

Supplementary material

A. Illustrative figures

An example of a functional dataset—the “Chinatown” dataset from the UCR [Chen et al. \(2015\)](#) repository—and evaluation points is plotted in Figure 1. Figure 2 depicts a sample isolation tree with a selected terminal node.

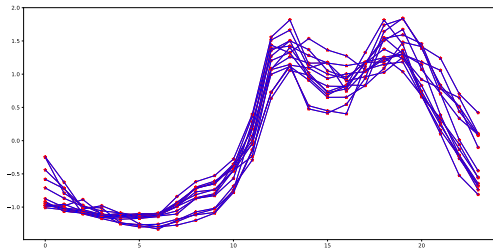


Figure 1: Example of a functional dataset. Red stars indicate the observed values (at the same equispaced time points), while the blue curves are obtained by linearly interpolating them.

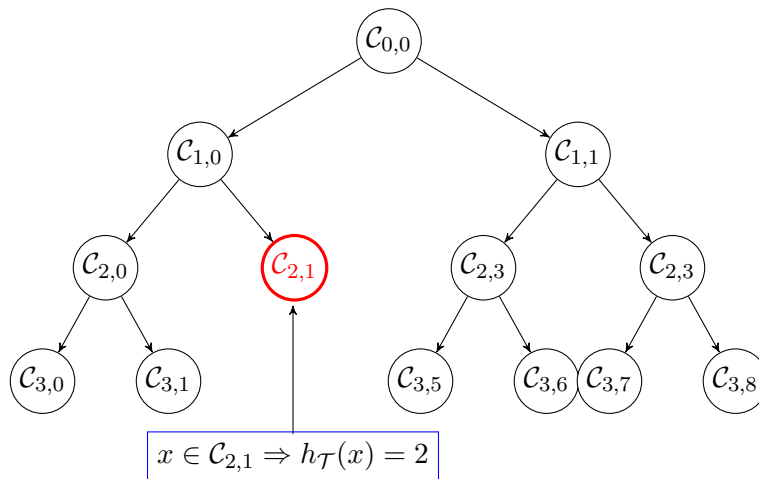


Figure 2: An example of a functional isolation tree structure denoted by \mathcal{T} . Here, $\mathcal{C}_{2,1}$ is a cell associated with a terminal node.

B. Presentation of used dictionaries

In this part, we define properly every dictionaries used in the paper.

Self-data dictionary (Self) consisting of the training dataset itself.

Brownian motion dictionary (B) is a combination of the space of continuous function $\mathcal{D} = \mathcal{C}([0, 1])$ and the Wiener measure \mathcal{W} on \mathcal{D} .

Brownian bridge dictionary (BB) is a combination of the space of continuous function $\mathcal{D} = \mathcal{C}([0, 1])$ and the Brownian bridge measure \mathcal{G} on \mathcal{D} .

Cosine dictionary (Cos) consisting of curves with the following forms:

$$\mathbf{x}_{a,\omega}(t) = a \cos(2\pi\omega t)$$

with $a \in [0, 1]$ and $\omega \in [0, 10]$.

Mexican hat wavelet dictionary (MHW) consists of the negative second derivatives of the normal density, shifted and scaled in a appropriate fashion:

$$\mathbf{x}_{\theta,\sigma}(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{t - \theta}{\sigma} \right)^2 \right) \exp \left(-\frac{(t - \theta)^2}{2\sigma^2} \right)$$

with $\theta \in [-0.8, 0.8]$ and $\sigma \in ([0.04, 0.2])$.

Dyadic indicator dictionary (DI) consisting of a set of indicator functions on the elements of binary partitioning, for a given J (chosen according to the granularity to be captured or from the discretisation considerations) having as elements $\{(\mathbf{x}_{k,j})_{0 \leq k < 2^j}\}_{1 \leq j \leq J}$:

$$\mathbf{x}_{k,j}(t) = \mathbb{1} \left(t \in \left[\frac{k}{2^j}, \frac{k+1}{2^j} \right] \right).$$

Uniform indicator dictionary (UI) consists of indicator function on $[a, b]$ where a and b are chosen uniformly on $[0, 1]$ such that $a < b$.

Dyadic indicator derivative (DI_d) consisting of a set of indicator functions on the elements of binary partitioning, for a given J (chosen according to the granularity to be captured or from the discretisation considerations) having as elements $\{(\mathbf{x}_{k,j})_{0 \leq k < 2^j}\}_{1 \leq j \leq J}$:

$$\mathbf{x}_{k,j}(t) = t \mathbb{1} \left(t \in \left[\frac{k}{2^j}, \frac{k+1}{2^j} \right] \right).$$

Uniform indicator derivative (UI_d) consists of functions $t \mapsto t$ on $[a, b]$ where a and b are chosen uniformly on $[0, 1]$ such that $a < b$.

C. Further discussion on the choice of dictionary

To illustrate the discussion on dictionaries, especially the incorporation of stochastic elements and external informations, we bring an example of the use of the *Brownian motion*

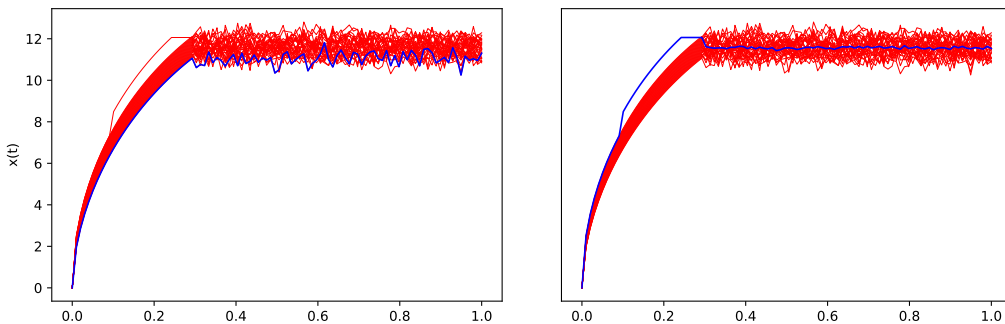


Figure 3: Synthetic data containing an isolated anomaly and the observation having the highest anomaly score, with dictionary being pure Brownian motion (left) and Brownian motion mixed with an indicator function in the area of interest in proportions 4 to 1 (right).

