"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518), Vol.6, 103-108 | Science Signpost Publishing

Unsupervised Anomaly Detection in Time Series with Convolutional-VAE

Emanuele La Malfa[1,a)] and Gabriele La Malfa[2,b)]

[1] *University of Oxford, Oxford, UK*
[2] *EMLYON Business School, Paris, FRANCE*

[a)] Corresponding author: emanuele.lamalfa@cs.ox.ac.uk
[b)] Corresponding author: gabriele.lamalfa@edu.emlyon.com

Abstract. We propose an unsupervised machine learning algorithm for anomaly detection that exploits self-learnt features of mono-dimensional time series. A Variational Autoencoder, where convolution takes place of dot product, is trained to compress each input to a low-dimensional point from a normal distribution, detecting an anomaly as low probability and high density sequence. We validate our work on different public datasets, obtaining results that shed new light on Variational Autoencoders applied to anomaly detection.

I. INTRODUCTION

Forecasting and anomaly detection represent two critical research topics in the analysis of continuous phenomena over time. These studies are applied to different disciplines: physics, healthcare, robotics, artificial intelligence, finance, product analysis, etc. Each discipline employs its own corpus of knowledge [1, 2], build on the top of many factors: from the nature of input variables and their relationships, to the presence of trends, seasonality in the time series etc. [3, 4]. Often physical processes exhibit patterns that can be modeled with simple functions that repeat themselves around their fundamental period, while for many others, like financial and economical series, random walks models [5] are employed and their predictability is still an open issue [6]. A fundamental factor, in terms of predictability of the process, is how information propagates through time. If from one side in Markovian processes the information useful to forecast the next time period depends only on the previous state of the system, on the other in chaotic system a small perturbation of the initial state may influence the behavior or the phenomenon in a remote future.

Sequential models applied to time series have been widely used in recent years in many disciplines [8, 9, 13]. Particular attention has been devoted to explore learning methods [14, 15], enabled by the capacity of those models to process unidimensional and multidimensional datasets, extract features autonomously [16] and model complex systems' dynamics. [17]. Malhotra et al. [20] use Long Short-Term Memory Networks (LSTM) [21] on physical time series: once forecasting is reliable, anomaly detection is based on modeling the prediction errors. On the other hands, Tsang et al. [22] provide a learning method applied to financial time series: after preprocessing data with a Symlet Wavelet Thresholding, a Stacked Autoencoder (SAE) is used for a pre-train session and finally an LSTM is used to forecast and identify anomalies. When it comes to financial time series forecasting, is often necessary a multivariate dataset which provides the missing information about the stochastic process that is not present in the mono variate market index. Laptev et al. [18] feed the features extracted from an autoencoder to a LSTM model, hence the model is used for anomaly detection. Being able to extract only the relevant features for the process may also benefit extreme event forecasting [19].

In this work, we use a Variational Autoencoder (VAE) [24], where dot product is replaced by convolution: this operation has been used extensively [8, 9, 10] for signal processing, hence it can enhance VAE so the model learns relevant features by compressing each input sequence to a point drawn from a low-dimensional gaussian distribution, hence labeling as anomalous dense timesteps (i.e. they are close each other) whose probability is low. To the best of our knowledge Convolutional VAE has been applied only recently to clustering problems [11, 12] while anomaly detection applied to physical time series is an original contribute of this work.

The paper is organized as follows: in Section II the Convolutional Variational Autoencoder (CVAE) model is described in details. In Section III the we present the results on several physical datasets, while the last Section is dedicated to the conclusion and future directions of this work.

II. METHOD

We approach the problem of anomaly detection in mono dimensional time series with a Variational Autoencoder [24] where the dot product, that involves the affine transformation between each stage input and the neural network's parameters, is replaced by convolution. The main idea behind this choice is that convolution is the state of the art method in many challenges where signals are involved [7], mainly due to ability of convolutional networks to build on top of the self-extracted features increasingly complex representations of the input that are used for tasks like classification, outliers detection etc. Differently from simple dot product, convolution is characterized by weights that are shared along the input, making it possible to spot patterns that are invariant to translations and rotations, i.e. robust to noise and perturbations.

