Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform

Kanarachos, S; Christopoulos, S-R.G; Chroneos, A. and Fitzgerald, M.E.

Original citation:
Kanarachos, S; Christopoulos, S-R.G; Chroneos, A. and Fitzgerald, M.E. (2017) Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform. *Expert Systems with Applications* Vol: 85 November 2017, pp.292-304. DOI: 10.1016/j.eswa.2017.04.028

http://dx.doi.org/10.1016/j.eswa.2017.04.028

Elsevier

Detecting anomalies in time series via a deep learning algorithm combining Wavelets, Neural Networks and Hilbert transform

Stratis Kanarachos[a], Stavros-Richard G. Christopoulos [a,b], Alexander Chroneos[a],and Michael E. Fitzpatrick[a]

[a] *stratis.kanarachos@coventry.ac.uk, ac0966@coventry.ac.uk, ab8104@coventry.ac.uk, ab6856@coventry.ac.uk, Faculty of Engineering, Environment and Computing, Coventry University, Priory Street, Coventry CV1 5FB, United Kingdom*

[b]*Solid Earth Physics Institute, Faculty of Physics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos 157 84, Athens, Greece*

Corresponding author: *Stratis Kanarachos, stratis.kanarachos@coventry.ac.uk Tel: +44(0)2477657720, Engineering & Computing Building - EC 4-07, Faculty of Engineering, Environment and Computing, Coventry University, 3, Gulson Road, Coventry, CV1 2JH*

Abstract

The quest for more efficient real-time detection of anomalies in time series data is critically important in numerous applications and systems ranging from intelligent transportation, structural health monitoring, heart disease, and earthquake prediction. Although the range of application is wide, anomaly detection algorithms are usually domain specific and build on experts' knowledge. Here a new signal processing algorithm –inspired by the deep learning paradigm – is presented that combines wavelets, neural networks, and Hilbert transform performs robustly and is transferable. The proposed neural network structure facilitates learning short and long-term pattern

interdependencies; a task usually hard to accomplish using standard neural network training algorithms. The paper provides guidelines for selecting the neural network's buffer size, training algorithm, and anomaly detection features. The algorithm learns online the system's normal behavior and does not require the existence of anomalous data, for assessing its statistical significance. This is essential for applications that require customization. The anomalies are detected by analyzing hierarchically the instantaneous frequency and amplitude of the residual signal. Its applicability is demonstrated through detection of anomalies in the Seismic Electric Signal activity, that is potentially important for earthquake prediction; and automated detection of road anomalies (e.g. potholes, bumps, etc.) using smartphone sensors. The evaluation of the anomaly detection algorithm is based on the statistical significance of the Receiver Operating Characteristic curve. Finally, we propose strategies for decision-making that may increase the efficiency of the application of the algorithm, and expedite evaluation of real-time data.

Keywords: anomaly detection; deep learning; receiver operating characteristics

Abbreviations

A: amplitude
ADF: Anomaly Detection Filter
AP: Anomalous Pulses
AUC: Area Under Curve
DNN: Deep neural network
DR: Dichotomous representation
$FN:$ False Negative

$FP:$ False Positive

FPr: False Positive rate

IMS-SEPI: Ioannina Measuring Station of the Solid Earth Physics Institute

L'$_s$-I, L', L: Length of dipoles

M: Dichotomous representation index

M_2: Anomalous pulses index

MEMS: Micro Electro-Mechanical Systems

NDEEF: Normalized Deflection of the Earth's Electric Field

N_q: polynomial degree

NN: Neural networks

ROC: Receiver Operating Characteristics

$S_{m,n}$: approximation coefficients

SES: Seismic Electric Signal

$T_{m,n}$: detail coefficients

$TN:$ True Negative

$TP:$ True Positive

TPr: True Positive rate

V2I: vehicle-to-infrastructure

V2V: vehicle-to-vehicle

$\boldsymbol{W}_m$**:** neural network interconnection matrices for the output layer

$\boldsymbol{V}_m$: neural network interconnection matrices for the hidden layer

WANEH: WAvelets, NEural networks and Hilbert transform

ΔV: Voltage difference

a_0: dilation parameter

b_0: location parameter

d_m: signal detail at scale m

d_{dm}: filter signal detail at scale m

$e:$ error

e_H: Hilbert transfrom of error e

k: estimator value

$k_i k$: threshold value

m: meter

m: parameter controlling the wavelet dilation

$m0$: arbitrary scale

n: parameter controlling the wavelet translation

n_h: number of hidden neurons

p: probability

q: scaling function shift

s: second

x: signal in time domain

x_d: filtered signal

x_m: approximation signal at scale m

y: neural network output

y_m: neural network uoutput at scale m

$\boldsymbol{\beta}_m$: is neural network bias vector

θ: instantaenous phase

λ: noise threshold

τ_m: scale dependent phase lag

φ: scaling function

$\varphi_{m,n}$: wavelet (father) basis

ψ: wavelet

$\psi_{m,n}$: wavelet (mother) basis

1. Introduction

Over the past few decades the advent of advanced computational methods in conjunction with the ever-increasing computational resources has led to advances in our understanding of the physical world (Choudhury et al., 2015; Fan, Osetskiy, Yip, & Yildiz, 2013; Rushton & Chroneos, 2014; Sun, Jayaraman, Chen, Persson, & Ceder, 2015). Detecting anomalies in the behaviour of systems and processes is significant for predicting their behaviour. This is critically important in systems as diverse as materials performance in hazardous environments such as nuclear reactors, autonomous vehicle suspension systems, disaster prevention due to earthquakes, social networks ("Introducing practical and robust anomaly detection in a time series," n.d.) as well as heart attack prevention (Ghahramani, 2015; Ikonomopoulos, Alamaniotis, Chatzidakis, & Tsoukalas, 2013; Nicholas V. Sarlis et al., 2015; P. A. Varotsos, Sarlis, Skordas, & Lazaridou, 2007). The above complex systems share a characteristic; lack of models that can accurately describe the system's behaviour. Therefore, for performing the anomaly detection task, it is required to rely on data-based approaches.

The body of literature work indicates the substantial interest of the research community in developing time series anomaly detection algorithms (Akhoondzadeh, 2015; Akouemo & Povinelli, 2015; Chen & Zhan, 2008; Georgoulas, Loutas, Stylios, & Kostopoulos, 2013; Harrou, Kadri, Chaabane, Tahon, & Sun, 2015; Li, Liu, & Zhang, 2015). In Pimentel *et al*., a comprehensive review of existing anomaly detection algorithms is provided. The authors pointed out that there is no universally accepted definition for anomaly and that anomalies can be distinguished into outliers, data points that are dissimilar to the remaining points in the data set, and anomaly patterns, a small

fraction of data which are different from the majority of "normal" data in the feature space. Furthermore, anomaly detection algorithms were classified into five major categories: Probabilistic, Distance-based, Reconstruction-based, Domain-based and Information-theoretic based.

In Akhoondzadeh an anomaly detection method for predicting earthquakes is presented. The proposed method, which is a combination of a reconstruction and statistical approach, detects ionospheric anomalies that occur a few days before and after earthquakes, a relatively slow phenomenon. Akhoondzadeh implements the Artificial Bee Colony Algorithm for training a predictor that estimates the future value of electron concentration. The prediction is made using three past samples. In case the cumulative prediction error falls outside a predefined range $\mu \pm k \cdot \sigma$ (where μ and σ are the mean and the standard deviation) the behaviour is considered as abnormal. According to the authors, the method performed satisfactorily in a number of earthquakes occasions.

A probabilistic approach for detecting anomalies in natural gas consumption data was proposed by Akouemo and Povinelli. First, a linear regression model is developed, mapping the natural gas consumption to the outside temperature, cooling and heating reference temperatures, cooling, and heating degree days and the natural gas consumption of the previous day. The regression model is location dependent as it depends on empirical coefficients such as degree days and threshold temperatures. Furthermore, the natural gas consumption appears to have a linear dependency from time as only the previous' day natural gas consumption is considered. After the coefficients of the linear regression are determined, they are used to compute the residuals of the data, by taking the difference between the actual and estimated values. The maximum and minimum

values of the set of residuals are used to discover anomalies. A statistical test is conducted, with the null hypothesis being that the extremum is not an outlier. If the probability for the extremum to belong to the same distribution as the remaining points in the residual data set is less than the probability of committing a type I error, then it is considered as abnormal. Type I occurs when the null hypothesis is rejected when true. The significance level was typically 1%.

Chen and Zhan proposed a distance-based anomaly detection algorithm for discovering infrequent patterns in time series. First, Haar wavelet decomposition is applied to reduce the noise level in the signal. Then the signal is analysed in multiple scales. The signal in each time scale is segmented and compared to previous time patterns of the same time scale. In case no match is found – difference larger than a predefined threshold – the pattern is considered anomalous. The method is based on two assumptions: a) The anomaly pattern is the most infrequent pattern and b) Signal decomposition in different scales and independent analysis is adequate for detecting anomalies. The algorithm was applied with success in standard benchmark case studies.

In Georgoulas *et al*. an early warning bearing fault detection scheme was proposed. The acceleration signal is first analysed in Intrinsic Mode Functions (IMF) using the Hilbert-Huang Transform and then processed by three independent anomaly detectors. Because of the continuous shaft rotation, it is fit to analyse the signal analysis in IMFs or frequencies. The authors employed an empirical rule for determining that the first and third IMF are the most important. To discover an anomaly the “opinions” of the detectors are combined using the majority voting scheme. The anomaly detectors utilised in the study are: a) a Gaussian detector that assumes a normal data distribution, b) a

Nearest Neighbour detector assuming that normal data instances occur in dense neighbourhoods, while anomalies occur far from the nearest neighbour and c) a Principal Components Analysis detector, which reconstructs a normal signal using subspaces capable of describing the normal dataset. Harrou et al. developed a detection scheme for detecting anomalies in emergency department calls. The method is a combination of Principal Component Analysis and the multivariate cumulative sum (MCUSUM) monitoring chart. The signal is reconstructed using PCA and then the residuals, when compared to the original signal, are calculated. The cumulative sum of deviations of each residual previously observed is compared to the nominal value. In case the difference is exceeding a predefined threshold the event is classified abnormal. The authors highlighted the importance of considering the multivariate signals' cross-correlation.

In Li *et al.* two versions of negative selection method are proposed for detecting anomalies on two synthetic datasets. The detector's performance was evaluated on the basis of true positives, true negatives, false positives and false negatives.

From the literature review, it is concluded that different problems were solved using different anomaly detection algorithm. The nature of the problems varied significantly in terms of the underlying dynamics, domain knowledge, embedded measurement noise, complexity of system's normal behaviour, and repeatability of normal or abnormal behaviour. The evaluation methods used are significantly different; some studies just focus on the number of true positives while others provide a comprehensive analysis of true positives, true negatives, false positives and false negatives. In all methods, a threshold was required for distinguishing normal from abnormal behaviour but only in a few cases the threshold determination was in detail

explained and linked to the detection method. Last but least, most anomaly detection algorithms are rarely tested as to whether they are transferable without significant tuning effort. In complex systems where the interactions and disturbances are unknown or too complex to model this is important.

In this paper, a new transferable anomaly detection method is presented. The method is a combination of Neural Networks, discrete wavelet analysis, and Hilbert transform. The threshold for classifying an event abnormal and its statistical significance are determined using the probabilistic Receiver Operating Characteristics (ROC) method. Our main contributions in the field of "Anomaly Detection" are a) the proposition of a unique Deep Neural Network structure for reconstructing the normal behaviour of a system and b) the feature selection of the anomaly detector in dependence of the probabilistic Receiver Operating Characteristics method. The first is important for detecting short and long term interdependencies using NNs; usually this is a hard task using standard training algorithms. The latter is important when modelling the normal behaviour of a complex system, where it is difficult to accurately reconstruct it.