dictionary. Let \mathcal{W} be the Wiener measure defined on $\mathcal{C}([0, 1])$ the space of continuous function on $[0, 1]$ and \mathcal{H} be the L_2 space. We define by *Brownian motion dictionary* (B) the *Split variables* space induced by $\nu = \mathcal{W}$ and $\mathcal{D} = \mathcal{C}([0, 1])$. Although seeming universal, this dictionary explores almost the entire argument space equivalently, and in practice can be unable to detect *isolated anomalies*. On Fig. 3 we plot the following synthetic dataset:

- 30 curves defined by $\mathbf{x}(t) = 30(1 - t)^q t^q$ on $t \in [0, 0.2]$ and $\mathbf{x}(t) = 30(0.8)^q 0.2^q + \mathcal{N}(0, 0.3^2)$ on $t \in [0.2, 0.7]$ with q equispaced in $[0.5, 0.55]$.
- 1 *abnormal* curve with the same shape but that is shifted at the beginning and whose continuation is deep in the 30 preceding curves.

One can see that the anomaly is not detected, that indicated as anomaly curve (the one with highest anomaly score) is on the fringe of the dataset though. Illustrative incorporation of the prior knowledge, in its simplified version, can consist, *e.g.*, in adding to the measure \mathcal{W} , a Dirac of the indicator function on the interval of interest $\tilde{\mathbf{x}}(t) = \mathbb{1}(t \leq 0.25)$ with weights: $\tilde{\mathcal{W}} := 0.2\mathcal{W} + 0.8\delta_{\tilde{\mathbf{x}}}$; this assigns the highest anomaly score to the desired observation. In the sequel, ν follows a uniform distribution if is not explicitly mentioned.

When having not enough prior knowledge, *e.g.* just knowing to stick to local features of functional data but not the precise interval, one would like to use a dictionary exploring different localities. To illustrate possible advantage of this approach, we use *Mexican hat wavelet* and *Dyadic indicator* dictionaries. Regard the “Chinatown” dataset [Chen et al. \(2015\)](#), which represents pedestrian count in Chinatown-Swanston St North for 12 months during year 2017. With 14 functions (working days) representing normal observations and taking 4 functions (weekends) as anomalies (Figure 4). One observes that while the Mexican hat wavelet dictionary correctly detects part of the anomalies, due to its smooth nature it is distracted by two normal curves with high deviation on the second half of the domain.

Having straight fronts and begin non-zero only in a small part of the domain, the dyadic indicator dictionary detects all four abnormal observations. Nevertheless, it is not adapted to scalar product that involves derivative. To adapt the dyadic indicator dictionary and the uniform indicator dictionary for the scalar product involving derivatives, we define their *slope* versions being $\mathbf{x}_{k,j}(t) = t\mathbb{1}\left(t \in \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]\right)$ and $\mathbf{x}(t) = t\mathbb{1}(t \in [t_1, t_2])$, respectively, with the notation defined above (*i.e.* their derivatives become indicator functions). Clearly, this list can be extended with further task-specific dictionaries.

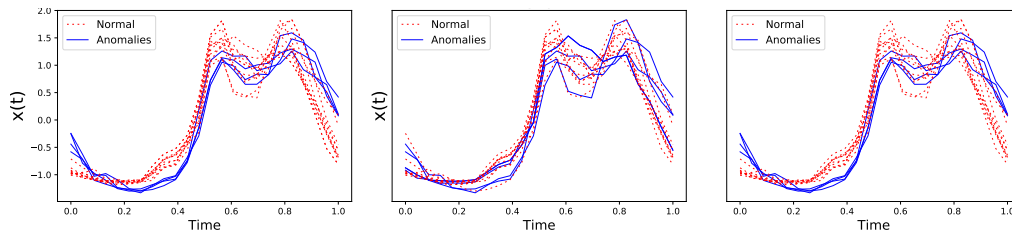


Figure 4: The “Chinatown” dataset, with normal observations in red and anomalies in blue: The data and the true anomalies (left), anomalies detected using the Mexican hat wavelet dictionary (middle), and anomalies detected using the dyadic indicator dictionary (right).

Before, we were considering dictionaries that are independent of data. Nevertheless one can use observations or their certain transform as a dictionary itself: projections on both normal and abnormal observations shall differ for normal ones and for anomalies; this suggests the *self-data dictionary* (Self). This can be extended to the *local self-data dictionary* which consists of the product of the self-data dictionary with the uniform indicator dictionary. As an example, we apply this to the “ECG5000” dataset plotted in Figure 5, where, different to the cosine dictionary, it allows to detect all abnormal observations.

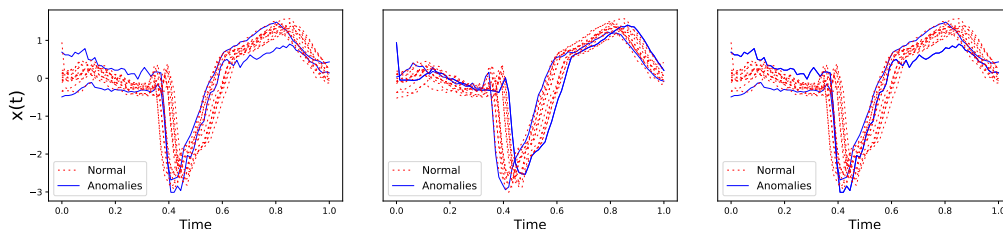


Figure 5: The “ECG5000” dataset, with normal observations in red and anomalies in blue: The data and the true anomalies (left), anomalies detected using the cosine dictionary (middle), and anomalies detected using the self data dictionary (right).