As the Variational Autoencoder learns to map each training input to a point belonging to a low-dimensional gaussian distribution, the model emphasizes the local characteristics of each sequence. Figure 1. shows separately the key points of both Variational Autoencoders and the convolution operation. In the next section the math behind the model and the CVAE architecture are described in details, while the full source code for the models employed in this work is provided[1].

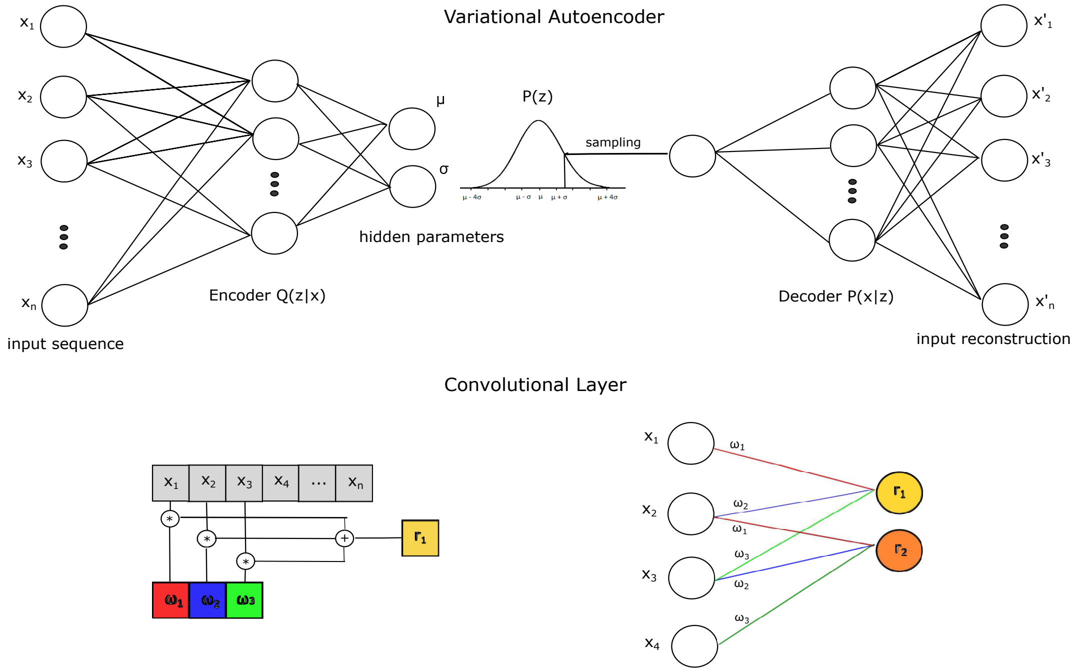


FIGURE 1: The top image shows a VAE architecture where the neural network's parameters are not shared (i.e. each layer is dense). The encoder learns a compressed representation of each input sequence by learning the hidden distribution's parameters (in our case, being it a standard distribution, a vector for the mean and a matrix for the standard deviation). From that representation, the decoder reconstructs the input through another step of affine transformations. The bottom image shows how convolution is employed in our work to replace the neural network's dense layers: in the convolutional layers the parameters are locally connected, hence the model is expected to learn a representation of the series which is invariant to translations and rotations.

[1] https://github.com/EmanueleLM/CVAE

Convolutional Variational Autoencoder

A Variational Autoencoder is used to make inference and learning with a probabilistic based method characterized by latent variables with a posterior intractable distribution.

By defining latent variables z that describe our data it is possible to obtain a generative model: formally, one can model the probability of the input data as $P(X) = \int P(X|z)P(z)dz$, where $P(z)$ is the probability distribution function of the latent variable z, also called *prior*, and $P(X|z)$ the conditional distribution of data.