As mentioned, the proposed method reconstructs only the normal behaviour of the system which is very important for applications in which anomalies are rare or not standard. Furthermore, the use of NN facilitates online training, beyond the point of deployment which is significant for applications where normal behaviour needs customization, for example, vehicle type or patient. The proposed method is successfully applied – without any manual tuning effort – to two diverse examples: the detection of anomalies in the Seismic Electric Signal activity, that is potentially important for

earthquake prediction and the automated road anomaly detection (for example potholes, bumps) using smartphones.

The rest of the paper is structured as follows: In Section 2 the proposed algorithm is presented and compared to existing anomaly detection methods. In Section 3 the application of the algorithm in two diverse cases is described in detail. In Section 4 the results are analyzed and discussed, while in Section 5 conclusions and future work are drawn.

2. Transferable anomaly detection in time series data: The WANEH algorithm

The anomaly detection method proposed in this paper is called WANEH. It combines WAvelets, NEural networks and Hilbert transform (WANEH) and its schematic is illustrated in **Fig. 1**.

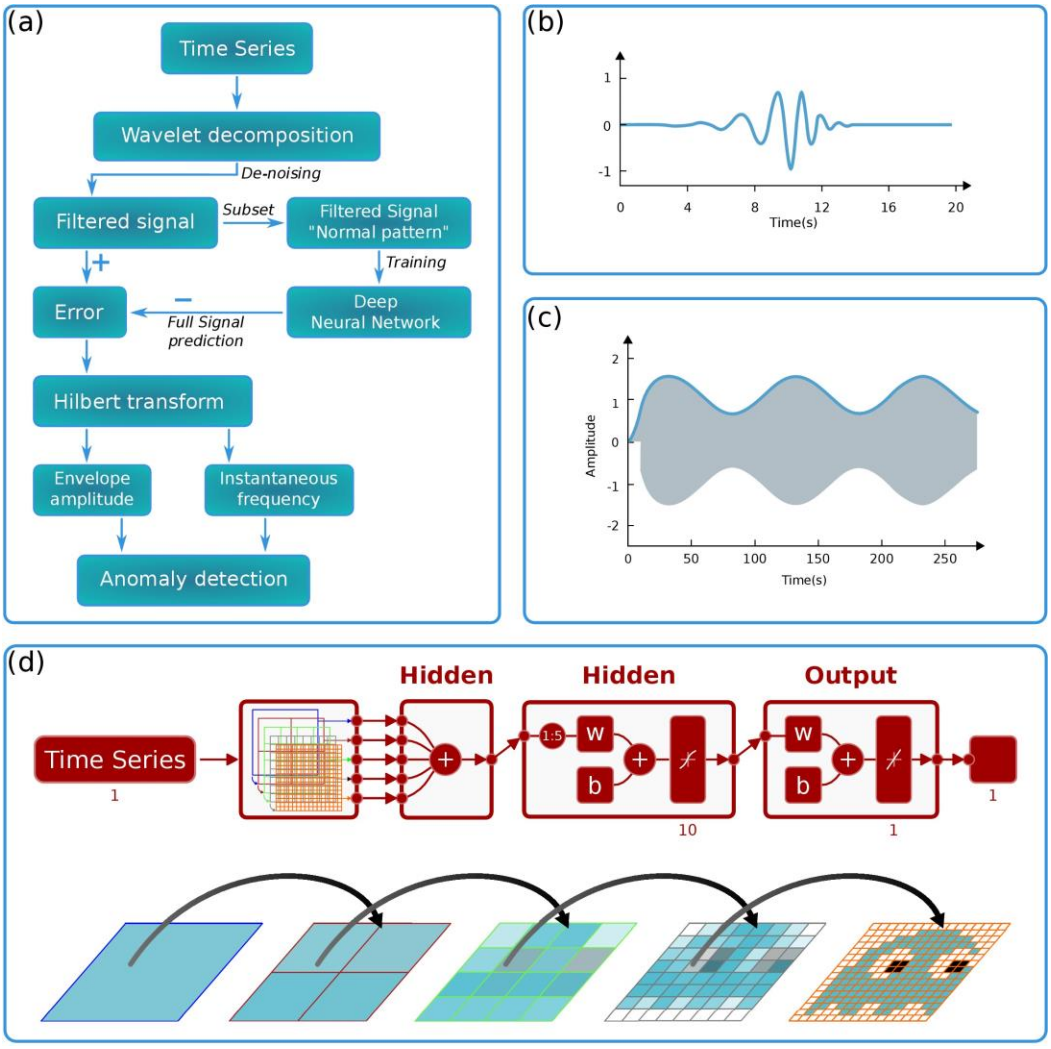

Fig. 1: WAvelets, NEural networks and Hilbert transform (WANEH) algorithm: a) Flow chart of proposed algorithm b) Daubechies 9 wavelet basis for de-noising the raw signal c) Energy temporal evolution d) Deep neural network architecture for learning the patterns in and between different time scales.

2.1 Multiresolution signal reconstruction using wavelet analysis.

Wavelets are important mathematical tools to analyse a time series $x(t)$. There are different methods fto applying waveles, for example, the continuous or discrete wavelet transform.

A wavelet $\psi_{m,n}$ grows and decays within a limited time period and the wavelet transform $T_{m,n}$ can decompose a signal into different scales with different levels of resolution through the dilation of a single prototype function known as the basis wavelet ψ, see **Fig. 1**b.

$$T_{m,n} = \int_{-\infty}^{\infty} x(t) \cdot \psi_{m,n}(t) \cdot dt \quad (1)$$

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \cdot \psi \cdot \left(\frac{t - n \cdot b_0 \cdot a_0^m}{a_0^m} \right) \quad (2)$$

where $T_{m,n}$ are the discrete wavelet transform values given on a scale-location grid of index m, n. The integers m, n control the wavelet dilation and translation respectively and are contained in the set of all integers, both positive and negative. a_0 is a specified fixed dilation step parameter set at a value greater than 1, and b_0 is the location parameter which must be greater than zero.

Common choices for a_0 and b_0 are 2 and 1 respectively. Equation (2) becomes:

$$\psi_{m,n}(t) = 2^{-m/2} \cdot \psi \cdot (2^{-m} \cdot t - n) \quad (3)$$

This power-of-two logarithmic scaling of both the dilation and translation steps is known as the dyadic grid arrangement. In this arrangement, the values $T_{m,n}$ are known as wavelet coefficients or detail coefficients. Some methods exploit detail coefficients for detecting anomalies.

By choosing an orthonormal wavelet basis, $\psi_{m,n}(t)$ it is possible to reconstruct the original signal in terms of the wavelet coefficients $T_{m,n}$ using the inverse discrete wavelet transform as follows:

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n} \cdot \psi_{m,n}(t) \quad (4)$$

Equation (4) is useful for reconstructing the signal $x(t)$ but not for obtaining a multiresolution of it. For this – under the assumption of dyadic grid arrangement and the use of orthonormal bases – the use of a scaling function $\varphi(t)$ is required:

$$\varphi_{m,n}(t) = 2^{-m/2} \cdot \varphi \cdot (2^{-m} \cdot t - n) \quad (5)$$

with the property

$$\int_{-\infty}^{\infty} \varphi_{0,0}(t) \cdot dt = 1 \quad (6)$$

The scaling function is convolved with signal $x(t)$ to produce the approximation coefficients $S_{m,n}$:

$$S_{m,n} = \int_{-\infty}^{\infty} x(t) \cdot \varphi_{m,n}(t) \cdot dt \quad (7)$$

and obtain a continuous approximation of signal $x_m(t)$, at scale m:

$$x_m(t) = \sum_{-\infty}^{\infty} S_{m,n} \cdot \varphi_{m,n}(t) \quad (8)$$

where $x_m(t)$ is a smooth, scaling-function-dependent, version of signal *x(t)*, at scale m. Using Equations (4) & (8) signal $x(t)$ is represented as a combined series expansion:

$$x(t) = \sum_{n=-\infty}^{n=\infty} S_{m0,n} \cdot \varphi_{m0,n}(t) + \sum_{m=-\infty}^{m0} \sum_{n=-\infty}^{\infty} T_{m,n} \cdot \psi_{m,n}(t) \quad (9)$$

Equation (9) expresses the original continuous signal as the combination of an approximation of itself, at arbitrary scale index $m0$, added to a succession of signal details from scales $m0$ down to $-\infty$. If we denote with $d_m(t)$ the signal detail, at scale m:

$$d_m(t) = \sum_{n=-\infty}^{\infty} T_{m,n} \cdot \psi_{m,n}(t) \quad (10)$$

then Equation (9) is rewritten as:

$$x(t) = x_{m0}(t) + \sum_{m=-\infty}^{m0} d_m(t) \quad (11)$$

$$x_{m-1}(t) = x_m(t) + d_m(t) \quad (12)$$

Equation (12) says that if the signal detail $d_m(t)$, at an arbitrary scale (index m), is added to the approximation $x_m(t)$, at that scale, we get the signal approximation $x_{m-1}(t)$ at an increased resolution. Equation (12) allows the multiresolution representation d_m of signal $x(t)$ at different scales m, see Fig. **2**. For further details on the selection of the scaling function, refer to (Addison, 2002). A key advantage of signal analysis via wavelets is that it allows local features of the signal to be studied, with a detail matching their scale.

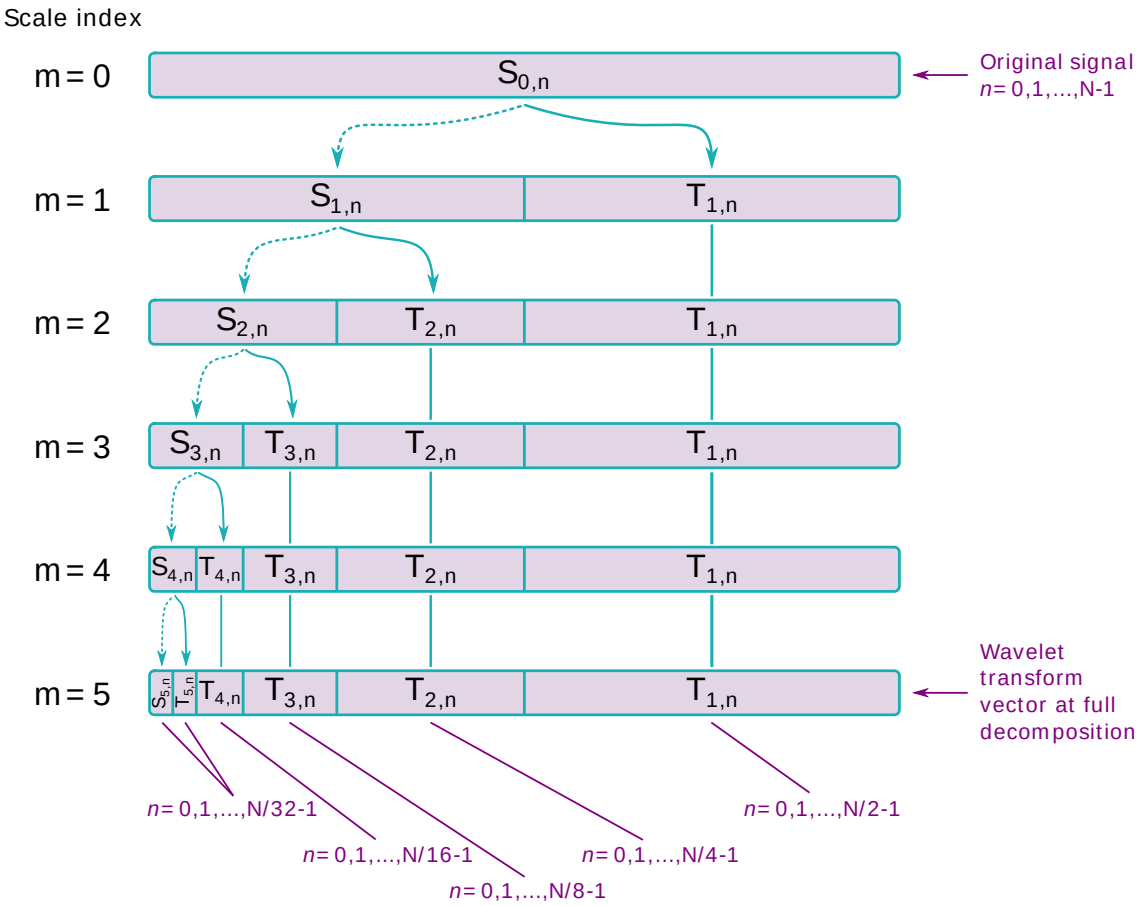

Fig. 2: Multiresolution representation of a signal using wavelet analysis using $m = 5$ scales

To obtain the de-noised $x_d(t)$ representation of signal $x(t)$ a threshold λ is defined and the detail coefficients $T_{m,n}$ are adjusted according to:

$$T_{m,n} = \begin{cases} 0, if\ |T_{m,n}| < \lambda \\ T_{m,n}, if\ |T_{m,n}| \geq \lambda \end{cases} \quad (13)$$

$$x_d(t) = x_{m0}(t) + \sum_{m=-\infty}^{m0} d_{dm}(t) \quad (14)$$

where $d_{dm}(t)$ is the filtered signal detail, at scale m. There are different ways for defining the threshold and in this paper the minimax approach was selected. For more details the readers are referred to (Donoho & Johnstone, 1998).

