6th IEEE International Conference on Big Data and Smart Computing

IEEE BigComp2019

Learning Representations from Healthcare Time Series Data for Unsupervised Anomaly Detection

João Pereira & Margarida Silveira

Signal and Image Processing Group
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal

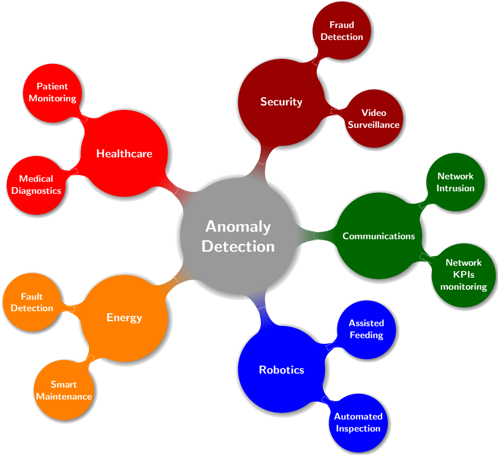
TÉCNICO LISBOA

Kyoto, Japan
February 28th, 2019

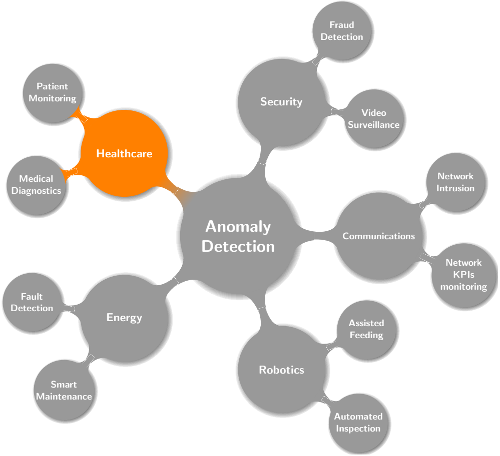
Institute for Systems and Robotics Lisboa

Introduction: Anomaly Detection

Anomaly detection is about finding patterns in data that do not conform to *expected* or *normal* behaviour.



Introduction: Anomaly Detection



- Most data in the world are **unlabelled**

anomaly labels

$$\text{Dataset } \mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)} \right) \right\}_{i=1}^{N}$$

- Annotating large datasets is difficult, time-consuming and expensive

- Time series have temporal structure/dependencies

$$\mathbf{x} = \left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T \right) \quad , \qquad \mathbf{x}_t \in \mathbb{R}^{d_\mathbf{x}}$$

Introduction: Main concepts

- Representation Learning;
- Autoencoders;
  - Variational Autoencoder (VAE)
- Recurrent Neural Networks (RNN);
  - Long Short-Term Memory Network (LSTM)

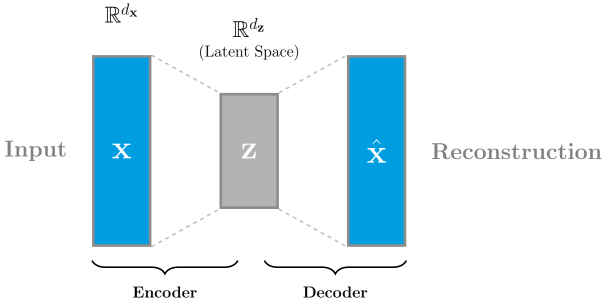
Introduction: Representation Learning

Learning good data representations is important.

- Representations are useful for downstream tasks (e.g., regression and classification);
- Make models more expressive and more accurate;
- Dismiss hand-designed features and representations;
- Neural networks are powerful representation learning models.