To conclude this, we provide a last example (see Fig. 6) where we highlight the impact of the scalar product choice. To illustrate the score change caused by different values of α ,

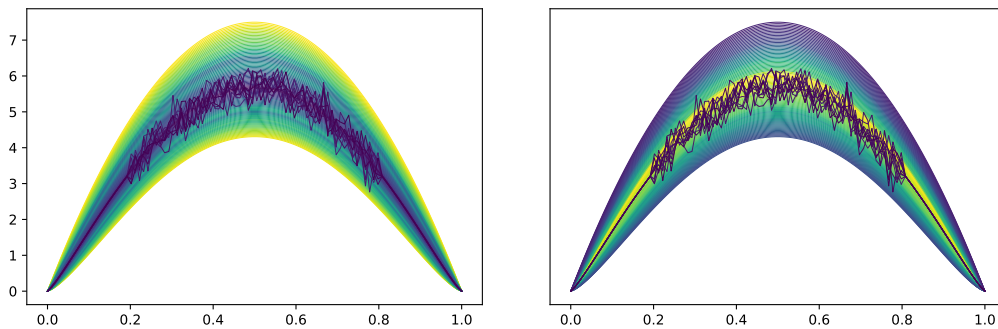


Figure 6: FIF anomaly scores for a sample of 100 curves with $\alpha = 1$ (left) and $\alpha = 0$ (right). Anomaly score increases from magenta to yellow in the left plot and decreases in the right plot.

we calculate the FIF anomaly scores with $\alpha = 1$ and $\alpha = 0$ for a sample consisting of 100 curves as follows (inspired by Cuevas et al. (2007), see Fig. 6):

- 90 curves defined by $\mathbf{x}(t) = 30(1 - t)^q t^q$ with q equispaced in $[1, 1.4]$,
- 10 *abnormal* curves defined by $\mathbf{x}(t) = 30(1 - t)^{1.2} t^{1.2}$ noised by $\varepsilon \sim \mathcal{N}(0, 0.3^2)$ on the interval $[0.2, 0.8]$.

One can see that even though the 10 noisy curves are abnormal for the majority of the data, they are considered as normal ones when only location is taken into account. On the other hand, they are easily distinguished with the high anomaly score when derivatives are examined.

C.1. Direction importance of finite size Dictionaries

Although feature importance have been tackled in supervised random trees (see *e.g.* Breiman (2001), Geurts et al. (2006)), this has not been adressed in the *Isolation Forest* literature (see Liu et al. (2008), Liu et al. (2012) and Hariri et al. (2018)). As a very randomized procedure, there is no incrementally way to define feature importance from the supervised setting. Nevertheless, it is a matter of interest in many anomaly detection applications to get interpretability of models, especially when dealing with functional data where many information are contained in curves. Thus, it is rewarding to get an *a posteriori* sparse representation of the dictionary \mathcal{D} which corresponds to the discriminating directions that have great importance in the construction of the model. Furthermore, it could bring some information on the distribution of normal data by studying the dispersion of the projection coefficients on a direction d (*e.g.* multi-modality). To extend this notion to the *Functional Isolation Forest* algorithm, we propose two ways to evaluate the importance of the elements of \mathcal{D} to discriminate anomaly curves. The general idea is to give importance to elements of \mathcal{D} which allows to discriminate between the sample. The naive idea is to add "+1" to the elements of \mathcal{D} where an instance of the node sample is isolated (except for the cells with

only two instances) such that good directions are those with a high score (after the forest construction). A clever one, more adaptive, would be to get weighted gain since curves isolated at nodes closer to the root should be more rewarding. To do this, we choose to give a reward depending on the size of the sample node where a curve is isolated. Precisely, the given reward is equal to the size of the node sample divided by the (sub)-sample used to build the tree. An example of the latter is given in Figure 7. The experiment is conducted on the real-world CinECGTorso dataset (more details in Section 4.2). We use FIF with the *Dyadic indicator dictionary* and the L_2 scalar product. As we can see, the two most important elements of the dictionary are indicator functions which localize the peak around $t = 0.4$ where anomalies are really different from the normal ones. These leads to some interpretability of a "black box" procedure.

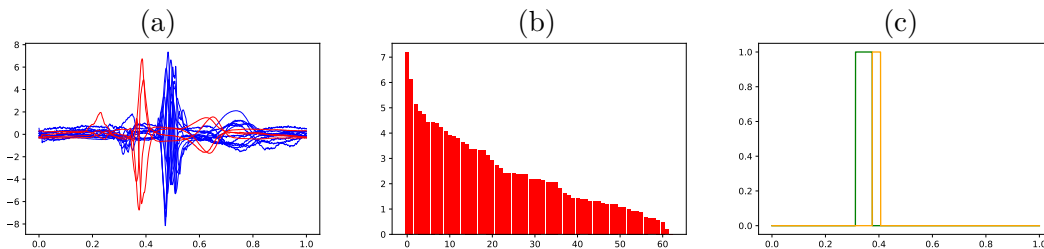


Figure 7: CinECGTorso training dataset (a), red curves correspond to anomalies while blue curves to normal data. The direction importances given by the "adaptive" way are represented by (b) and the two most important functions (from the dyadic dictionary) used by FIF to build the model are plotted in (c).

D. Study of the parameters of FIF

In this section, we present results of a simulation study of the variance of the FIF algorithm. The experiments were conducted on the datasets (a) and (b) from Section 4 (see also Figure 3 of Section 4), for each of the four specified observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ using the following settings (except varying parameter):

Dictionary: Gaussian wavelets (negative second derivative of the standard Gaussian density) with random variance selected in an uniform way in $[0.2, 1]$ and a translation parameter selected randomly in $[-4, 4]$. We fixed the size of the dictionary to 1000.

Scalar product: L^2 dot product.

Size of the dataset: $n = 500$.

Subsampling size: $\psi = 64$.

The number of trees: $N = 100$.

The height limit: fixed to $l = \lceil \log_2(\psi) \rceil$.

The figures below indicate boxplots of the FIF anomaly score, over 100 runs. Empirical study of the FIF anomaly score and its variance when increasing the number of F-trees is depicted in Figure 8.