In this study, the wavelet basis function comes from the Daubechies wavelet family. Daubechies wavelets besides Equations (5) and (6) satisfy also:

$$\sum_{k=0}^{N_q-1} (-1)^q \cdot c_q \cdot q^m = 0 \quad (15)$$

for integers $m = 0, 1, 2, \ldots, N_q/2 - 1$ and q (q represents the scaling function shift along the time axis). Daubechies wavelets can approximate signals which are polynomial up to degree $N_q/2 - 1$. For example, Table 1 lists the polynomial degrees supported by two different Daubechies wavelet types.

Table 1. Polynomial degrees supported by Daubechies wavelets db2 (Haar) and db9

Daubechies wavelet	Polynomial Degree
Db2	0
Db9	3

2.2 Deep Temporal Neural Networks Architecture.

Recent advances have demonstrated the excellent performance of Deep Neural Networks (DNNs) in modelling static data and performing tasks such as image classification and protein structure prediction (Heffernan et al., 2015; Hinton, 2007). The main idea behind deep learning is to first learn the hidden patterns in raw data and then combine this information to perform the prediction or classification. DNNs have not been applied extensively to time series data but this field is currently gaining increasing attention (Längkvist, Karlsson, & Loutfi, 2014).

A novel deep temporal neural network architecture for identifying the temporal structure of the filtered signal $x_d(t)$ is illustrated in **Fig. 1**d. The DNN is structured in three layers. The inputs to the first layer are the filtered signal details $d_{dm}(t)$, for all scales m. The first layer comprises an ensemble of stacked neural networks. For each scale m, a standard feedforward neural network is trained for identifying the temporal structure of the signal detail $d_{dm}(t)$.

$$y_m = \sum_{r=1}^{n_h} w_{mir} \cdot \sigma_m \cdot \left(\sum_{j=1}^{pn} v_{mrj} \cdot x_{mj} + \beta_{mr} \right) \quad (16)$$

or in matrix form:

$$y_m = \boldsymbol{W}_m \cdot \sigma_m \cdot (\boldsymbol{V}_m \cdot \mathbf{x}_m + \boldsymbol{\beta}_m) \quad (17)$$

where $\mathbf{x}_m = [d_{dm}(t), d_{dm}(t - \tau_m), d_{dm}(t - 2 \cdot \tau_m), \dots\ ] \in R^p$ is the input., τ_m is a scale dependent constant , $y_m = d_{dm}(t + 1)$ is the output at time $t + 1$ and the nonlinear

operation $\sigma(\quad)$ is taken elementwise. In this paper, the nonlinear operation $\sigma(\quad)$ is a logistic sigmoid. For more details on the different options available, refer to (Rojas, 1996). The interconnection matrices are $\boldsymbol{W}_m \in R^{l \times n_h}$ for the output layer, $\boldsymbol{V}_m \epsilon R^{n_h \times pn}$ for the hidden layer, $\boldsymbol{\beta}_m \in R^{n_h}$ is the bias vector with n_h the number of hidden neurons.

In the second layer, the m feedforward neural networks are combined to learn the temporal structure of signal $x_d(t)$:

$$y = \mathbf{W} \cdot \sigma \cdot \left(\mathbf{V}_2 \cdot \sigma \cdot \left(\sum_0^m \mathbf{y}_m \right) + \boldsymbol{\beta}_2 \right) \tag{18}$$

where $\mathbf{y}_m = [y_m(t), y_m(t - \tau_m), y_m(t - 2 \cdot \tau_m), \dots] \epsilon R^p$, τ_m is a scale dependent constant, and $y = y(t + 1)$ is the DNN output at time $t + 1$.

2.3 Hilbert transform.

The error signal *e* is defined as the difference of the filtered signal $x_d(t)$ from DNN's output $y(t)$: $e = x_{filt} - x_{NN}$.

$$e = x_d - y \tag{19}$$

The envelope A and instantaneous frequency f_{inst} of the error signal $e(t)$ are the features used for anomaly detection. For this the Hilbert transform is utilized:

$$e_H(t) = lim_{\varepsilon \to 0} \left[\frac{1}{\pi} \cdot \int_{-\infty}^{t-\varepsilon} \frac{e(t)}{x - t} \cdot dt + \frac{1}{\pi} \cdot \int_{t+\varepsilon}^{+\infty} \frac{e(t)}{x - t} \cdot dt \right] \tag{20}$$

where $e_H(t)$ is the Hilbert transform. Hilbert transform is the convolution of $e(t)$ with a reciprocal function $1/x - t$, thus Hilbert transform emphasizes the local properties of $e(t)$. If $\hat{e}(\omega)$ represents the Fourier transform of $e(t)$, then the Hilbert transform is:

$$e_H(t) = \mathcal{F}^{-1}\{-j \cdot sgn(\omega) \cdot \hat{e}(\omega)\} \quad (21)$$

where $\mathcal{F}^{-1}$ represents inverse Fourier transform (Goswami & Hoefel, 2004). The instantaneous phase $\theta(t)$, frequency $\dot{\theta}(t)$, and amplitude $A(t)$ of $e(t)$ are defined:

$$\theta(t) = arctan\left\{\frac{e_H(t)}{e(t)}\right\} \quad (22)$$

$$\dot{\theta}(t) = \frac{d\theta}{dt}$$

$$A(t) = \sqrt{e(t)^2 + e_H(t)^2} \quad (23)$$

A schematic of an envelope detector is given in **Fig. 1**c (thick solid line).

2.4 Probabilistic Receiver Operating Characteristics.

The ROC method is used to compare the results of the proposed algorithm with a dichotomous time series characterizing the specific anomaly. In each example, the latter time series is determined by independent methods. This comparison is used for the evaluation of both the statistical significance and the efficiency of the method.

2.5 Comparison to existing methods

The proposed methods focuses on the early and accurate detection of anomalies and not in their classification. Furthermore, we are only interested in detecting anomalies in patterns, not in the detection of outlier points. Many existing anomaly detection methods require datasets containing pattern anomalies, which are difficult to produce in in real life problems. Some anomalies are rare, for example, earth's seismic electric signals, or too expensive to collect. Besides, unknown anomaly patterns might emerge. Contrary, the proposed method is based on the approximation of the system's normal behaviour and does not require for its development prior anomalous data.

As mentioned in (Pimentel, Clifton, Clifton, & Tarassenko, 2014), anomaly detection methods can be broadly classified in five categories: Probabilistic, Distance-based, Reconstruction-based, Domain-based and Information-theoretic based. Each category has its own strengths and weaknesses. Probabilistic methods require large amounts of data, thus don't perform well when the anomalies training set is small. Distance-based methods, which include nearest-neighbour and clustering approaches, require the definition of an appropriate distance measure for the given data. It is hard to define such a metric, especially in high-dimensional problems. Furthermore, distance-based methods usually require manual selection of parameters, so it is not possible to use them for automatically constructing a model of normality. Distance based methods are also computationally expensive in the test phase. Domain-based methods, which include Support Vector Machines, do not make any assumptions on data distribution and detect anomalies using only a small number of data, closest to the boundary. Therefore, they can be trained using relatively small database sizes, and training is fast. However, for the same reason, they are sensitive to outliers. Information-theoretic based methods are

highly dependent on the choice of the information theoretic measure, and it may be difficult to associate score with an anomaly. Additionally, information-theoretic based methods are computationally intensive in the test phase.

The proposed method belongs to the reconstruction-based anomaly detection methods. They typically do not make any assumptions regarding the properties of the data distribution. On the other hand, their structure depends on parameters that need to be optimized. The optimisation method choice is vital for the model performance (Kanarachos, Griffin, & Fitzpatrick, 2017; Piotrowski, 2014; Piotrowski & Napiorkowski, 2011). Furthermore, NN performance depends heavily on its structure. For example, NNs cannot learn easily long-term dependencies and are prone to overfitting. (Martens & Sutskever, 2012). To this end, a NN structure is proposed based on signal reconstruction using wavelets. Wavelets offer a better trade-off in the time/frequency resolution of the signal compared to Fourier Transform or Short Fourier Transform and their performance in detecting temporal anomalies, this paper's focus, is much better (Gao & Yan, 2011). Reconstructing the signal at multiple scales has several advantages a) Easier NN training, because NNs are required to learn less and more coherent features compared to when learning the complete signal and b) Easier learning of long term and short term interdependencies, as the signals at different scales represent the signal’s short and long term temporal structure.

In the anomaly detection phase, it is not proposed to use detectors at each scale because a) In many practical applications, some scales are not informative. Thus, it is more difficult to train a detector for each scale and then combine them and b) The interdependencies between different scales are not explicitly considered. If anomaly

detection is performed separately at each scale, it is implicitly assumed that the signals are independent. The reason for using Hilbert transform is because it focuses on the local features of the signal. In case the frequency content is indicative of an anomaly, this will be utilised in the anomaly detection. In the opposite case, as it happens in many nonlinear systems, the detection is based on the amplitude of the residual signal.

With regard to the anomaly detection threshold, no data distribution assumptions are made. The Receiver Operating Characteristics (ROC) method allows a comprehensive evaluation of the anomaly detector's performance, distinguishing between true positives, true negatives, false positives and false negatives. Furthermore, the particular ROC implementation facilitates –in a computational efficient manner, near real time– recalculating the confidence in the method's performance, when new anomalies are detected. In the future, it is envisaged to link the proposed ROC evaluation method to the training phase of the DNNs.

WANEH is computationally intensive in the training phase. Learning the system's normal behaviour is not trivial and may require several iterations. On the other hand, in the test phase the DNN implementation is computationally is very efficient. Although there are efficient techniques for analysing a signal in its wavelets components, its real-time implementation is still a challenge. Last but not least, it is emphasized that it is very likely domain specific methods to perform better than the proposed algorithm. However, the focus of this paper is on the development of a transferable anomaly detection method, which requires minimal tuning effort when applied in different domains.

3. Applications

3.1 Seismic Electric Signal application

It was previously demonstrated by P. A. Varotsos & Alexopoulos (1986) that when a critical stress is achieved on an ionic solid the existing electric dipoles orient cooperatively. This behavior results in the emission of an electric signal, which is transient (P. A. Varotsos, Sarlis, & Skordas, 2011). Thus, before an earthquake takes place, since the stress in the focal area (containing ionic materials) gradually increases, when a critical value is approached we may obtain characteristic Seismic Electric Signals (SES) (P. Varotsos et al., 1996; P. Varotsos & Lazaridou, 1991). These SES are low frequency (≤ 1 Hz) variations of the Earth's electric field (P. A. Varotsos & Alexopoulos, 1984) usually consisting of rectangular pulses. The suggested mechanism (P. A. Varotsos & Alexopoulos, 1986) of SES generation is as follows. In ionic solids, in addition to the common inherent defects in the lattice, there are always extrinsic defects owing to aliovalent impurities. Aliovalent impurities attract nearby intrinsic defects, forming electric dipoles the orientation of which can be changed by defect migration. Initially the orientations of these dipoles are random, but as the applied pressure increases a critical value is approached above which cooperative orientation of these dipoles occurs (Varotsos, Sarlis, & Skorda, 2011).(P. A. Varotsos et al., 2011).