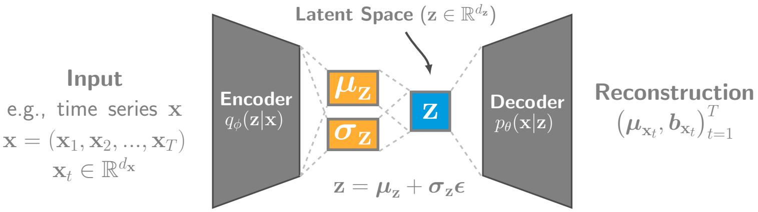
Autoencoders

- Aim to reconstruct their input $\mathbf{x}$
- Two parts: an *encoder* and a *decoder*



- Parameterized by a feed-forward NN, a CNN, a RNN, ...
- Loss function measures the quality of the reconstructions
- Often under-complete ($d_{\mathbf{z}} < d_{\mathbf{x}}$) $\rightarrow$ dimensionality reduction

- Deep generative model rooted in Bayesian inference



$$p_\theta(\mathbf{x}) = \int_\mathbf{z} p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} \qquad p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})}{p_\theta(\mathbf{x})}$$

The evidence and the posterior are intractable! 🙁

Kingma & Welling, **Auto-Encoding Variational Bayes**, ICLR'14

Rezende *et al.*, **Stochastic Backpropagation and Approximate Inference in Deep Generative Models**, ICML'14

VAE Training Objective

Objective: Maximize the Evidence Lower Bound (ELBO)

Reconstruction term

Regularization term (KL loss)

$$\log p_\theta(\mathbf{x}) \geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\big)}_{:=\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi;\ \mathbf{x})}$$

$\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence between the approximate posterior and the prior.

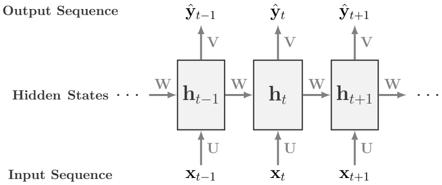
Recurrent Neural Networks

What if data are not i.i.d. in time?

(e.g., time series, text, videos)

RNNs capture the temporal dependencies of the data

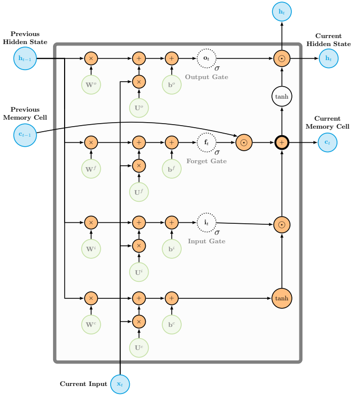
- Real-valued hidden state $\mathbf{h}_t$
- Feedback connection
- Parameters shared across timesteps



$$\mathbf{h}_t = f(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1})$$

f is often a $\tanh$ or $\mathrm{sigmoid}$

- Proposed to solve the vanishing gradient problem
- New cell and three gates



Hochreiter & Schmidhuber, **Long Short-Term Memory**, Neural Computation'97

Graves *et al.*, **Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition**, ICANN'05

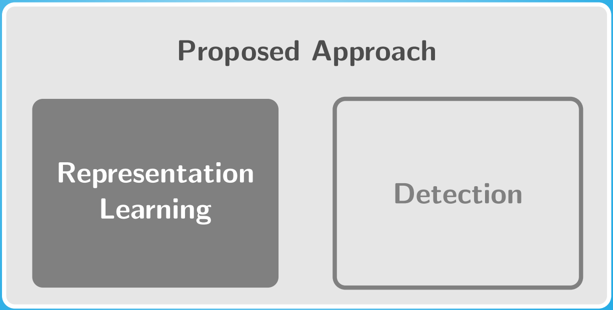
The Principle in a Nutshell

- Based on a **Variational Autoencoder**;
- Encoder and decoder are Bi-LSTMs;
- Train a VAE on mostly **normal** data;
- Learns a normal data manifold;
- Anomaly detection in the latent (representations) space.