In order to obtain $P(z)$, one can use the conditional probability $P(z|X)$: unfortunately approximating this distribution is often hard, hence variational inference approximates $P(z|X)$ with another tractable distribution $Q(z|X)$. This approximation problem can be optimized by a convolutional neural network where the first half layers, i.e. the *encoder*, map X to the low-dimensional gaussian distribution $P(z)$ employing $Q(z|X)$, while the second part of the network rebuilds (hence the name *decoder*) the input by approximating $P(X|z)$ from its low-dimensional representation z.

In order to rebuild the original input sequence, the *deconvolution* operation is employed [26], while the hidden representation of each input sequence is obtained with the so-called *reparametrization trick* [27].

The network's parameters are learnt through backpropagation, by minimizing the Kullback-Leibler (KL) divergence between encoder and the intractable prior distributions, namely $D_{KL}[Q(z)\|P(z|X)]$. The objective function, known as Evidence Lower Bound (ELBO), takes the following form: $logP(x) - D_{KL}[Q(z|X)\|P(z|X)] = E[logP(X|z)] - D_{KL}[Q(z|X)\|P(z)]$. In the last equation $E[logP(X|z)]$ measures the reconstruction error from the input sequence, while $D_{KL}[Q(z|X)\|P(z)]$ accounts for the divergence between the encoder and the prior function.

In our work we train the CVAE on data D_{train} that does not contain anomalies and test it on unseen data D_{test} that instead may contain anomalies: in this way, when the compressed representation of an input sequence z from D_{test} is very different from the pool of patterns seen so far in the time series, it will be assigned a low probability and marked as anomalous.

Since minimizing the CVAE loss is computationally expensive even in the condition of defining a trainable encoder function, we explored several normalizations techniques: from l2 regularization, that is known to benefit CVAE [11], to l1 (that induces sparsity in the solution), to a combination of both l1 and l2 regularizations. Moreover, we have experienced that assigning importance weights to the different loss' terms (reconstruction error and KL divergence between decoder and prior) benefits the anomaly detection.