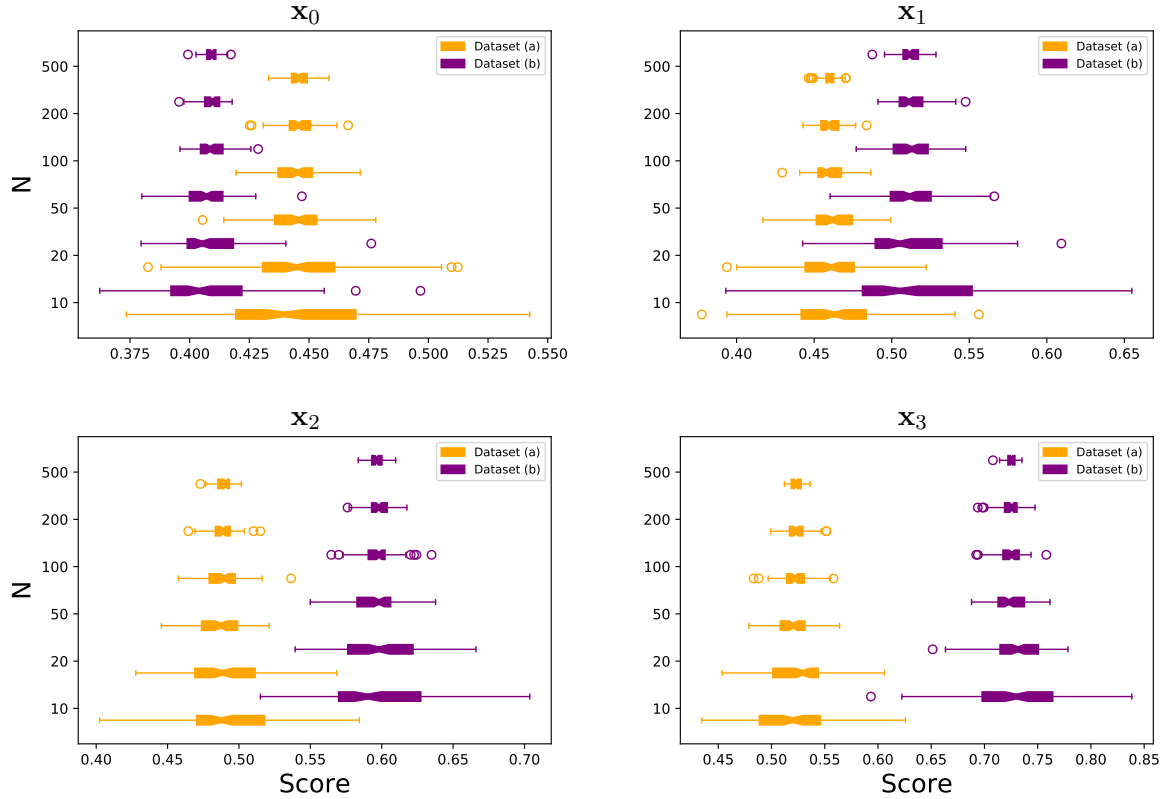


Figure 8: Boxplot (over 100 repetitions) of the FIF score for the observations x_0, x_1, x_2, x_3 for different sample sizes. The orange boxplots represent the dataset (a) while the purple boxplots represent the dataset (b).

Empirical study of the FIF anomaly score and its variance when increasing the subsample size is depicted in Figure 9.

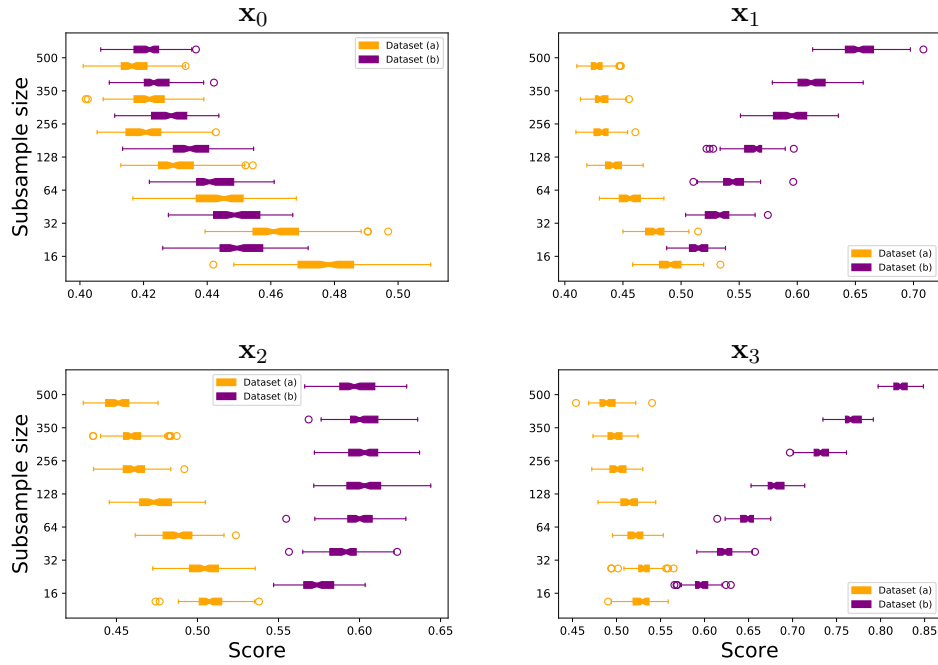


Figure 9: Boxplot (over 100 repetitions) of the FIF score for the observations x_0, x_1, x_2, x_3 for different sample sizes. The orange boxplots represent the dataset (a) while the purple boxplots represent the dataset (b).

Empirical study of the FIF anomaly score and its variance when increasing the height limit of the F- \hat{t} tree is depicted in Figure 10.