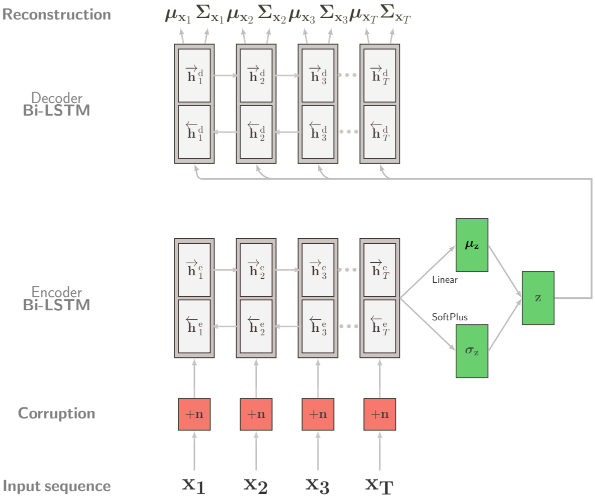
Proposed Approach

Representation Learning

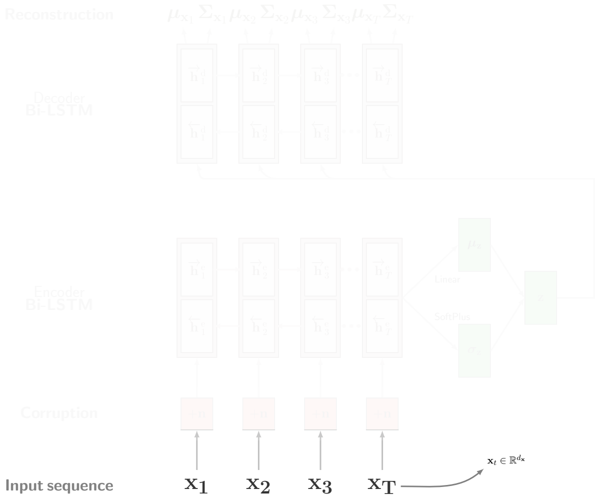
Detection


Proposed Approach
Representation Learning
Detection

Representation Learning



Representation Learning



Denoising Autoencoding Criterion

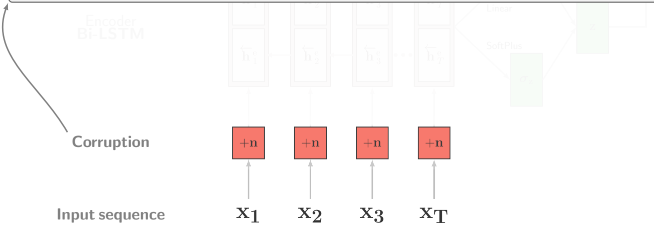
Corruption process: additive Gaussian noise

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \mathbf{x} + \mathbf{n} \quad , \quad \mathbf{n} \sim \mathrm{Normal}(\mathbf{0}, \sigma_{\mathbf{n}}^2 \mathbf{I})$$

♣

Vincent *et al.*, **Extracting and Composing Robust Features with Denoising Autoencoders**, ICML'08

Bengio *et al.*, **Denoising Criterion for Variational Auto-Encoding Framework**, ICLR'15



Learning temporal dependencies

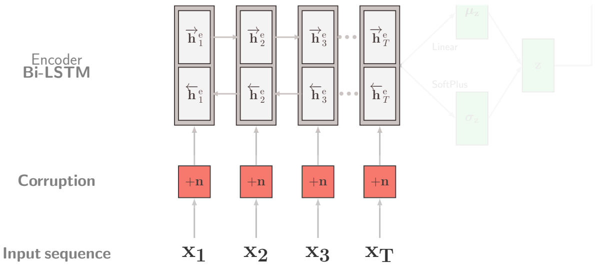
Bidirectional Long-Short Term Memory network

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right]$$

- 256 units, 128 in each direction
- Sparse regularization, $\Omega(\mathbf{z}) = \lambda \sum_{i=1}^{d_\mathbf{z}} |z_i|$

Hochreiter *et al.*, **Long-Short Term Memory**, Neural Computation'97

Graves *et al.*, **Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition**, ICANN'05



Variational Latent Space

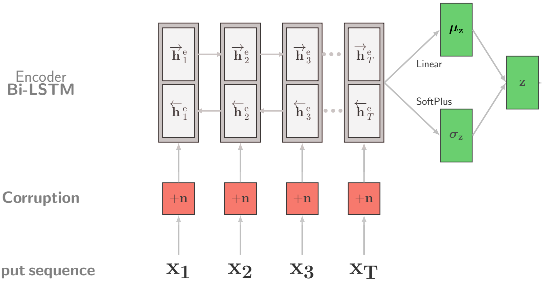
Variational parameters derived using neural networks

$$(\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\sigma}_{\mathbf{z}}) = \text{Encoder}(\mathbf{x})$$