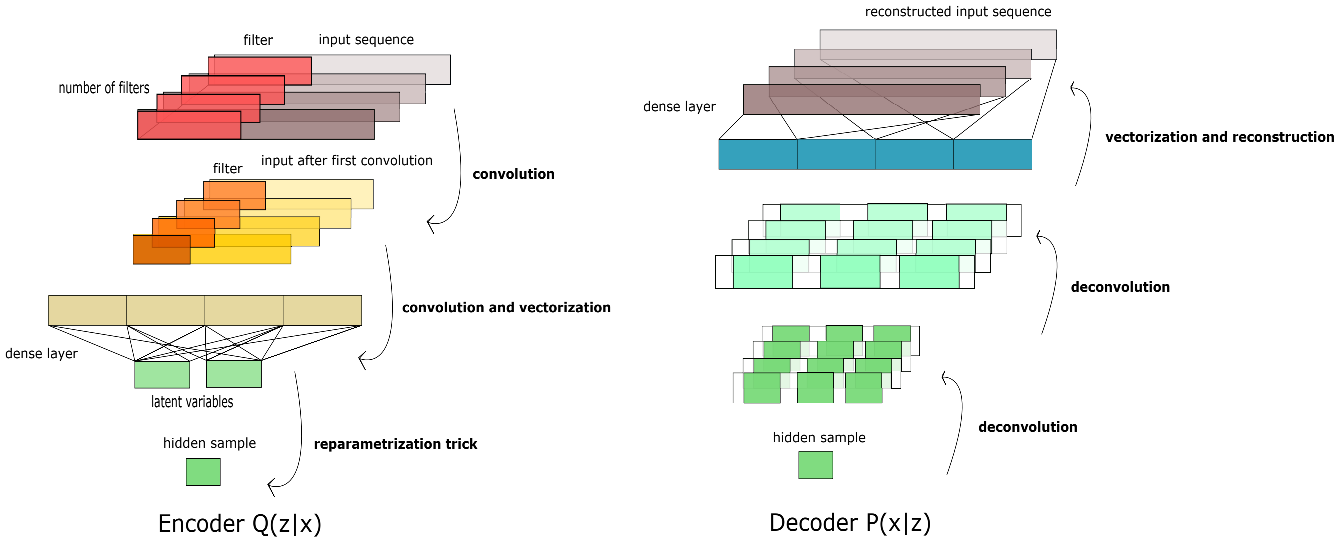


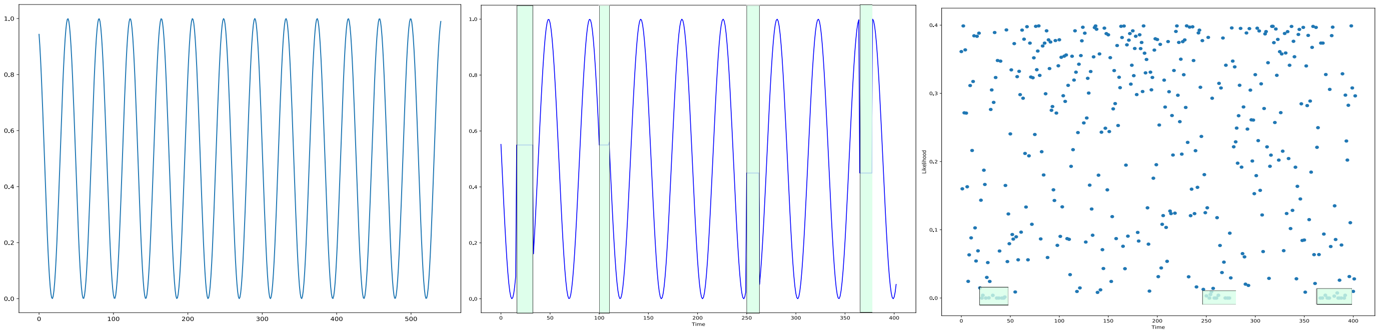
FIGURE 2: The CVAE model proposed in this work: at each stage of the encoder (left part of the image), convolution replaces dot product (fully connected layers), except for the layer before the *bottleneck*, i.e. where the *reparametrization trick* is used to sample from the latent distribution. At each stage of the decoder (right part of the image), deconvolution is used to reconstruct the input sequence, except for the last layer that is dense. Fully connected layers are necessary to make the input match the dimension of respectively its reconstruction and the latent variables distribution.

III. Anomaly Detection

We introduce a method to detect anomalies based on both probability and density of the candidate sequences: given a series of input sequences $s = (x_1, .., x_n)$, s is anomalous if the following two conditions on s hold. First, the probability that CVAE assigns to each sequence is below a threshold τ, namely $\forall\ x_i \ \in \ s\ P_M(x_i) \leq \tau$, for example the 5^{th} and 95^{th} percentiles, being z drawn from a normal distribution. The density based condition requires that for each sequence in s, the temporal distance between each couple (x_i, x_{i+1}) is a fraction of the entire sequence length, namely $\forall\ x_i, x_{i+1} \ \in \ s\ d(x_i, x_{i+1}) \leq k|x|,\ \ s.t.\ k \in (0, 1)$, where $d(\cdot, \cdot)$ is a measure of the distance on the x-axis (the *time* axis) between two candidate sequences, while $|x|$ is the length of each input sequence.

We test our model on three publicly available physical[2] datasets, plus a synthetic one. All the datasets are arranged so the anomalies are not present in train/validation, while in the test part one or more anomalies need to be detected. As regards the synthetic dataset, it is a *sin* function (the model for each data point is $y = sin(t + \rho)$) where some flat zones are introduced in the test set, as it was thought to explicitly show how CVAE algorithm discovers and highlights anomalies (see Figure 3 for respectively the train set (a), the test set (b) and each sequence's likelihood (c)). Even if all the 4 datasets come in the form of univariate quasi predictable time-series, they set different challenges: data from Space Shuttle Marotta Valve (Figure 4) contain a localized anomaly which can be easily spotted when a sequence is long enough to capture the unseen pattern, while in the Power Consumption dataset (Figure 5) the anomaly can be spotted if the algorithm *captures* the 7 days periodicity, finally spotting in the test set that two out of five consumption's peaks are not present at the end of the sequence. Data are preprocessed with normalization methods and subsampled (up to a factor of 5) when possible, to speedup computation.