In the present study, we apply our algorithm in order to detect anomalies of the Earth's electric field and thereby SES activity. Here we investigate as an example the case of SES activity that was recorded on the 18th and 19th April 1995, a few weeks before the M6.6 1995 Kozani-Grevena earthquake at 40.2°N21.7°E on 13 May 1995 at 08:47 UT. The SES time series that we analyse were recorded at the Ioannina measuring

station of the Solid Earth Physics Institute (IMS-SEPI). Figure 3 shows a 5h excerpt of measurements on a "normal" day.

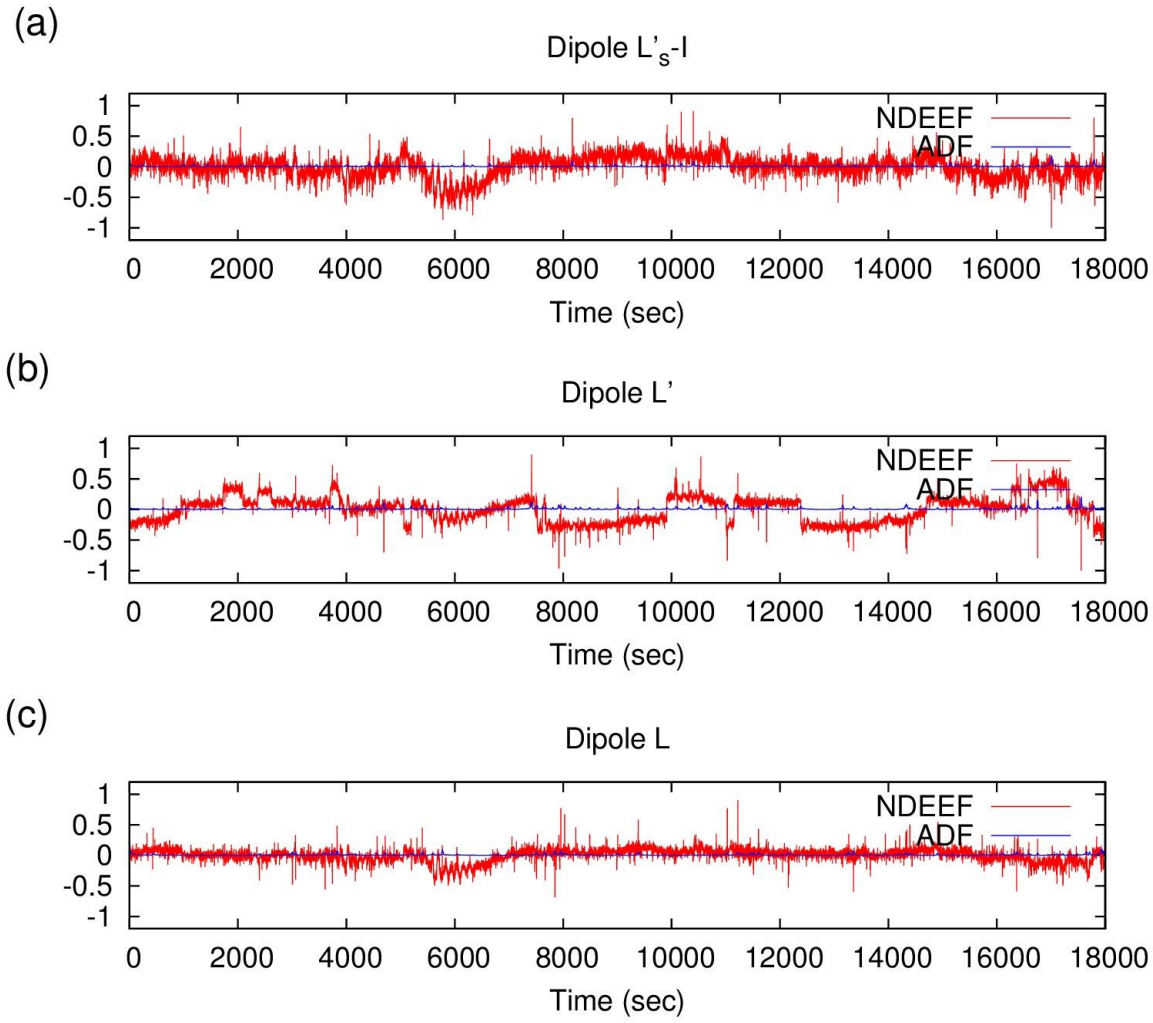

Fig. 3: Monitoring Earth's electrical activity: The normalized deflection of the Earth's electric field (NDEEF) (red) together with the Anomaly Detection Function (ADF) outcome (blue) for the dipoles L'$_s$-I, L' and L in (a), (b), and (c), respectively.

This station has several short and long dipoles in a number of directions (see map in Fig. 1 of supplemental material of Ref. (P. A. Varotsos, Sarlis, & Skordas, 2003) available from http://ftp.aip.org/epaps/phys_rev_lett/E-PRLTAO-91-007338/SupInfo.pdf). The SES activity can be distinguished from human-generated noise using the $\Delta V/L$ criterion:

$\Delta V/L$ is approximately the same when measured in parallel short and long dipoles) (Varotsos, & Lazaridou, 1991),(P. A. Varotsos & Lazaridou, 1991) thus if short and long dipoles are operating simultaneously one can identify SES activity. The application of the algorithm therefore focused on the long dipoles L'$_s$-I, L' and L since they are almost parallel and exhibit similar behaviour (see Fig. 4). The dichotomous (or binary) representation (DR) of an SES (Varotsos, Sarlis, & Skordas, 2002),(P. A. Varotsos, Sarlis, & Skordas, 2002) which is a two-valued (ON/OFF) time series that acquires the value OFF, e.g. $x_{DR}(t) = 0$, when there is no SES pulse at time *t* and the value ON, e.g. $x_{DR}(t) = 1$, when there is an ongoing SES pulse at that time is here compared with our outcome (in contrast, e.g. Ref. (P. A. Varotsos et al., 2002), the dichotomous representation of an SES activity can be represented by comparing simultaneously the value of the deflection of the electric field measured from short and long dipoles at time *t* with a threshold and indicates the existence of an SES pulse).

In Fig. 3 a 5h excerpt of the measurements of a "normal" day without anomalies in the SES are presented as-received from the three long dipoles at IMS-SEPI. The electric field is not constant and exhibits fluctuations that seem (quasi-)random. Additionally, we can observe spikes, and different values of the potential difference on each dipole. It is a challenge to detect anomalies against such a non-stable background with an Anomaly Detection Filter (ADF) algorithm. The proposed deep neural network learns not only how the signal behaves at different time scales but also the interdependencies between the scales themselves. This is a unique approach compared to other neural network or wavelet based methods (Alexandridis & Zapranis, 2013; Kocadağlı & Aşıkgil, 2014).

In Figs 4 and 5 the recorded normalized deflection (P. A. Varotsos, Sarlis, & Skordas, 2003) of the electric field of the earth (red lines), the Dichotomous Representation of the SES pulses (green lines) and the Anomaly Detection Filter outcome are presented for the long dipoles L'_s-I (Fig. 4a and Fig. 5a), L′ (Fig. 4b and Fig. 5b) and L (Fig. 4c and Fig. 5c) (blue lines) for 5h excerpts on 18 and 19/04/95. As we can see the ADF outcome and DR match well in all cases. We can better inspect the quality of matching, as example, in the zoom area of Fig. 3b if we focus on the local maxima of ADF and the DR pulses.

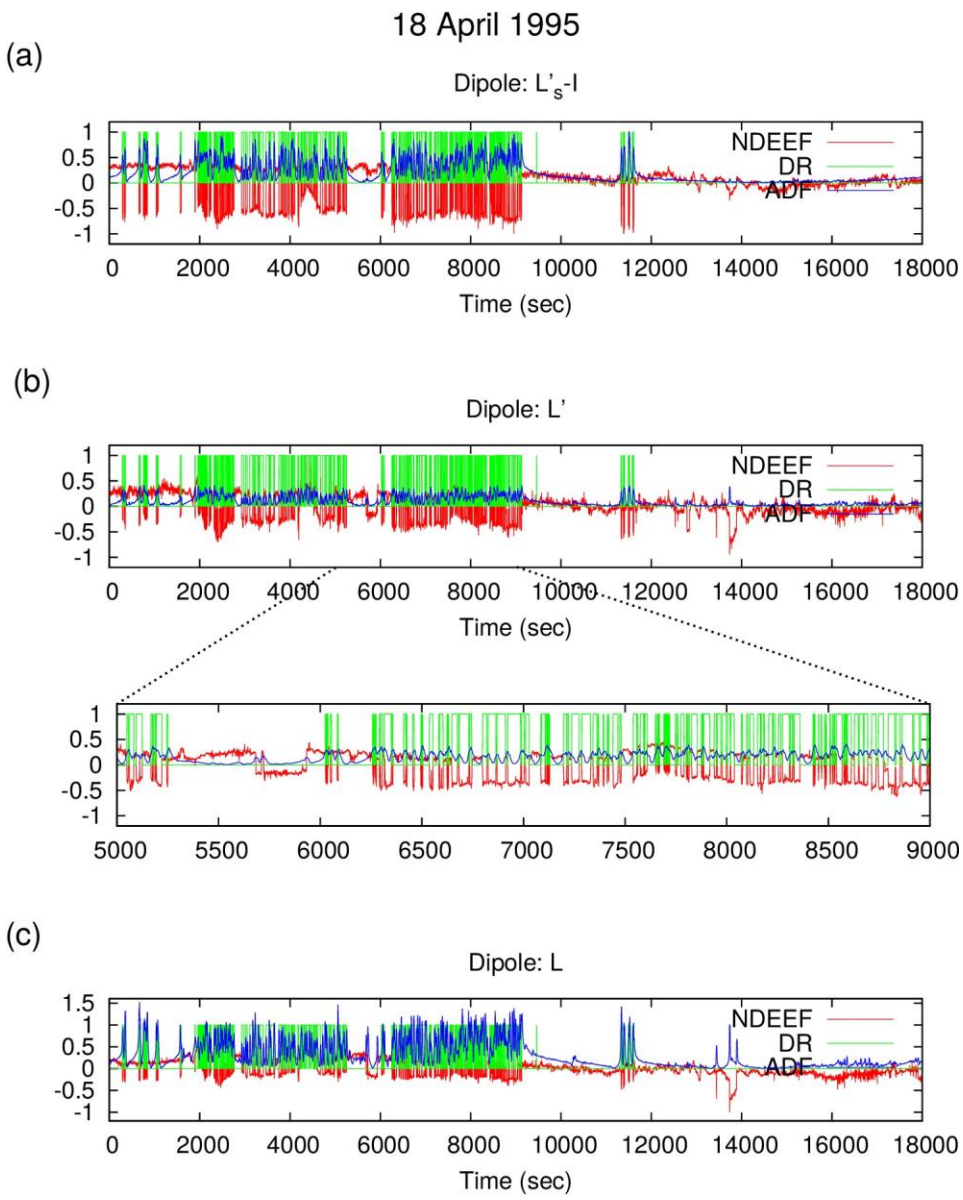

Fig. 4: Abnormal Earth Activity on 18/04/95: The normalized deflection of the Earth's electric field (NDEEF) (red), the dichotomous representation (DR) of SES (green) together with the Anomaly Detection Function (ADF) outcome (blue) for the dipoles L'$_s$-I, L' and L in (a), (b), and (c), respectively, on 18/04/95, that preceded the 13/05/95 M6.6 Kozani-Grevena earthquake. In panel (b) there is also a zoom area between 5000 and 9000 sec.