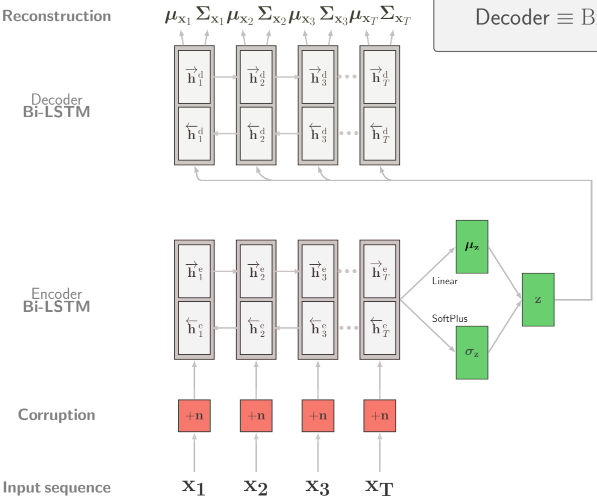
Sample from the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$

$$\mathbf{z} = \boldsymbol{\mu}_{\mathbf{z}} + \boldsymbol{\sigma}_{\mathbf{z}} \odot \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$$

Kingma & Welling, **Auto-Encoding Variational Bayes**, ICLR'14



Representation Learning



Decoder

Another Bi-LSTM.

$$\text{Decoder} \equiv \text{Bi-LSTM}(\mathbf{z})$$

Representation Learning

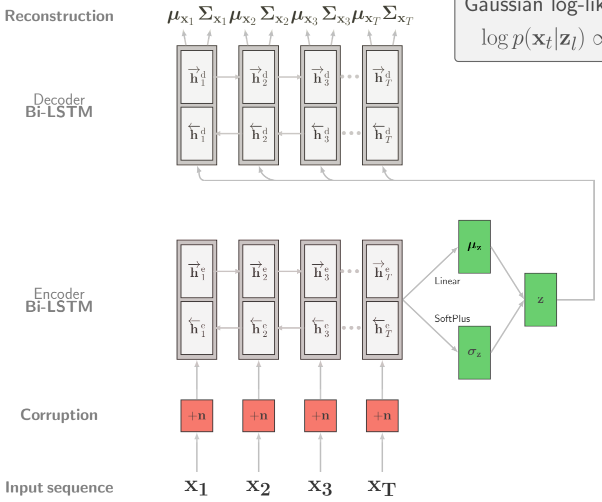
Output

Reconstruction parameters:

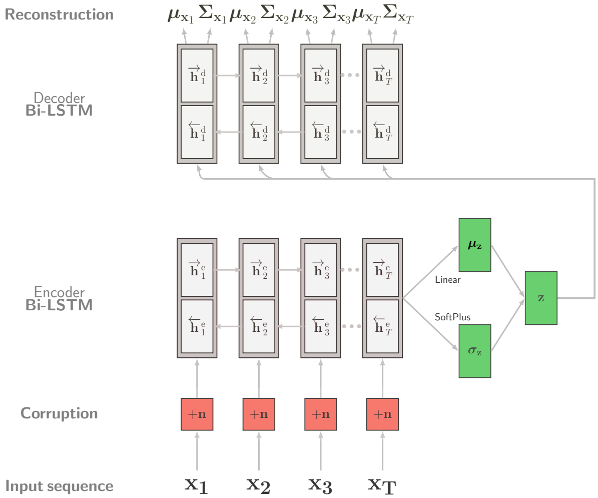
$$\left(\boldsymbol{\mu}_{\mathbf{x}_t}, \boldsymbol{\Sigma}_{\mathbf{x}_t}\right)_{t=1}^{T}$$

Gaussian log-likelihood:

$$\log p(\mathbf{x}_t|\mathbf{z}_l) \propto \|\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x}_t}\|_2^2$$



Representation Learning



Loss Function

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -\mathbb{E}_{\mathbf{z}\sim\tilde{q}_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta\big(\mathbf{x}|\mathbf{z}\big)\Big] + \lambda_{\mathrm{KL}}\mathcal{D}_{\mathrm{KL}}\big(\tilde{q}_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\big)$$

λ_{KL} weights the trade-off between reconstruction quality and KL regularization over the latent representation $\mathbf{z}$.