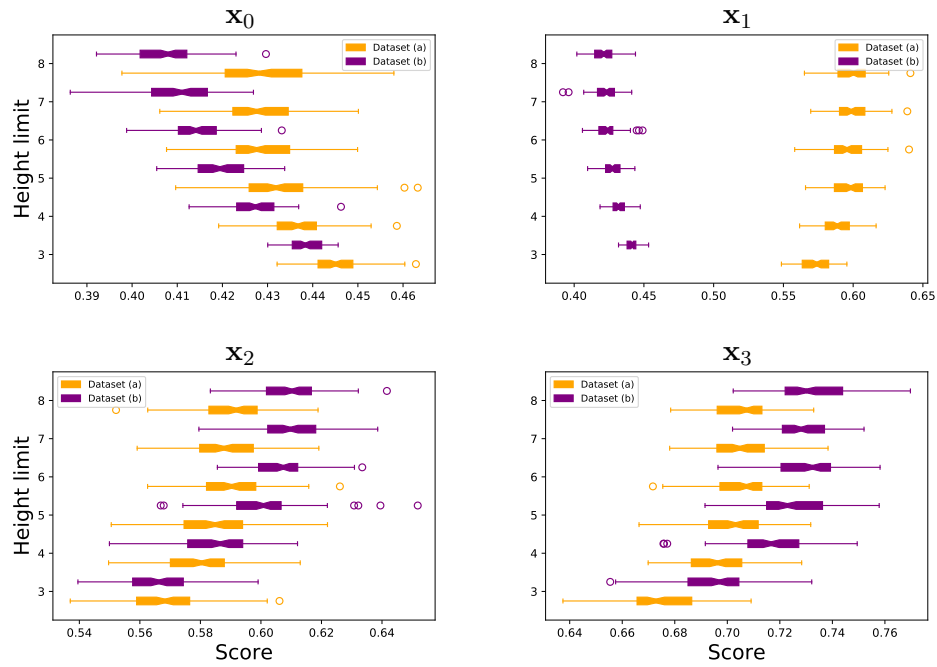


Figure 10: Boxplot (over 100 repetitions) of the FIF score for the observations x_0, x_1, x_2, x_3 for different sample sizes. The orange boxplots represent the dataset (a) while the purple boxplots represent the dataset (b).

Taking finite size versions of the infinite *gaussian wavelets* dictionary, an empirical study of the FIF anomaly score and its variance when increasing the size of the dictionary is depicted in Figure 11.

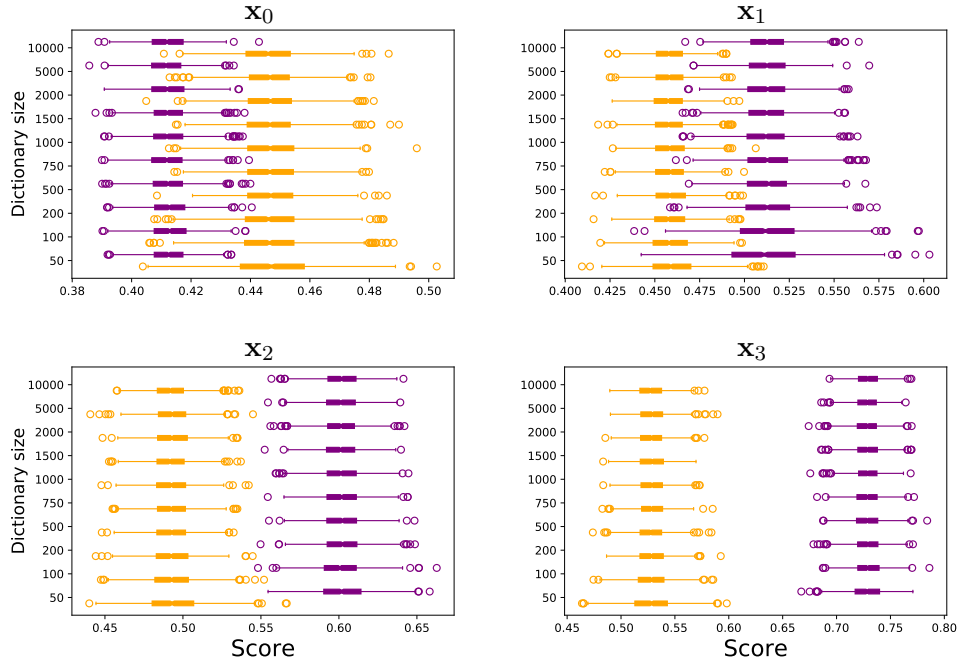


Figure 11: Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different sample sizes. The orange boxplots represent the dataset (a) while the purple boxplots represent the dataset (b).

Empirical study of the FIF anomaly score for a variety of dictionaries with the L_2 scalar product of the derivatives is depicted in Figure 12.