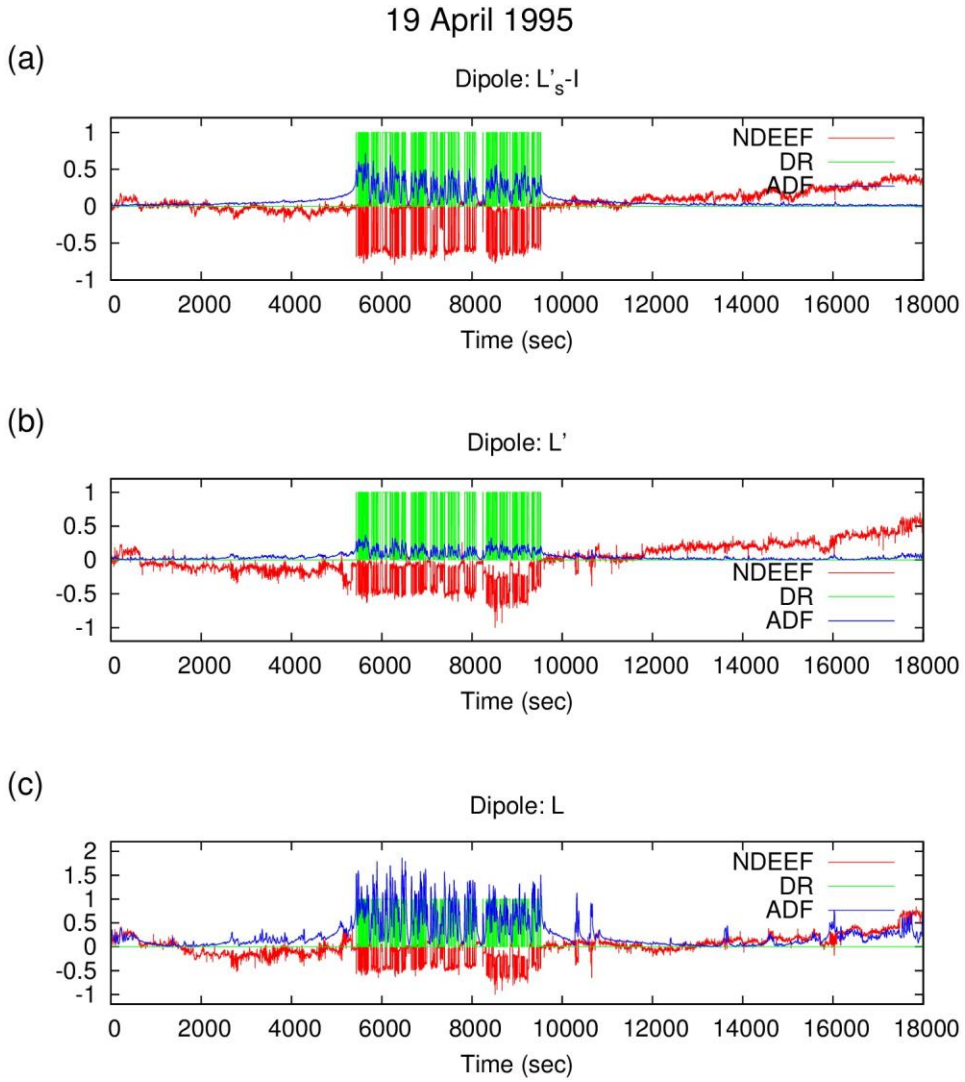

Fig. 5: Abnormal Earth Activity on 19/04/95: The normalized deflection of earth's electric field (NDEEF) (red), the Dichotomous Representation (DR) of the Seismic Electrical Signal (green) together with the Anomaly Detection Filter (ADF) outcome (blue) for the dipoles L'$_s$-I, L' and L in (a), (b), and (c), respectively, on 19/04/95.

In order to evaluate the SES anomaly detection we employ a Receiver Operating Characteristics (ROC) analysis (Fawcett, 2006). The ROC method can be used to evaluate an estimator (N. V. Sarlis, Christopoulos, & Bemplidaki, 2015); in this case as estimator k we are using the value of the ADF outcome. The DR index M takes value $M = 1$ when there is an SES pulse and $M = 0$ when there is not (see schematic diagram

in Fig. 6). The ROC graph (e.g., see Fig. 7) depicts the Hit rate (or True Positive rate) on the Y-axis and the False Alarm rate (or False Positive rate) on the X-axis. Here, we examine if the estimator k lies over a given threshold k_i. There are two classes for a Dichotomous Response, "Positive" when $M = 1$ and "Negative" when $M = 0$; and there are also two hypothesized classes "Yes" if $k \geq k_i$ and "No" if $k < k_i$. Therefore, there are four classifications: TP (True Positive) when $M = 1$ and $k \geq k_i$; FP (False Positive) when $M = 0$ and $k \geq k_i$; FN (False Negative) when $M = 1$ and $k < k_i$; and TN (True Negative) when $M = 0$ when $k < k_i$. Thus, the True Positive rate (TPr) is the ratio of TPs over the total number of "Positives" $P = TP + FN$; and the False Positive rate (FPr) is the ratio of FPs over the total number of "Negatives" $Q = FP + TN$. These are summarized in Fig. 6. For each value of k_i we obtain an operating point in the ROC graph.

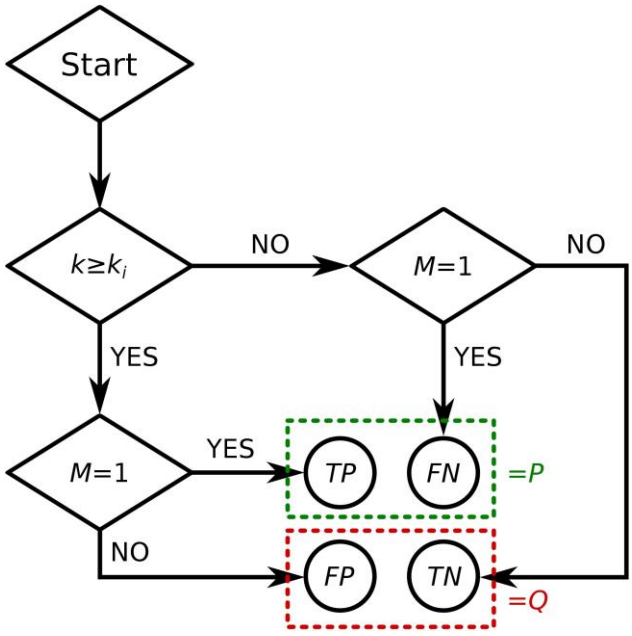

Fig. 6: ROC schematic: Diagram for Receiver Operating Characteristics classifications.

A random estimator will be located in the region close to the diagonal, where the true positive and false positive rates are roughly equal. An approach to evaluate the statistical significance of an ROC curve is to calculate the Area Under the Curve (AUC). Recently, for the statistical significance of the ROC curves a new visualization scheme has been proposed (N. V. Sarlis & Christopoulos, 2014) based on k-ellipses and with this technique among others, using the areas under the curves of the k-ellipses, one can measure the probability (p-value) for given values of P and Q to obtain an ROC curve by chance when ascribing "Yes" or "No" randomly.

The ROC graphs arising from the detection of the SES activity using the present method (i.e., the estimator k) are depicted in Fig. 7a for 18/04/95 and in Fig. 7b for 19/04/95; in both diagrams the ROC diagrams for the L's-I, L' and L dipoles are shown. We can observe similar behavior in all dipoles, which reveals the very good performance of the Anomaly Detection Filter regardless of the noisy background. In particular, the Areas Under the Curve for 18th of April are $AUC_{L's\text{-}I} = 0.8771$, $AUC_{L'} = 0.8266$, $AUC_{L} = 0.8445$ and for 19th of April are $AUC_{L's\text{-}I} = 0.9410$, $AUC_{L'} = 0.9269$, $AUC_{L} = 0.9054$: additionally the p-value in all of these cases is much smaller than 10^{-8} (10^{-8} is the accuracy of the VISROC.f FORTRAN code (N. V. Sarlis & Christopoulos, 2014)).

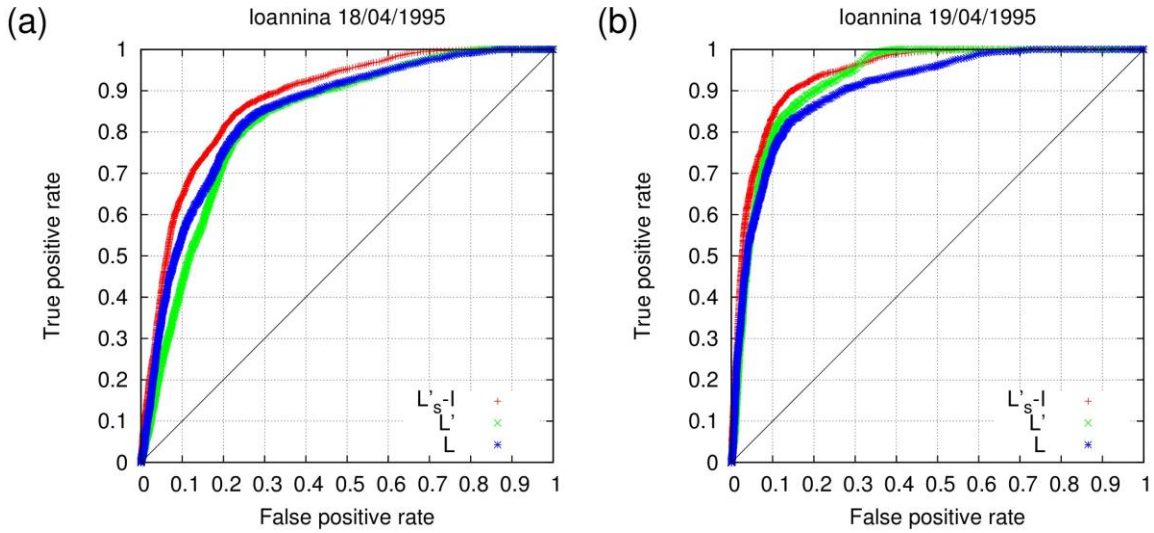

Fig. 7: Earth's electric activity- ROC output: Receiver Operating Characteristics when using the Anomaly Detection Filter outcome as estimator for the dipoles L'$_s$-I (red pluses), L' (green crosses) and L (blue asterisks) on (a) 18/04/95 and (b) 19/04/95.

Finally, we note the very good efficiency of the Anomaly Detection Filter since we obtain $TPr \approx 86\%, 81\%, 82\%$ with $FPr \approx 25\%$ for 18th April and $TPr \approx 90\%$, $86\%, 83\%$ with $FPr \approx 15\%$ for 19th April for the L'$_s$-I, L' and L dipoles, respectively. At this point it is useful to mention that the results for 19th April are better, and one of the reasons for this is that, as we can see in Fig. 5, some anomalous pulses (see the panels (b) and (c) between 13000 sec and 14300 sec at L' and L dipoles) were detected on 18 April but there is no SES activity since the $\Delta V/L$ criterion is not fulfilled. This shows that a vector generalization of the ADF algorithm may increase the already high SES anomaly detection efficiency. The fact that the algorithm is fast can be useful for decision-making during real-time data acquisition: a computer can analyze data excerpts and decide automatically on the basis of the estimator k whether this excerpt should be transmitted to a central station or not. Moreover, in an updated setup this decision may trigger the activation of a higher sampling rate.