Training Framework

Optimization & Regularization

- About 270k parameters to optimize
- *AMS-Grad* optimizer[1]
- Xavier weight initialization[2]
- Denoising autoencoding criterion[3]
- Sparse regularization in the encoder Bi-LSTM[4]
- KL cost annealing[5]
- Gradient clipping[6]

Training executed on a single GPU (NVIDIA GTX 1080 TI)

---

[1] Reddi, Kale & Kumar, **On the Convergence of Adam and Beyond**, ICLR'18

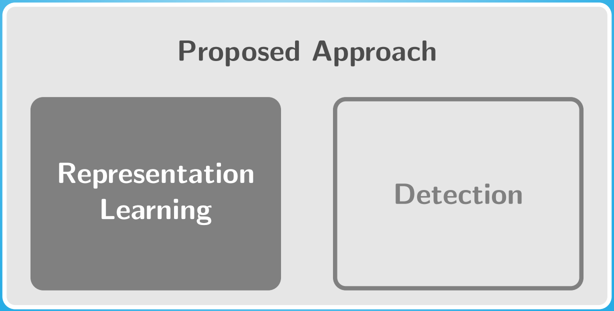
[2] Bengio *et al.*, **Understanding the Difficulty of Training Deep Feedforward Neural Networks**, AISTATS'10

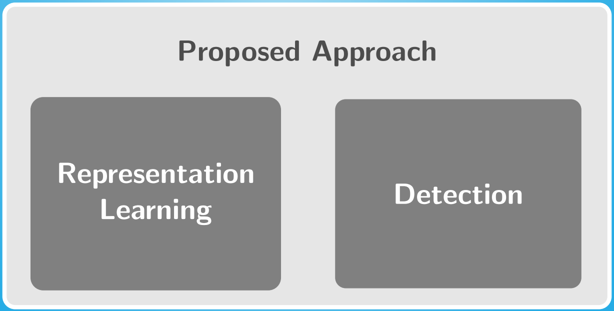
[3] Bengio *et al.*, **Denoising Criterion for Variational Auto-Encoding Framework**, AAAI'17

[4] Arpit *et al.*, **Why Regularized Auto-Encoders Learn Sparse Representation?**, ICML'16

[5] Bowman, Vinyals *et al.*, **Generating Sentences from a Continuous Space**, SIGNLL'16

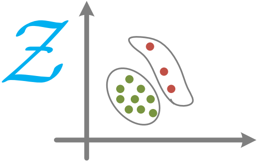
[6] Bengio *et al.*, **On the Difficulty of Training Recurrent Neural Networks**, ICML'13


Proposed Approach
Representation Learning
Detection

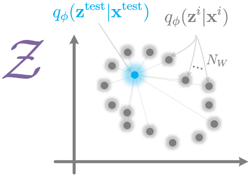

Proposed Approach
Representation Learning
Detection

Based on the representations in the z-space.

- **Clustering**



- **Wasserstein Metric (W)**

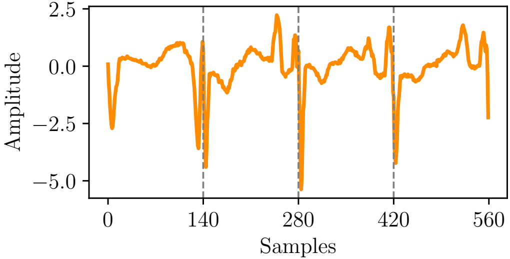


$$\text{score}(\mathbf{z}^{\text{test}}) = \text{median}\{W(\mathbf{z}^{\text{test}}, \mathbf{z}^i)^2\}_{i=1}^{N_W}$$