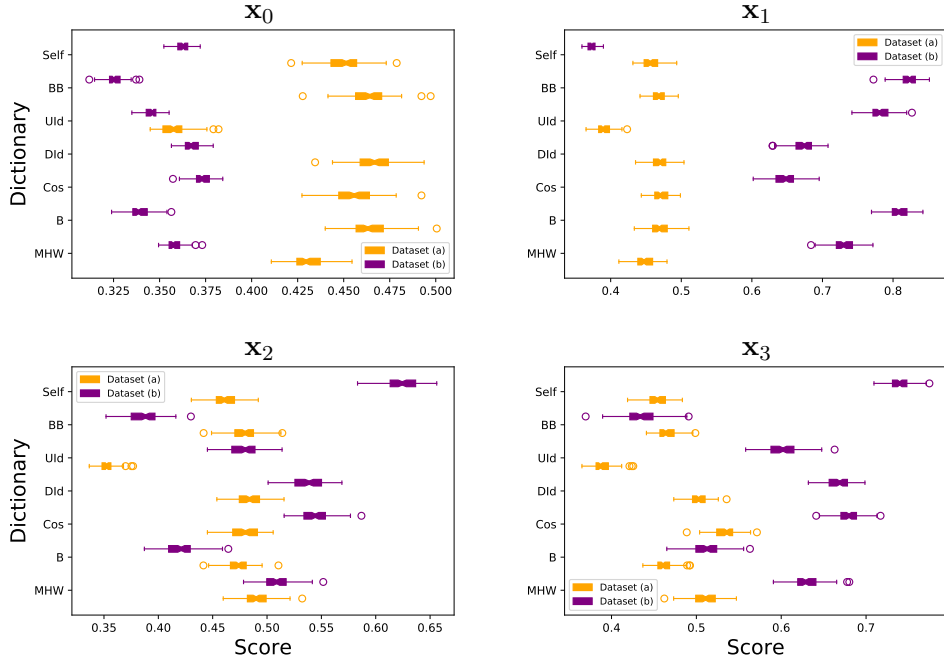


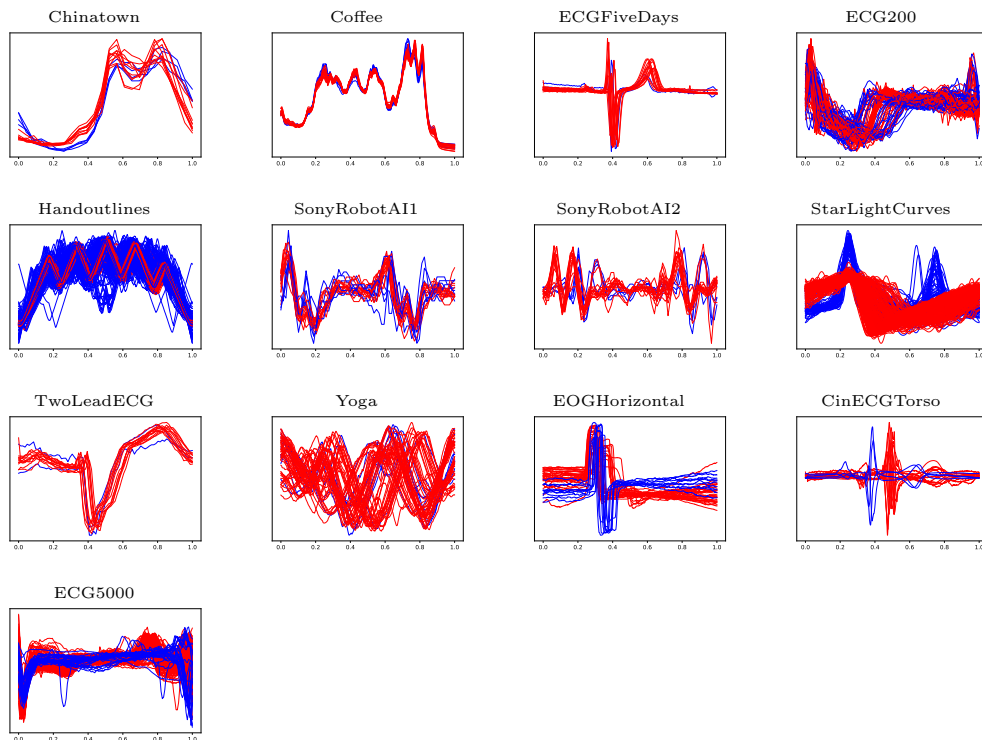
Figure 12: Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different dictionaries using the L_2 scalar product of the derivatives. The orange boxplots represent the dataset (a) while the purple boxplots represent the dataset (b).

Analysis of the results of Section C In a first experiment, we show the boxplots of the score estimated by FIF when increasing the number of F-*trees* and observe that, as expected, the variance diminishes when N grows (see Figure 8). We also see in Figure 9 that with an increasing subsample size ψ the FIF anomaly score increases for anomalies since these are more often present in the subsample and thus isolated faster (with shorter path length) when calculating the score than when they were absent in the subsample; this effect is reciprocal for normal observations. A similar behavior is observed with increasing height limit l in Figure 10. The variance of the score tends to slightly increase with ψ and l because of more observations/branching possibilities. If the dictionary is sufficiently rich, its size does not influence the FIF anomaly score and its variance stabilizes relatively fast while growing the size of the dictionary (see Figure 11) which encourages the use of massive (and infinite size) dictionaries.

E. Complementary results on the performance comparison

E.1. benchmark datasets

Here, we plot the thirteen benchmark train datasets used in the experiment. Anomalies are represented by blue color while normal data are drawn in red.



E.2. Functional depth

In this part, we present further 8 functional depths which are outperformed (on an average) by the two depth functions presented in Section 4 on the 13 real-world datasets and we display their AUC performance. SFD corresponds to *simplicial integrated depth*, HFD to *Halfspace integrated depth*, RP-SD to the *random projection method with simplicial depth*, RP-RHD to the *random projection method with random halfspace depth*, fAO to the *functional adjusted outlyingness*, fDO to the *functional directional outlyingness* to and fbd to *functional bagdistance*. The reader is referred to Cuevas et al. (2007); Fraiman and Muniz (2001); Hubert et al. (2015) for the bibliography on employed functional data depth notions.

