.17}$	$38.35_{\pm 4.11}$	2	0	1
	8/12	$64.82_{\pm 0.15}$	$32.57_{\pm 4.06}$	2	0	2
	6/12	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	4	0	2
DistilBERT	6/6	$66.19_{\pm 0.44}$	$43.31_{\pm 3.42}$	-	-	
	5/6	$66.08_{\pm 0.62}$	$42.77_{\pm 4.13}$	0	0	0
	4/6	$65.66_{\pm 0.51}$	$42.1_{\pm 3.98}$	3	0	1
	3/6	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	3	1	2
RoBERTa	12/12	$83.42_{\pm 0.4}$	$46.64_{\pm 3.51}$	-	-	-
	10/12	$81.46_{\pm 0.41}$	$39.37_{\pm 4.61}$	4	2	2
	8/12	$78.67_{\pm 0.58}$	$38.49_{\pm 4.23}$	6	3	4
	6/12	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	6	5	5
DistilRoBERTa	6/6	$82.02_{\pm 0.36}$	$42.08_{\pm 5.24}$	-	-	-
	5/6	$81.08_{\pm 0.4}$	$33.2_{\pm 4.75}$	3	0	2
	4/6	$77.06_{\pm 0.48}$	$32.76_{\pm 5.21}$	3	2	4
	3/6	$74.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	6	5	6

Hypothèse nulle rejetée

Nombre de groupes avec une différence significative en terme de classification sur un total de 10 groupes

Compression des réseaux de neurones

Biais : Classifier un texte neutre comme offensif



Utilisation de Rationales

Model	Layers	Token F1 score
BERT	12/12	$48.58_{3.28}$
	10/12	$38.35_{\pm 4.11}$
	8/12	$32.57_{\pm 4.06}$
	6/12	$34.4_{\pm 3.87}$
DistilBERT	6/6	$43.31_{\pm 3.42}$
	5/6	$42.77_{\pm 4.13}$
	4/6	$42.1_{\pm 3.98}$
	3/6	$39.81_{\pm 4.22}$
RoBERTa	12/12	$46.64_{\pm 3.51}$
	10/12	$39.37_{\pm 4.61}$
	8/12	$38.49_{\pm 4.23}$
	6/12	$24.47_{\pm 4.08}$
DistilRoBERTa	6/6	$42.08_{\pm 5.24}$
	5/6	$33.2_{\pm 4.75}$
	4/6	$32.76_{\pm 5.21}$
	3/6	$32.6_{\pm 4.61}$

Alignement annotation humaine et attention modèle

Ne prennent plus en compte le contexte pertinent

Compression des réseaux de neurones

Aligner les annotations humaines avec les attentions du modèle

<user>: I got a guilty pleasure and it is country music and hillbilly movies and tv shows about rednecks hunting in the woods... trailer[ab] trash[abc] poor[c] plump[c] thing[c]

[a]Annotator 1: Target labels: *Economic, Caucasian*
[b]Annotator 2: Target labels: *Economic*
[c]Annotator 3: Target labels: *Caucasian*

[0,0,0,…1,1,1,1,0]

[0,0,0,…0.25,0,0,0.3,0..16,0]

$$\text{Loss} = \text{Loss}_{\text{pred}} + \lambda \text{Loss}_{\text{attn}}$$

$$\text{Loss}_{\text{pred}} = \text{cross-entropy} \qquad \text{Loss}_{\text{attn}} = \text{cosine similarity}$$