3.2 Intelligent Transportation Application

Modern road vehicles are equipped with hundreds of sensors that monitor both the vehicle's behavior and the surrounding environment. Smartphones, which are nowadays widespread utilized, have numerous embedded sensors and contribute also to monitoring the vehicle and driver's states. Recent advances in mobile, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication enabled the connection of these sensors into the Internet of Things (Digital Agenda for Europe, The Internet of Things, 2015). The combination and utilization of information from multiple vehicles is expected to have a major impact on transportation safety, comfort and efficiency. However, the number of vehicles involved and the amount of data that can be gathered is massive, and this has implications. A potential remedy lies in the detection and isolation of the critical information at vehicle level.

Here, we present the automated detection of road anomalies (e.g. potholes, bumps) using smartphone sensors. The analysis is based on measurements collected using the in-built accelerometers of a smartphone while driving in Coventry city centre, UK (see Fig. 8). The smartphone was fixed using a mobile holder on the vehicle's front window and oriented at an angle of about 30˚ with respect to the vertical. The sampling rate was 10Hz.

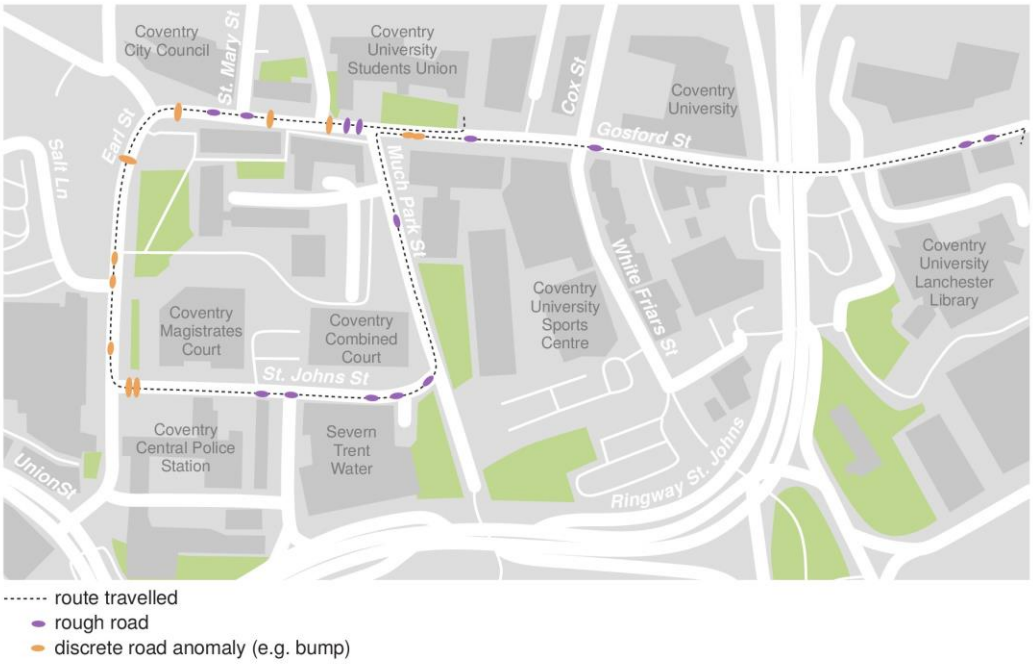

Fig. 8: Road anomaly informed map: Detection and visualization of road anomalies. The schematic was drawn for this paper using Adobe Illustrator.

As observed in Fig. 9a the accelerometer's signal is noisy and therefore detection of road anomalies at low speeds is difficult. The vehicle speed under which the experiments took place was below 10 m s^{-1}. Road anomalies, Fig. 9b, were labeled by correlating acceleration signals, video images recorded from within the vehicle, and obstacle verbal descriptions during the test runs.

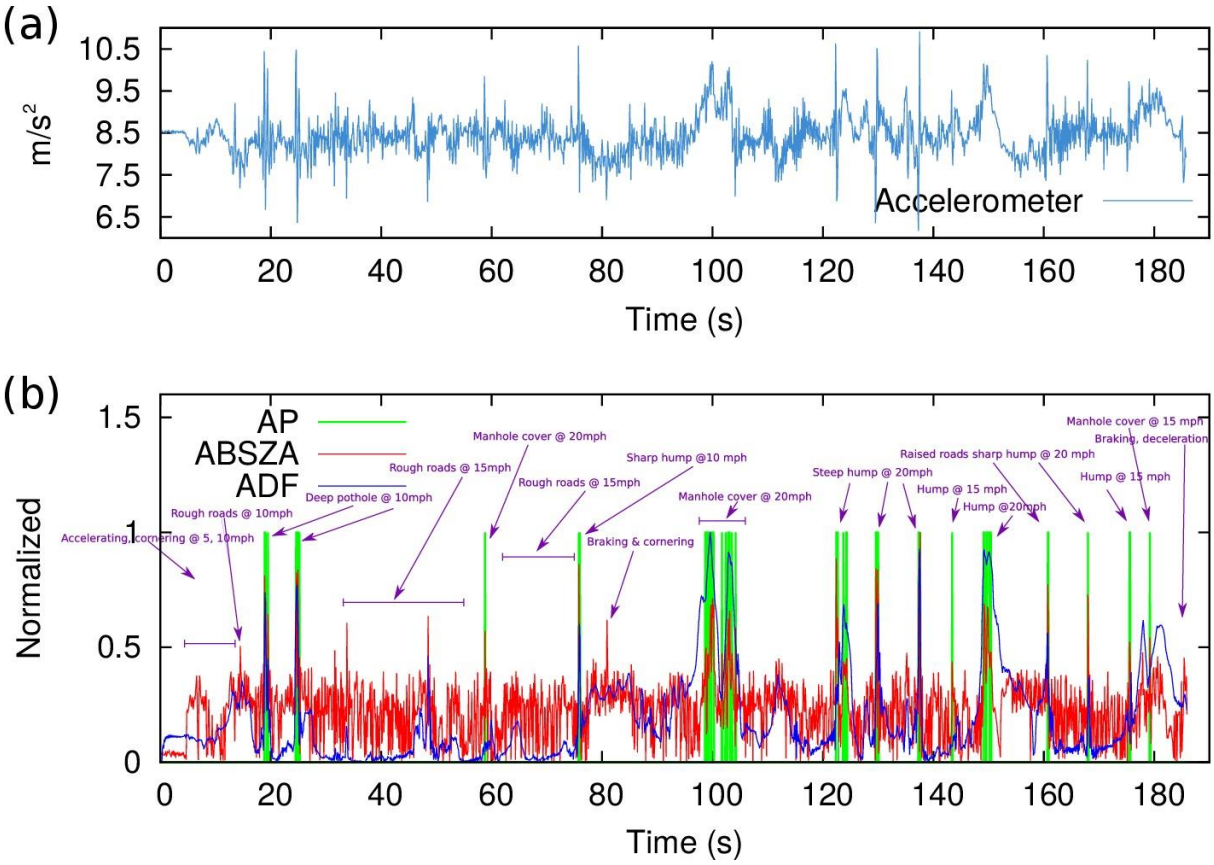

Fig. 9: Acceleration signal and road anomaly detection: (a) The recorded time series from the smartphone accelerometer whilst driving in Coventry City Centre. (b) The Anomalous Pulses (AP) (green), the normalized absolute value of *z*-score of the recorded values from the accelerometer (ABSZA) together with the normalized Anomaly Detection Filter (ADF) outcome (blue) for the case of road anomaly detection. The blue signal is shifted three places with respect to the original cyan signal as the neural network uses a buffer of size 3.

The vehicle dynamics case is completely different to the Seismic Electric Signal one described in the previous section. The system behaves deterministically and the response depends mainly on the type of road disturbances and the vehicle's dynamical properties (springs, masses and dampers). The challenge here is that the vehicle dynamical properties are in general unknown and costly to obtain. Additionally, a number of vehicle parameters may vary: e.g. the passengers' total mass. Furthermore, the sensors used for

measuring the vibrations are low-cost Micro Electro-Mechanical Systems (MEMS) and differ between smartphones. Thus, the sensor quality, sampling rate, and the support conditions differ. Last but not least, handling manoeuvres such as braking or cornering induce disturbances to the system's response and may fool the anomaly detector.

The acceleration signal, shown in Fig. 9a, was de-noised using the Daubechies 9 mother wavelet up to the 8th level. The DNN structure was built on the wavelet decomposition. The nonlinear autoregressive neural network was trained using the vertical acceleration signal obtained while driving on smooth road surfaces. It is highlighted that as we use a buffer of size 3 we shift the neural network's output three places when plotting *y* in Fig. 9b. The shifted signal is also used for the remaining operations that involve the calculation of the error function *e*, the Hilbert Transform, and the ROC analysis.

In order to indicate humps, manhole covers and potholes in the time series, we define a threshold ($Th = 2.16$) on the absolute value of z-score of the recorded values from the accelerometer and we rejected all the cases that are not marked as such in Fig. 9b. Thus, we made a series of Anomalous Pulses (AP): the AP index M_2 takes value $M_2 = 1$ when there is a road anomaly and $M_2 = 0$ when there is not (see Fig. 9b). Finally, we employ the Receiver Operating Characteristics analysis to evaluate the detection (see Fig. 10, where the yellow, green and black lines depict k-ellipses for p-values of 1%, 5% and 10% respectively (N. V. Sarlis & Christopoulos, 2014)). The Area Under the Curve here is $AUC = 0.9078$ and the p-value is much smaller than 10^{-8}, as in the case of the seismic electrical application. We can observe again the very good efficiency of the Anomaly

Detection Function since we obtain a True Positive Rate of ≈ 81% with a False Positive Rate ≈ 12%.

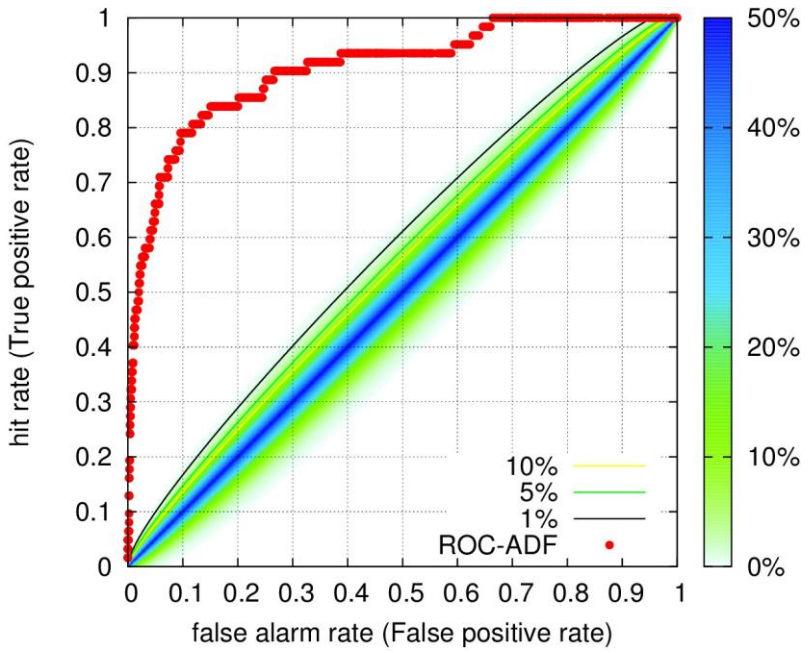

Fig. 10: Road anomalies – ROC output: Receiver Operating Characteristics plot when using the Anomaly Detection Filter (ADF) outcome as an estimator for the detection of road anomalies. The k-ellipses for p-values 10%, 5% and 1% are depicted with yellow, green and black solid lines respectively.

Despite the structural uncertainties and the low signal-to-noise ratio the anomaly detector performs quite robustly. At higher vehicle speeds, where the signal-to-noise ratio increases, the algorithm performs even better. Another point for discussion is the low sampling rate of the smartphone used in this example. At a speed of 10 m/s and a sampling rate of 10 Hz the spatial resolution is about 1 m, which means that the acquired signal is – to a certain extent – already filtered. Many smartphones -currently offered in the market- can acquire acceleration signals with sample rates of up to 100 Hz.