The results obtained with CVAE algorithm are reported in following images [3].



(a)

(b)

(c)

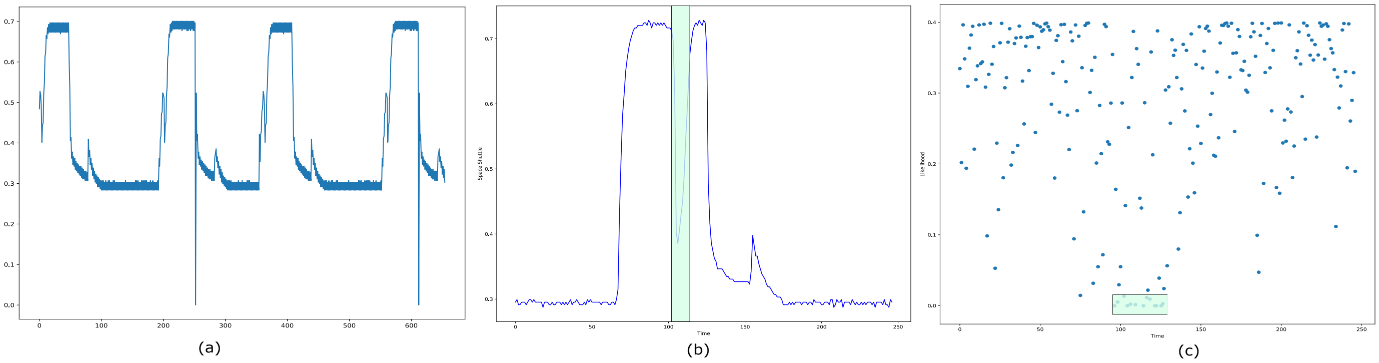
FIGURE 3: Anomaly detection on the synthetic dataset: it is a simple sin function where in the test set (figure (b)) some flat lines has been substituted to the origin function (the green vertical bars highlight each anomalous zone). The CVAE algorithm is able to spot the anomalies (green rectangles on figure (c), where on y-axis it is reported the probability of each timestep) assigning low probability to each sequence that has been artificially modified.

IV. CONCLUSION

We have presented an unsupervised method for anomaly detection that exploits two different concepts: Variational Autoencoders and Convolutional Neural Networks. We have shown that anomalies are highlighted as high-density/low-probability points. We reserve to extend the analysis to other datasets: in future works the aim is to apply those methods also to multidimensional financial time series, to capture the highly complex relations between features and hidden variables.

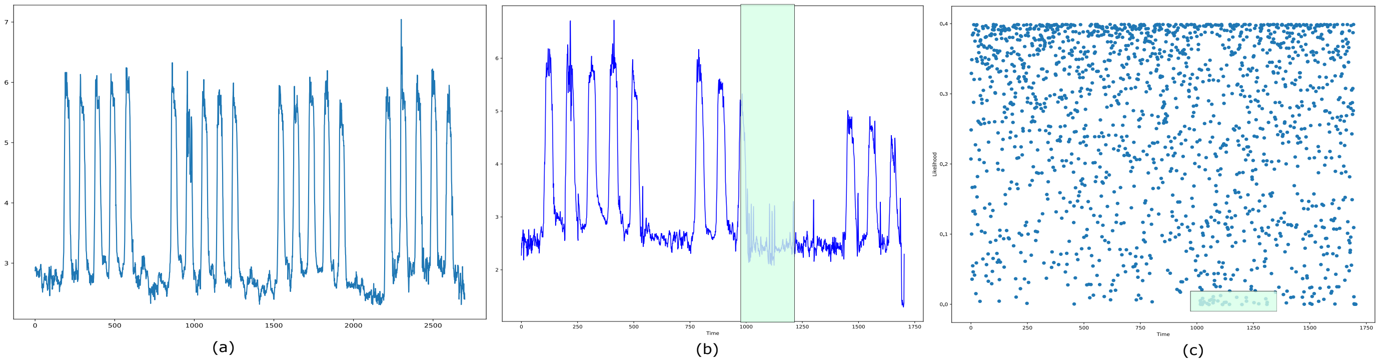
[2] http://www.cs.ucr.edu/~eamonn/discords.