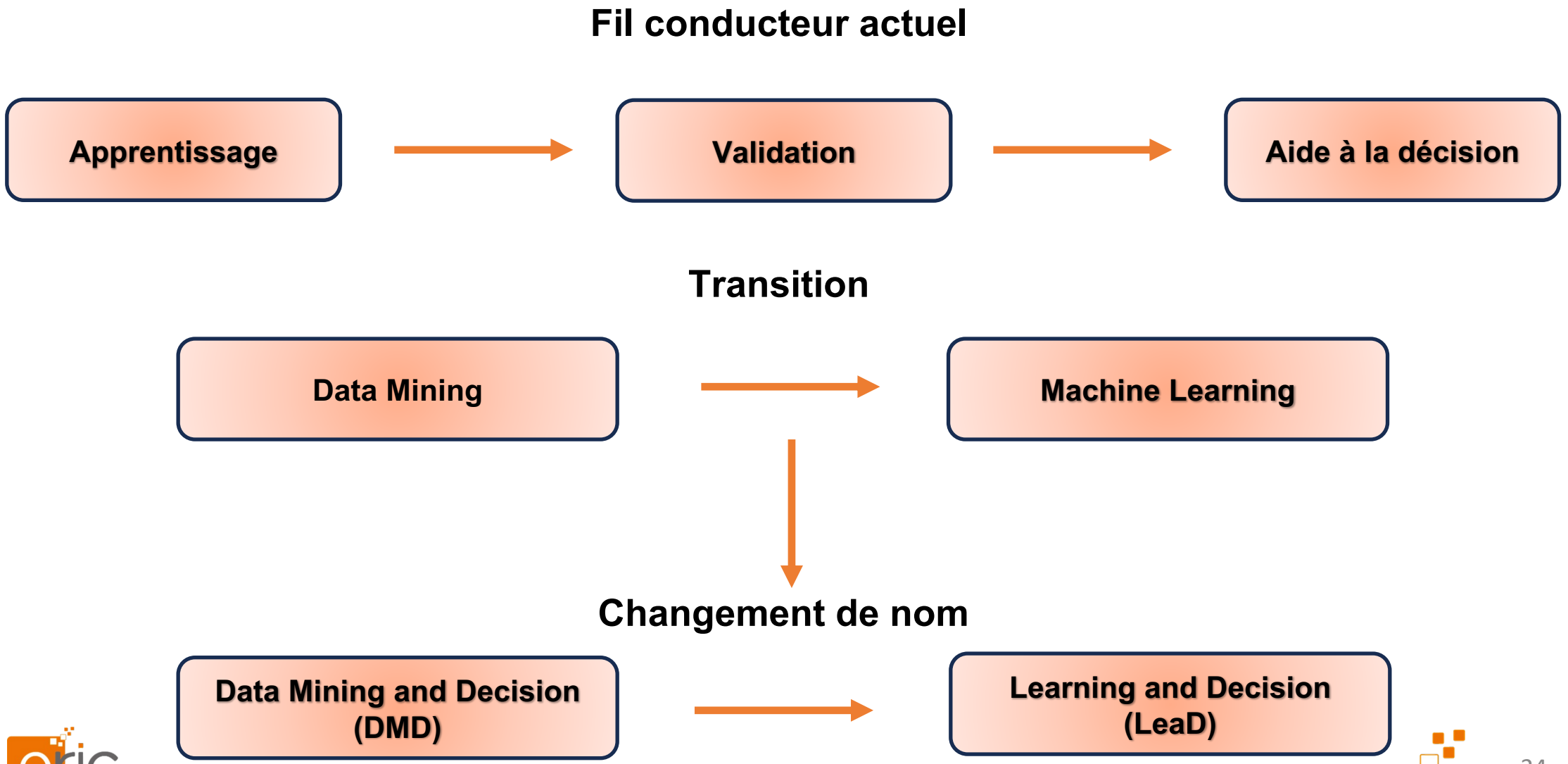
Model	λ	F1 score	Token F1 score	Subgroup AUC
BERT (6/12)	0	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	$0.59_{\pm 0.01}$
	0.01	$65.12_{\pm 0.38}$	$36.3_{\pm 4.01}$	$0.707_{\pm 0.11}$
	0.1	$65.92_{\pm 0.24}$	$39.26_{\pm 3.91}$	$0.784_{\pm 0.07}$
	1	$66.61_{\pm 0.17}$	$45.54_{\pm 3.29}$	$0.803_{\pm 0.12}$
DistilBERT (3/6)	0	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	$0.768_{\pm 0.24}$
	0.01	$64.35_{\pm 0.51}$	$40.4_{\pm 3.04}$	$0.748_{\pm 0.16}$
	0.1	$65.11_{\pm 0.7}$	$41.03_{\pm 3.28}$	$0.794_{\pm 0.31}$
	1	$66.71_{\pm 0.22}$	$42.67_{\pm 3.14}$	$0.796_{\pm 0.28}$
RoBERTa (6/12)	0	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	$0.519_{\pm 0.21}$
	0.01	$80.86_{\pm 0.22}$	$33.19_{\pm 3.28}$	$0.612_{\pm 0.29}$
	0.1	$78.58_{\pm 0.23}$	$36.49_{\pm 4.11}$	$0.681_{\pm 0.17}$
	1	$82.38_{\pm 0.26}$	$40.52_{\pm 3.81}$	$0.691_{\pm 0.14}$
DistilRoBERTa (3/6)	0	$71.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	$0.62_{\pm 0.08}$
	0.01	$79.14_{\pm 0.47}$	$34.41_{\pm 4.11}$	$0.634_{\pm 0.04}$
	0.1	$81.25_{\pm 0.33}$	$36.51_{\pm 3.5}$	$0.635_{\pm 0.08}$
	1	$81.96_{\pm 0.51}$	$43.02_{\pm 4.14}$	$0.65_{\pm 0.09}$

Performances et fairness (Subgroup AUC) avec contrainte d'alignement

Présentation équipe DMD

- Bilan de l'équipe (2019 – 2024)
- Contribution 1 : Détection d'outliers dans les données fonctionnelles
- Contribution 2 : Compression des réseaux de neurones
- **Trajectoire de l'équipe (2027 – 2031)**

DMD - Trajectoire



DMD - Trajectoire

Données complexes et hétérogènes

Graphes, Séries Temporelles et TAL

Données distribuées et confidentialité des données

Apprentissage : frugalité – transfert - renforcement

Alternatives aux modèles profonds

Apprentissage fédéré

Compression des modèles et biais

Apprentissage par renforcement

Apprentissage par transfert

Incertitude et équité

Bornes en généralisation

PAC(-Bayes), Prédiction Conforme

Fairness

Poursuivre les collaborations industrielles