4. Results and discussion

The present study introduces WANEH signal processing algorithm that combines wavelets, neural networks and Hilbert transform to detect anomalies in time series data. The following non-application-specific components summarize the algorithm:

4.1 Sensor induced noise rejection

Wavelet decomposition is applied to remove from signal x the sensor induced noise. The resulting signal x_d is then used to build the DNN. A comparison between different wavelet bases -including db2, db4, db5 and db9- has shown that db9 performs more robustly. This is probably due to that db9 approximates well polynomials of 3rd degree. Although in many cases decomposition up to the 5th level is adequate, it was found that the 8th level decomposition provided a better estimation of the long term dependencies. It is well known that a pseudo-frequency f_{wm} is assigned to each scale m:

$$f_{wm} = \frac{f_c}{a_0^m \cdot \Delta}$$	(24)

where f_c is the center frequency of the wavelet and Δ is the sampling period. Thus, the higher scale m of wavelet decomposition, the better the description of long term signal dependencies. Equation (24) also demonstrates Heisenberg's principle: the larger the scale m, the finer the frequency resolution obtained. On the other hand, due to wavelet's dilation, the temporal resolution is coarser. The de-noised signal x_d is obtained using Equations (13) and (14).

4.2 Deep Learning

Here a nonlinear autoregressive neural network (NN) is employed for predicting x_d. The main assumption is that the NN approximates only the system's normal behaviour. Under this assumption, in the case of an anomaly, it is expected that the NN will fail to predict accurately the system's output x_d. The error $e = x_d - y$ is then used to detect the anomaly.

The actual implementation is based on the Deep Learning paradigm (LeCun, Bengio, & Hinton, 2015).(LeCun, Bengio, & Hinton, 2015). Instead of training only one neural network, a collection of multiple stacked neural networks is employed to learn the signal's x_d underlying temporal structure. (Längkvist et al., 2014).

For this, a novel deep temporal neural network is proposed. The architecture is given in **Fig. 1**d. It is well known that NNs are weak in learning long-term trends, and the proposed architecture overcomes the problem. The first part is a set of stacked neural networks that models x_d at different time scales. The second part is an autoregressive neural network consisting of 10 hidden neurons with nonlinear (log-sigmoid) activation functions and a three-layer buffer. Although the exact number of hidden layers and buffer size are problem-dependent, it was found that relatively simple neural networks (number of neurons less than 10) cannot describe temporal dynamics sufficiently. Different standard neural network training algorithms were evaluated – including Levenberg-Marquardt, BFGS Quasi-Newton, Fletcher-Powell Conjugate Gradient and Bayesian regularization; among them, Bayesian regularization backpropagation performed more robustly. Numerous numerical experiments with different sized buffers have shown that a

large buffer size decreases the detector's performance. In the problems studied, a buffer size with three to five elements presented the best performance. The implication of using a three-layer buffer is that anomaly detection can only start when at least three samples at each scale are acquired.

4.3 Feature extraction

The instantaneous energy and frequency content of signal *e* are extracted using Hilbert transform. Hilbert transform is useful for identifying instantaneous frequency changes also in the higher frequency spectrum, where wavelet transform is not performing well. In case the instantaneous frequency is not useful to detect anomalies, the signal's envelope is exploited instead.

4.4 Probabilistic inference

Here a threshold is derived based on the statistical significance of the classifier's Receiver Operating Characteristics (ROC) (N. V. Sarlis & Christopoulos, 2014). This approach requires only few anomalous data as opposed to other methods (e.g. Christopoulos & Sarlis, 2017). We calculated that the statistical significance is far beyond chance and the efficiency is very high. Using probabilistic ROC method and decreasing in these cases the threshold k_i of the estimator k, one can find the value of threshold k_i (threshold of ADF outcome in these cases) that we have the best efficiency or a specific efficiency (e.g. the value of k_i in order to have FPr less than 10%) and the corresponding p-value to found the statistical significance of these results even with small total numbers of "Positives" and "Negatives". Practically, based on the results of probabilistic ROC we

can trust the detection for values of ADF greater or equal to a threshold that gives the best efficiency knowing in the same time the probability to have these results by chance.

4.5 Numerical results and comparison to existing methods

The performance of the algorithm is illustrated by two examples, characterized by completely different underlying dynamics. In the SES activity case, the time series signal NDEEF is non-stationary. The signal range for both “normal” and “abnormal” earth activity is the same. The record exhibits frequent spikes, not periodic, which can mislead the anomaly detector. To our knowledge, this is the first time where an anomaly detector for abnormal Seismic Electrical Activity was successfully applied.

In the intelligent transportation case, the noise to signal ratio is high due to the smartphone's sensor characteristics. Although, in this case, the dynamic system is less complex the properties of its components are uncertain and variable. Manoeuvres such as braking can introduce significant pitch and therefore can potentially mislead the anomaly detector. For the road anomaly case, different detection methods, using smartphone signals, have been proposed.

In the following, a comparison is made between WANEH, Method A (Vittorio et al., 2014) and method B (Cong et al., 2013). Method A refers to an acceleration-based anomaly detection method, classified as a probabilistic method. If acceleration exceeds a certain threshold then it is classified as an anomaly. Method B is based on signal’s wavelet decomposition and exploitation of the detail coefficients. Support Vector Machines are employed to map the detail coefficients to the anomalies. Method B is classified as a clustering technique.

Although it is possible to develop Method A by studying only the system's normal behaviour, this is not possible with Method B. For completeness of the comparison, two different versions of Method B were implemented. In the first version, method B1, the Support Vector Machine is trained using data that include all road anomalies. In the second version, method B2, data that include only half of the anomalies are used for training the Support Vector Machine. The ROC for WANEH, method A, method B1 and B2 are shown in Fig. 11 with red, purple, orange, and dark blue solid circles respectively. As observed the three other methods cannot obtain hit rate higher than approximately 41% and their performance is very close (if not identical) to that of the current method when focusing on operating points with very small false alarm rate. In simple words, the current method performs equally good as the other three methods, but it additionally allows the selection of higher hit rates (at the expense of course of higher false alarm rates).

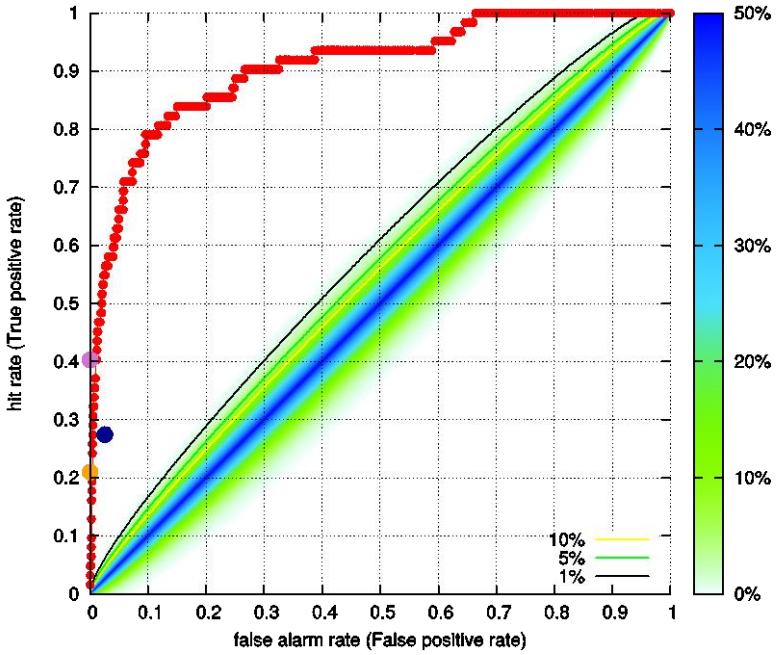

Fig. 11: Comparison between WANEH (red solid circles), Method A (dark blue solid circle) and two versions of Method B, B1 (orange solid circle) and B2 (purple solid circle).

To further elaborate the necessity of DNNs, a comparison is made between the performances of WANEH when using DNNs and when using shallow autoregressive NNs. For the latter case, we trained an autoregressive neural network with buffer size three and ten hidden neurons to follow signal x. In Fig. 12 the results obtained using WANEH with DNNs and WANEH without DNNs with the red and dark purple solid circles, respectively, are illustrated. As observed the AUC of the WANEH with DNNs ($AUC = 0.9078$) is grater than the AUC of the WANEH without DNNs ($AUC = 0.7691$). In simple words, the use of the WANEH with DNNs improves the method.

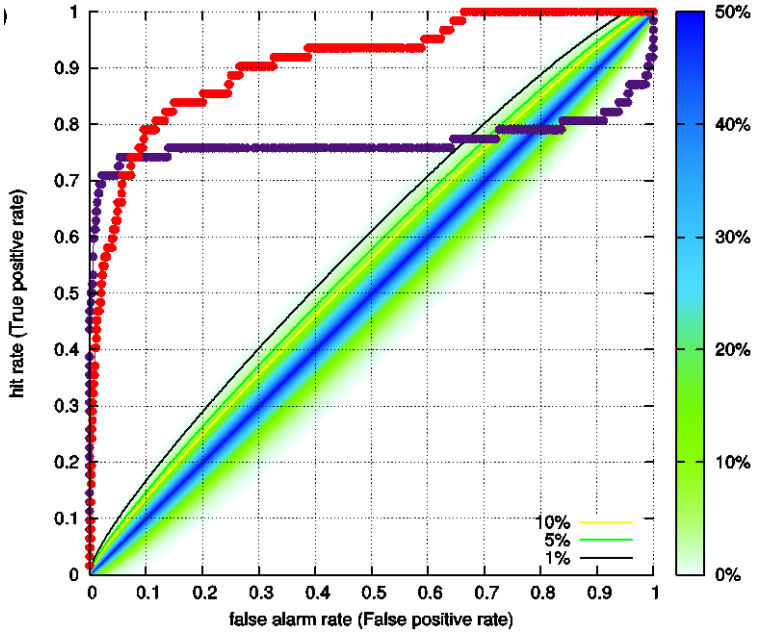

Fig. 12: Comparison between WANEH with DNNs and without DNNs (shallow autoregressive neural network) with red and dark purple solid circles, respectively.

In conclusion, the proposed anomaly detection method is transferable, without significant manual tuning effort. Thus, it can be utilized in different applications, without employing domain-specific knowledge. The method is based on learning the system's normal behaviour, which can be challenging in certain applications. The method is computationally expensive in the training phase, as it is required to learn the system's normal behaviour. In the testing phase its application is near real-time. In the road anomaly case, WANEH performed better compared to methods known from the literature. In other applications, where the interactions between system components are well understood, it is expected that domain-specific methods perform better.

5. Conclusions and future work

Anomaly detection in time series data is a significant problem with applications in many different domains including medicine, physics, engineering, and finance. Although many methods have been proposed up to now most of them share some common disadvantages; they are application specific, require large amounts of (existing) anomalous data, are not (near) real-time capable, can't distinguish complex anomalies or provide a confidence measure over the classifier's outcome.

In this study a systematic method, designed to overcome the aforementioned shortcomings, is presented. The method, inspired by the deep learning paradigm, combines neural networks, wavelet analysis and Hilbert transform to distinguish anomalous from normal operation. The probabilistic Receiver Operating Characteristics (ROC) method is employed for determining the anomaly detection threshold and the statistical significance of the detector's outcome. A theoretical comparison between the proposed method and other categories of anomaly detection methods, highlighting advantages and disadvantages, was provided.

Our main contributions to the field of "Anomaly Detection" are a) the proposition of a unique Deep Neural Network structure for reconstructing the normal behaviour and b) the time-frequency analysis of the residual signal in relation to probabilistic ROC. The anomaly detection method is based solely on signals describing the system's normal behaviour. The method facilitates online learning and is adaptive because the training phase extend beyond the point of deployment. These attributes are important for

applications in which anomalies are rare or not standard and for applications where normal behaviour requires customization, for example, vehicle type or patient.