[3] Since the algorithm is fully unsupervised, one may obtain the parameters τ and k as the parameters that enhance anomaly detection on one or more synthetic datasets. On the other hands, one may wish to find τ, k so they maximize the F_β score between anomalous and non-anomalous sequences on validation, but this would make the problem partially supervised.



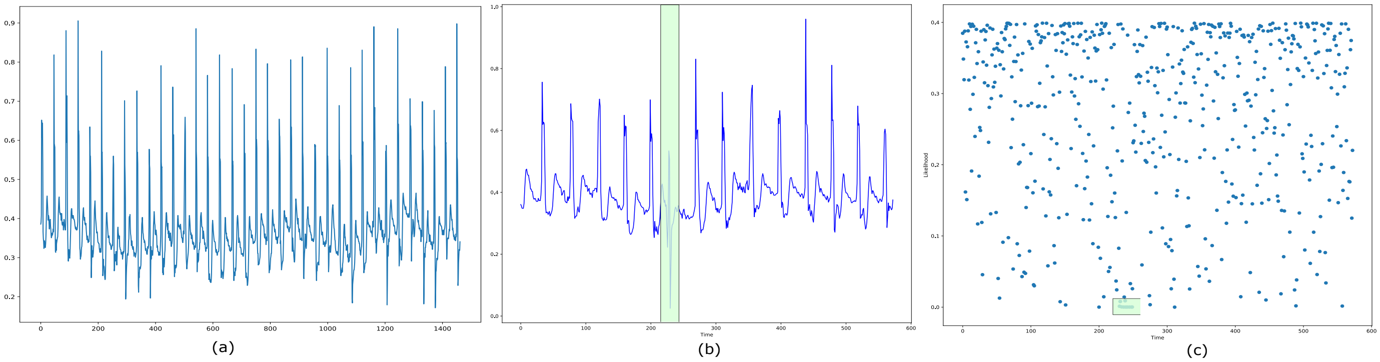
(a) (b) (c)

FIGURE 4: Anomaly detection on the space shuttle dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b)) there's a phase of decompression and compression that constitutes an anomaly that is detected by the CVAE algorithm with certainty (green rectangles on figure (c), where on y-axis it is indicated the probability of each timestep).



(a) (b) (c)

FIGURE 5: Anomaly detection on the power consumption dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b), where the green vertical bar highlight the anomalous zone) the anomalous sequence is identified by two out of five missing peaks of consumption. The CVAE algorithm spots it with a high density zone of low probability sequences (green rectangles on figure (c), where on y-axis it is indicated the probability of each timestep).



(a) (b) (c)

FIGURE 6: Anomaly detection on the ecg dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b), where the green vertical bars highlight the anomalous zone) there's an anomalous sequence between two non-anomalous peaks. The CVAE algorithm spots it with a high density zone of low probability sequences (green rectangles on figure (c), where on y-axis it is indicated the probability of each timestep).

REFERENCES

[1] V. Chandola, A. Banerjee, V. Kumar, *Anomaly detection: A survey*, ACM Computing Surveys (CSUR), Vol. 41, Issue 3, 2009.

[2] V. J. Hodge, J. Austin, *A Survey of Outlier Detection Methodologies*, Artificial Intelligence Review, Vol. 22, Issue 2, pp. 85126, 2004.

[3] R. H. Shumway, D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, Springer International Publishing, 2017.

[4] R. S. Tsay, *Analysis of Financial Time Series*, John Wiley & Sons, 3rd Edition, Hoboken, 2010.