Experiments & Results

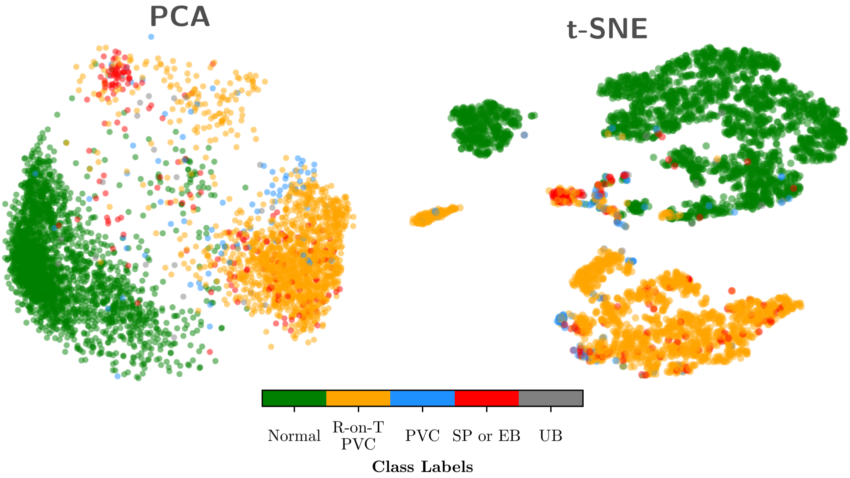
Time Series Data

Electrocardiogram (ECG)



- Dataset ECG5000: available in the UCR Time Series Classification Archive [Keogh *et al.*, 2015];
- One heartbeat ≈ 140 samples;
- 5000 sequences;
- Labelled, 5 classes annotated.

Each datapoint $\rightarrow$ a sequence of length T



Scores using clustering, Wasserstein distance and a support vector machine.
All trained on the representations provided by the model.

Metric	Hierarchical	Spectral	k-Means++	Wasserstein	SVM
AUC	0.9569	0.9591	0.9591	**0.9819**	0.9836
Accuracy	0.9554	0.9581	**0.9596**	0.9510	0.9843
Precision	**0.9585**	0.9470	0.9544	0.9469	0.9847
Recall	0.9463	0.9516	**0.9538**	0.9465	0.9843
F_1-score	0.9465	0.9474	**0.9522**	0.9461	0.9844

Scores using clustering, Wasserstein distance and a support vector machine.
All trained on the representations provided by the model.

	Unsupervised				Supervised
Metric	**Hierarchical**	**Spectral**	**k-Means++**	**Wasserstein**	**SVM**
AUC	0.9569	0.9591	0.9591	**0.9819**	0.9836
Accuracy	0.9554	0.9581	**0.9596**	0.9510	0.9843
Precision	**0.9585**	0.9470	0.9544	0.9469	0.9847
Recall	0.9463	0.9516	**0.9538**	0.9465	0.9843
F_1-score	0.9465	0.9474	**0.9522**	0.9461	0.9844

Comparison with other works:

Source	S/U	Model	AUC	Acc	F_1
Proposed	S	VRAE+SVM	**0.9836**	**0.9843**	**0.9844**
	U	VRAE+Clust/W	**0.9819**	**0.9596**	**0.9522**
Lei *et al.*, 2017	S	SPIRAL-XGB	0.9100	-	-
Karim *et al.*, 2017	S	F-t ALSTM-FCN	-	0.9496	-
Malhotra *et al.*, 2017	S	SAE-C	-	0.9340	-
Liu *et al.*, 2018	U	oFCMdd	-	-	0.8084

- score not reported in the mentioned paper
S/U ≡ **S**upervised/**U**nsupervised

Conclusions & Future Work

- Effective on detecting anomalies in time series data;
- Unsupervised;
- Can be applied on data containing also some anomalous data;
- Suitable for both **univariate and multivariate** data;
- General - works with other kinds of sequential data (e.g., **text**, **videos**);

Acknowledgements

FCT
20 anos
Fundação para a Ciência e a Tecnologia

Institute for Systems and Robotics
Lisboa

SIPG
SIGNAL AND IMAGE PROCESSING GROUP

TÉCNICO LISBOA

Acknowledgements

Thank you for your attention!

mail@joao-pereira.pt
msilveira@isr.tecnico.ulisboa.pt