FIF-SUPPLEMENTARY

| Datasets | db35 | db36 | db37 | db38 | dmev | haar | rbio1.1 | rbio1.3 | rbio1.5 | rbio2.2 | rbio2.4 | rbio2.6 | rbio2.8 | rbio3.1 | rbio3.3 | rbio3.5 | rbio3.7 | rbio3.9 | rbio4.4 | rbio5.5 | rbio6.8 | sym2 |
|-----------------|------|------|------|------|------|------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|------|
| Chinatow | 0.71 | 0.69 | 0.77 | 0.74 | 0.74 | 0.71 | 0.71 | 0.69 | 0.72 | 0.78 | 0.74 | 0.72 | 0.81 | 0.75 | 0.73 | 0.74 | 0.61 | 0.72 | 0.79 | 0.67 | 0.81 | 0.66 |
| Coffee | 0.50 | 0.38 | 0.41 | 0.50 | 0.54 | 0.53 | 0.58 | 0.46 | 0.50 | 0.45 | 0.67 | 0.49 | 0.63 | 0.33 | 0.42 | 0.49 | 0.56 | 0.49 | 0.54 | 0.53 | 0.54 | 0.47 |
| ECGFiveDays | 0.84 | 0.85 | 0.83 | 0.84 | 0.82 | 0.88 | 0.86 | 0.88 | 0.81 | 0.85 | 0.65 | 0.90 | 0.82 | 0.92 | 0.82 | 0.79 | 0.87 | 0.84 | 0.88 | 0.77 | 0.78 | 0.81 |
| ECG200 | 0.78 | 0.80 | 0.78 | 0.83 | 0.79 | 0.78 | 0.80 | 0.76 | 0.78 | 0.72 | 0.68 | 0.70 | 0.72 | 0.81 | 0.71 | 0.76 | 0.75 | 0.72 | 0.75 | 0.73 | 0.82 | 0.79 |
| Handoutlines | 0.72 | 0.72 | 0.70 | 0.74 | 0.75 | 0.70 | 0.65 | 0.70 | 0.71 | 0.75 | 0.69 | 0.71 | 0.73 | 0.74 | 0.71 | 0.73 | 0.73 | 0.67 | 0.72 | 0.70 | 0.68 | 0.75 |
| SonyRobotAI1 | 0.86 | 0.87 | 0.83 | 0.84 | 0.87 | 0.74 | 0.80 | 0.87 | 0.80 | 0.86 | 0.76 | 0.81 | 0.79 | 0.81 | 0.77 | 0.85 | 0.79 | 0.84 | 0.79 | 0.80 | 0.86 | 0.80 |
| SonyRobotAI2 | 0.82 | 0.85 | 0.86 | 0.81 | 0.84 | 0.88 | 0.86 | 0.86 | 0.85 | 0.83 | 0.85 | 0.86 | 0.85 | 0.81 | 0.88 | 0.92 | 0.87 | 0.88 | 0.83 | 0.85 | 0.81 | 0.86 |
| StarLightCurves | 0.75 | 0.70 | 0.73 | 0.73 | 0.73 | 0.77 | 0.76 | 0.79 | 0.77 | 0.77 | 0.75 | 0.74 | 0.78 | 0.72 | 0.76 | 0.73 | 0.74 | 0.75 | 0.77 | 0.76 | 0.73 | 0.74 |
| TwoLeadECG | 0.69 | 0.58 | 0.62 | 0.67 | 0.57 | 0.70 | 0.66 | 0.68 | 0.71 | 0.63 | 0.63 | 0.67 | 0.66 | 0.66 | 0.61 | 0.73 | 0.71 | 0.62 | 0.62 | 0.64 | 0.66 | 0.63 |
| Yoga | 0.57 | 0.60 | 0.59 | 0.56 | 0.59 | 0.58 | 0.59 | 0.59 | 0.59 | 0.58 | 0.58 | 0.58 | 0.59 | 0.56 | 0.60 | 0.58 | 0.60 | 0.57 | 0.55 | 0.56 | 0.59 | 0.56 |
| EOGHorizontal | 0.74 | 0.71 | 0.71 | 0.70 | 0.68 | 0.69 | 0.70 | 0.64 | 0.74 | 0.68 | 0.65 | 0.66 | 0.68 | 0.71 | 0.75 | 0.72 | 0.68 | 0.67 | 0.69 | 0.76 | 0.72 | 0.74 |
| CinECGTorso | 0.42 | 0.48 | 0.38 | 0.51 | 0.47 | 0.48 | 0.53 | 0.42 | 0.44 | 0.41 | 0.36 | 0.35 | 0.41 | 0.30 | 0.39 | 0.30 | 0.49 | 0.36 | 0.41 | 0.42 | 0.42 | 0.44 |
| ECG5000 | 0.95 | 0.95 | 0.96 | 0.95 | 0.95 | 0.95 | 0.94 | 0.94 | 0.95 | 0.96 | 0.94 | 0.95 | 0.95 | 0.95 | 0.94 | 0.93 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |

| Datasets | sym3 | sym4 | sym5 | sym6 | sym7 | sym8 | sym9 | sym10 | sym11 | sym12 | sym13 | sym14 | sym15 | sym16 | sym17 | sym18 | sym19 | sym20 |
|-----------------|------|------|------|------|------|------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Chinatow | 0.67 | 0.70 | 0.74 | 0.67 | 0.80 | 0.72 | 0.70 | 0.74 | 0.76 | 0.65 | 0.79 | 0.74 | 0.71 | 0.83 | 0.69 | 0.62 | 0.77 | 0.66 |
| Coffee | 0.65 | 0.62 | 0.64 | 0.63 | 0.49 | 0.50 | 0.41 | 0.38 | 0.60 | 0.46 | 0.45 | 0.56 | 0.55 | 0.56 | 0.46 | 0.46 | 0.56 | 0.55 |
| ECGFiveDays | 0.90 | 0.81 | 0.81 | 0.81 | 0.90 | 0.89 | 0.81 | 0.75 | 0.78 | 0.79 | 0.80 | 0.88 | 0.83 | 0.82 | 0.81 | 0.83 | 0.78 | 0.85 |
| ECG200 | 0.76 | 0.76 | 0.81 | 0.72 | 0.76 | 0.77 | 0.72 | 0.78 | 0.77 | 0.78 | 0.82 | 0.79 | 0.81 | 0.78 | 0.77 | 0.81 | 0.83 | 0.80 |
| Handoutlines | 0.75 | 0.73 | 0.70 | 0.74 | 0.71 | 0.74 | 0.69 | 0.71 | 0.73 | 0.72 | 0.73 | 0.70 | 0.69 | 0.74 | 0.72 | 0.75 | 0.72 | 0.72 |
| SonyRobotAI1 | 0.75 | 0.85 | 0.82 | 0.76 | 0.82 | 0.76 | 0.86 | 0.86 | 0.83 | 0.83 | 0.90 | 0.86 | 0.83 | 0.88 | 0.90 | 0.88 | 0.84 | 0.83 |
| SonyRobotAI2 | 0.87 | 0.86 | 0.84 | 0.85 | 0.86 | 0.86 | 0.84 | 0.85 | 0.81 | 0.87 | 0.87 | 0.91 | 0.88 | 0.85 | 0.85 | 0.87 | 0.85 | 0.87 |
| StarLightCurves | 0.76 | 0.77 | 0.77 | 0.76 | 0.76 | 0.75 | 0.77 | 0.76 | 0.77 | 0.76 | 0.75 | 0.74 | 0.70 | 0.74 | 0.76 | 0.73 | 0.78 | 0.71 |
| TwoLeadECG | 0.66 | 0.68 | 0.58 | 0.62 | 0.71 | 0.69 | 0.69 | 0.64 | 0.66 | 0.69 | 0.64 | 0.67 | 0.65 | 0.64 | 0.68 | 0.63 | 0.67 | 0.70 |
| Yoga | 0.58 | 0.57 | 0.59 | 0.60 | 0.58 | 0.59 | 0.58 | 0.58 | 0.58 | 0.59 | 0.59 | 0.57 | 0.57 | 0.58 | 0.59 | 0.58 | 0.58 | 0.59 |
| EOGHorizontal | 0.72 | 0.68 | 0.68 | 0.70 | 0.70 | 0.71 | 0.67 | 0.70 | 0.69 | 0.71 | 0.70 | 0.70 | 0.69 | 0.65 | 0.69 | 0.68 | 0.72 | 0.65 |
| CinECGTorso | 0.45 | 0.44 | 0.51 | 0.41 | 0.49 | 0.56 | 0.39 | 0.48 | 0.49 | 0.37 | 0.46 | 0.48 | 0.43 | 0.58 | 0.36 | 0.41 | 0.48 | 0.40 |
| ECG5000 | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |

F. Multivariate Functional Isolation Forest

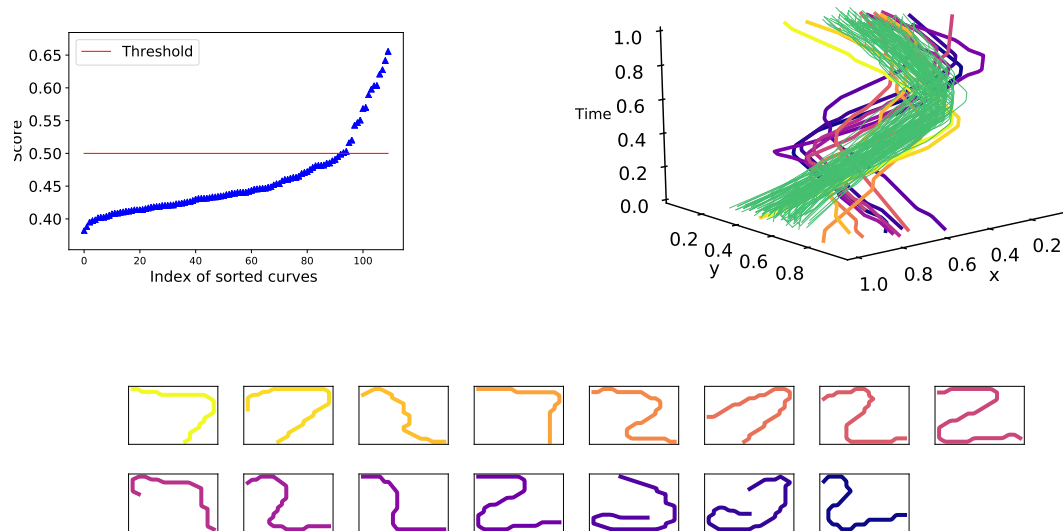


Figure 13: FIF anomaly scores for a sample of 110 digits (100 seven and 10 two). Left plot corresponds to the sorted score of these curves. Right plot represents the digits in three dimensions, green ones correspond to normal data, anomalies score increases from orange to dark red. Bottom plot shows the fifteen detected anomalies.

FIF can be easily extended to the multivariate functional data, *i.e.* when the quantity of interest lies in \mathbb{R}^d for each moment of time:

$$\begin{aligned} X : \Omega &\longrightarrow (\mathcal{H}([0, 1]))^{\otimes d} \\ \omega &\longmapsto ((X_1(\omega))_{t \in [0, 1]}, \dots, (X_d(\omega))_{t \in [0, 1]}) \end{aligned}$$

For this, the coordinate-wise sum of the d corresponding scalar products is used to project the data onto a chosen dictionary element:

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}^{\otimes d}} := \sum_{i=1}^d \langle f^{(i)}, g^{(i)} \rangle_{\mathcal{H}}.$$

Further, a dictionary should be defined in $(\mathcal{H}([0, 1]))^{\otimes d}$. This can be done, *e.g.*, by either componentwise application of one or several univariate dictionaries from Sections 3 and B, or by constructing of special d -variate ones. For illustration purposes, regard the following example constructed based on the MNIST [Lecun et al. \(1998\)](#) dataset. First, we extract the digits' contours (skeletons) using skimage python library [van der Walt et al. \(2014\)](#). Then each observation is transformed into a curve in $(L_2([0, 1]) \times L_2([0, 1]))$ (one

vertical and one horizontal coordinates) using length parametrization on $[0, 1]$. We construct the problem by taking 100 curves from class 7 and adding 10 observations from class 2. We apply FIF with two-dimensional *sinecosine* dictionary and the following scalar product: $\langle \mathbf{f}, \mathbf{g} \rangle_{(L_2)^{\otimes d}} = \sum_{i=1}^d \langle f^{(i)}, g^{(i)} \rangle_{L_2}$. *sinecosine* is construct as a direct extension of *cosine* dictionary introduced for FIF by selecting randomly cosine or sine function on each coordinates. Figure 13 shows anomaly detection using the visual elbow rule to define the threshold. Among those detected, five digits are indeed 7s, but do not resemble them and thus are identified as anomalies.

References

- L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, July 2015. URL www.cs.ucr.edu/~eamonn/time_series_data/.
- A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007.
- R Fraiman and G Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, Dec 2001.
- P Geurts, D Ernst, and L Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr 2006.
- S. Hariri, M. Carrasco Kind, and R. J. Brunner. Extended isolation forest. *ArXiv e-prints*, 2018. URL <https://arxiv.org/abs/1811.02141>.
- M. Hubert, P.J. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, 2015.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE Computer Society, 2008.
- F. T. Liu, K. M. Ting, and Z. Zhou. Isolation-based anomaly detection. In *ACM Transactions on Knowledge Discovery from Data (TKDD)*, volume 6, pages 1–39, 2012.
- S van der Walt, J.L. Schönberger, J Nunez-Iglesias, F Boulogne, J.D. Warner, N Yager, E Gouillart, and T Yu. scikit-image: image processing in python. *PeerJ*, 2, 2014.