[5] E. F. Fama, *Random Walks in Stock Market Prices*, Financial Analysts Journal, Vol. 21, no. 5, pp. 55-59, 1965.

[6] A. W. Lo, A. C. MacKinlay, *A Non-Random Walk Down Wall Street*, Princeton University Press, 2002.

[7] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, *A survey of deep neural network architectures and their applications*, Neurocomputing, 234, 11-26, 2017.

[8] L. Deng, *Three classes of deep learning architectures and their applications: a tutorial survey*, APSIPA Transactions on Signal and Information Processing, 2012.

[9] Y. LeCun, Y. Bengio, G. Hinton, *Deep learning*, Nature, Vol. 521, no. 7553, 2015.

[10] Y. LeCun, Y. Bengio, *Convolutional networks for images, speech, and time series*, The handbook of brain theory and neural networks 3361 (10), 1995.

[11] X. Guo, X. Liu, E. Zhu, J. Yin, *Deep clustering with convolutional autoencoders*, International Conference on Neural Information Processing Springer, Cham, 2017.

[12] Aytekin, C., Ni, X., Cricri, F., Aksu, E. *Clustering and Unsupervised Anomaly Detection with l 2 Normalized Deep Auto-Encoder Representations*, International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE, 2018

[13] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer-Verlag Berlin Heidelberg, Vol. 385, 2012.

[14] K. Yamanishi, J. Takeuchi, *A unifying framework for detecting outliers and change points from time series*, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, Issue 4, pp. 482-492, 2006.

[15] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu , *Recurrent Neural Networks for Multivariate Time Series with Missing Values*, Scientific Reports 8, Article Number 6085, 2018.

[16] M. Assaad, R. Bon, H. Cardot, *A New Boosting Algorithm for Improved Time-Series Forecasting with Recurrent Neural Networks*, Information Fusion 9, pp. 41-55, 2008.

[17] O. Ogunmolu, X. Gu, S. Jiang, N. Gans, *Nonlinear systems identification using deep dynamic neural networks*, arXiv: 1610.01439, 2016.

[18] N. Laptev, J. Yosinski, L. E. Li, S. Smyl, *Time-series Extreme Event Forecasting with Neural Networks at Uber*, Uber AI, ICML, 2017.

[19] L. de Haan, A. Ferreira, *Extreme Value Theory: An Introduction*, Springer Series in Operations Research and Financial Engineering, 2006.

[20] P. Malhotra, L. Vig, G. Shroff, A. Puneet, *Long short term memory networks for anomaly detection in time series*, ESANN 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[21] S. Hochreiter, J. Schmidhuber, *Long short-term memory*, Neural Computation, Vol. 9, Issue 8, pp. 1735-1780, 1997.

[22] G. Tsang, J. Deng, X. Xie, *Recurrent Neural Networks for Financial Time-Series Modelling*, 24th International Conference on Pattern Recognition, ICPR, 2018.

[23] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, G. Shroff, A. Puneet, *LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection*, ArXiv: 1607.00148, 2016.

[24] D. P. Kingma, M. Welling, *Auto-Encoding Variational Bayes*, arXiv: 1312.6114, 2014.

[25] J. Armstrong, F. Collopy, *Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons*, International Journal of Forecasting, Vol. 8, Issue 1, pp. 69-80, 1992.

[26] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, *Deconvolutional networks*, Cvpr (Vol. 10, p. 7), 2010.

[27] D. J. Rezende, S. Mohamed, *Variational inference with normalizing flows*, arXiv preprint arXiv:1505.05770, 2015.

[28] S. Makridakis, *Accuracy Measures: Theoretical and Practical Concerns*, International Journal of Forecasting, Vol. 9, Issue 4, pp. 527-529, 1993.

[29] R. J. Hyndman, A. B. Koehler, *Another Look at Measures of Forecast Accuracy*, International Journal of Forecasting, Vol. 22, Isuue 4, pp. 679-688, 2006.