The algorithm is fast; however it is not real-time as it requires a number of samples and processing time to detect an anomaly. Nonetheless, decision-making strategies based on the results of this algorithm can expedite real-time data analysis and acquisition. The training phase can be computationally demanding, as it is required to learn the normal system behaviour. In some applications, this may be challenging. The selection of optimisation method is crucial for the performance of the method. This is a research field where significant contributions are expected, in the forthcoming years. In this paper, guidelines were given considering the most popular NN training methods.

The efficacy and multidisciplinary nature of this algorithm has been examined for two diverse examples where near real-time detection of anomalies is important: in the case of Seismic Electrical Activity for taking precautionary measures to minimize the effects of an imminent disaster; and for road anomaly detection to highlight locations requiring maintenance and repair. The statistical significance of the classifier's outcome has been evaluated using probabilistic Receiver Operating Characteristics and revealed the good performance of the algorithm.

In the road anomaly case, the proposed algorithm was compared to two known road anomaly detection methods a) one based on a probabilistic method and b) one based on wavelet decomposition and Support Vector Machines. The numerical results confirm the good performance of the proposed WANEH algorithm.

Future work includes the automated calculation of the detection threshold by coupling the anomaly detection filter with the probabilistic Receiver Operating

Characteristics. Furthermore, it is envisaged to further improve the success rate of the anomaly detection filter by combining the outcome from different sources-signals e.g. acceleration in the longitudinal and vertical direction.

Acknowledgments

S.-R.G.C., A.C. and M.E.F are grateful for funding from the Lloyd's Register Foundation, a charitable foundation helping to protect life and property by supporting engineering-related education, public engagement and the application of research. S.-R.G.C would like to express his sincere thanks to the Director of the SEPI Professor Panayiotis Varotsos for useful discussions and for providing the data of IMS-SEPI.

Author contributions

S.K. and S.-R.G.C. performed the calculations. All the authors analysed the results and contributed to the discussion and preparation of the paper.

Additional information

Competing financial interests: The authors declare no competing financial interests.

References

Addison, P. (2002). *The Illustrated Wavelet Transform Handbook*. Informa UK Limited. https://doi.org/10.1201/9781420033397

Akhoondzadeh, M. (2015). Application of Artificial Bee Colony algorithm in {TEC} seismo-ionospheric anomalies detection. *Advances in Space Research*, *56*(6), 1200–1211. https://doi.org/http://dx.doi.org/10.1016/j.asr.2015.06.024

Akouemo, H. N., & Povinelli, R. J. (2015). Probabilistic anomaly detection in natural gas time series data. *International Journal of Forecasting*. https://doi.org/http://dx.doi.org/10.1016/j.ijforecast.2015.06.001

Alexandridis, A. K., & Zapranis, A. D. (2013). Wavelet neural networks: A practical guide. *Neural Networks*, *42*, 1–27. https://doi.org/http://dx.doi.org/10.1016/j.neunet.2013.01.008

Chen, X., & Zhan, Y. (2008). Multi-scale anomaly detection algorithm based on infrequent pattern of time series. *Journal of Computational and Applied Mathematics*, *214*(1), 227–237. https://doi.org/http://dx.doi.org/10.1016/j.cam.2007.02.027

Choudhury, S., Aguiar, J. A., Fluss, M. J., Hsiung, L. L., Misra, A., & Uberuaga, B. P. (2015). Non-uniform Solute Segregation at Semi-Coherent Metal/Oxide Interfaces. *Scientific Reports*, *5*, 13086.

Christopoulos, S.-R. G., & Sarlis, N. V. (2017). An Application of the Coherent Noise Model for the Prediction of Aftershock Magnitude Time Series. *Complexity*, *2017*, 27. https://doi.org/10.1155/2017/6853892

Cong, F., Hautakangas, H., Nieminen, J., Mazhelis, O., Perttunen, M., Riekki, J., & Ristaniemi, T. (2013). Applying Wavelet Packet Decomposition and One-Class Support Vector Machine on Vehicle Acceleration Traces for Road Anomaly Detection. In C. Guo, Z.-G. Hou, & Z. Zeng (Eds.), *Advances in Neural Networks*

– *ISNN 2013: 10th International Symposium on Neural Networks, Dalian, China, July 4-6, 2013, Proceedings, Part I* (pp. 291–299). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39065-4_36

Digital Agenda for Europe, The Internet of Things - Digital Agenda for Europe - European Commission. (2015). Retrieved from https://ec.europa.eu/digital-agenda/en/internet-things

Donoho, D. L., & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, (3), 879–921. https://doi.org/10.1214/aos/1024691081

Fan, Y., Osetskiy, Y. N., Yip, S., & Yildiz, B. (2013). Mapping strain rate dependence of dislocation-defect interactions by atomistic simulations. *Proceedings of the National Academy of Sciences*, *110*(44), 17756–17761.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, *27*, 861–874.

Gao, R. X., & Yan, R. (2011). *Wavelets* (1st ed.). Springer US. https://doi.org/10.1007/978-1-4419-1545-0

Georgoulas, G., Loutas, T., Stylios, C. D., & Kostopoulos, V. (2013). Bearing fault detection based on hybrid ensemble detector and empirical mode decomposition. *Mechanical Systems and Signal Processing*, *41*(1–2), 510–525. https://doi.org/http://dx.doi.org/10.1016/j.ymssp.2013.02.020

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459.

Goswami, J. C., & Hoefel, A. E. (2004). Algorithms for estimating instantaneous frequency. *Signal Processing*, *84*(8), 1423–1427. https://doi.org/10.1016/j.sigpro.2004.05.016

Harrou, F., Kadri, F., Chaabane, S., Tahon, C., & Sun, Y. (2015). Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, *88*, 63–77. https://doi.org/http://dx.doi.org/10.1016/j.cie.2015.06.020

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., … Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, *5*, 11476.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434. https://doi.org/http://dx.doi.org/10.1016/j.tics.2007.09.004

Ikonomopoulos, A., Alamaniotis, M., Chatzidakis, S., & Tsoukalas, L. H. (2013). Gaussian Processes for State Identification in Pressurized Water Reactors. *Nucl. Technol*, *182*, 1–12.

Introducing practical and robust anomaly detection in a time series. (n.d.). Retrieved March 29, 2017, from https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series

Kanarachos, S. (2012). A new min-max methodology for computing optimised obstacle avoidance steering manoeuvres of ground vehicles. *International Journal of Systems Science*, 45(5),

Kanarachos, S., Griffin, J., & Fitzpatrick, M. E. (2017). Efficient truss optimization using the contrast-based fruit fly optimization algorithm. *Computers & Structures*, *182*, 137–148. https://doi.org/10.1016/j.compstruc.2016.11.005

Kocadağlı, O., & Aşıkgil, B. (2014). Nonlinear time series forecasting with Bayesian neural networks. *Expert Systems with Applications*, *41*(15), 6596–6610. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2014.04.035

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, *42*, 11–24. https://doi.org/http://dx.doi.org/10.1016/j.patrec.2014.01.008

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, D., Liu, S., & Zhang, H. (2015). Negative selection algorithm with constant detectors for anomaly detection. *Applied Soft Computing*, *36*, 618–632. https://doi.org/http://dx.doi.org/10.1016/j.asoc.2015.08.011

Martens, J., & Sutskever, I. (2012). Training Deep and Recurrent Networks with Hessian-Free Optimization. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade: Second Edition* (pp. 479–535). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_27

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Piotrowski, A. P. (2014). Differential Evolution algorithms applied to Neural Network training suffer from stagnation. *Applied Soft Computing*, *21*, 382–406. https://doi.org/10.1016/j.asoc.2014.03.039

Piotrowski, A. P., & Napiorkowski, J. J. (2011). Optimizing neural networks for river flow forecasting – Evolutionary Computation methods versus the Levenberg–

Marquardt approach. *Journal of Hydrology*, *407*(1–4), 12–27. https://doi.org/10.1016/j.jhydrol.2011.06.019

Rojas, R. (1996). *Neural Networks: A Systematic Introduction* (1st ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-61068-4

Rushton, M. J. D., & Chroneos, A. (2014). Impact of uniaxial strain and doping on oxygen diffusion in CeO2. *Scientific Reports*, *4*, 6068.

Sarlis, N. V., & Christopoulos, S.-R. G. (2014). Visualization of the significance of Receiver Operating Characteristics based on confidence ellipses. *Comput. Phys. Commun.*, *185*, 1172–1176.

Sarlis, N. V., Christopoulos, S.-R. G., & Bemplidaki, M. M. (2015). Change ΔS of the entropy in natural time under time reversal: Complexity measures upon change of scale. *EPL*, *109*(1), 18002. https://doi.org/http://dx.doi.org/10.1209/0295-5075/109/18002

Sarlis, N. V., Skordas, E. S., Varotsos, P. A., Nagao, T., Kamogawa, M., & Uyeda, S. (2015). Spatiotemporal variations of seismicity before major earthquakes in the Japanese area and their relation with the epicentral locations. *Proc. Natl. Acad. Sci. USA*, *112*(4), 986–989. https://doi.org/10.1073/pnas.1422893112

Sun, W., Jayaraman, S., Chen, W., Persson, K. A., & Ceder, G. (2015). Nucleation of metastable aragonite CaCO3 in seawater. *Proceedings of the National Academy of Sciences*, *112*(11), 3199–3204. https://doi.org/10.1073/pnas.1423898112

Varotsos, P. A., Sarlis, N. V., & Skordas, E. S. (2002). Long-range correlations in the electric signals that precede rupture. *Phys. Rev. E*, *66*, 11902. https://doi.org/10.1103/PhysRevE.66.011902

Varotsos, P. A., Sarlis, N. V., & Skordas, E. S. (2003). Attempt to distinguish electric signals of a dichotomous nature. *Phys. Rev. E*, *68*, 31106. https://doi.org/10.1103/PhysRevE.68.031106

Varotsos, P. A., Sarlis, N. V., & Skordas, E. S. (2011). *Natural Time Analysis: The new view of time. Precursory Seismic Electric Signals, Earthquakes and other Complex Time-Series*. Berlin Heidelberg: Springer-Verlag.

Varotsos, P. A., Sarlis, N. V., Skordas, E. S., & Lazaridou, M. S. (2007). Identifying sudden cardiac death risk and specifying its occurrence time by analyzing electrocardiograms in natural time. *Appl. Phys. Lett.*, *91*, 64106.

Varotsos, P., & Alexopoulos, K. (1984). Physical Properties of the variations of the electric field of the earth preceding earthquakes, I. *Tectonophysics*, *110*, 73–98. https://doi.org/10.1016/0040-1951(84)90059-3

Varotsos, P., & Alexopoulos, K. (1986). *Thermodynamics of Point Defects and their Relation with Bulk Properties*. Amsterdam: North Holland.

Varotsos, P., Eftaxias, K., Lazaridou, M., Nomicos, K., Sarlis, N., Bogris, N., … Kopanas, J. (1996). Recent earthquake prediction results in Greece based on the observation of Seismic Electric Signals. *Acta Geophys. Pol.*, *44*, 301–327.

Varotsos, P., & Lazaridou, M. (1991). Latest aspects of earthquake prediction in Greece based on Seismic Electric Signals. *Tectonophysics*, *188*, 321–347. https://doi.org/10.1016/0040-1951(91)90462-2

Varotsos, P. V., Sarlis, N. V., & Skordas, E. S. (2003). Electric Fields that "arrive" before the time derivative of the magnetic field prior to major earthquakes. *Phys. Rev. Lett.*, *91*, 148501. https://doi.org/10.1103/PhysRevLett.91.148501

Vittorio, A., Rosolino, V., Teresa, I., Vittoria, C. M., Vincenzo, P. G., & Francesco, D. M. (2014). Automated Sensing System for Monitoring of Road Surface Quality by Mobile Devices. *Transportation: Can We Do More with Less Resources? – 16th Meeting of the Euro Working Group on Transportation – Porto 2013*, *111*, 242–251. https://doi.org/10.1016/j.sbspro.2014.01.057