Anomaly Detection in Electrocardiogram Readings with Stacked LSTM Networks

Thill, M.; Däubener, S.; Konen, W.; Bäck, T.H.W.; Barancikova, P.; Holena, M.; ... ; Rosa, R.

Anomaly Detection in Electrocardiogram Readings with Stacked LSTM Networks

Markus Thill[1], Sina Däubener, Wolfgang Konen[1], and Thomas Bäck[2]

[1] TH Köln – Cologne University of Applied Sciences, 51643 Gummersbach, Germany,
{markus.thill, wolfgang.konen}@th-koeln.de
[2] Leiden University, LIACS,
2333 CA Leiden, The Netherlands,
t.h.w.baeck@liacs.leidenuniv.nl

Abstract: Real-world anomaly detection for time series is still a challenging task. This is especially true for periodic or quasi-periodic time series since automated approaches have to learn long-term correlations before they are able to detect anomalies. Electrocardiography (ECG) time series, a prominent real-world example of quasi-periodic signals, are investigated in this work. Anomaly detection algorithms often have the additional goal to identify anomalies in an unsupervised manner.

In this paper we present an unsupervised time series anomaly detection algorithm. It learns with recurrent Long Short-Term Memory (LSTM) networks to predict the normal time series behavior. The prediction error on several prediction horizons is used to build a statistical model of normal behavior. We propose new methods that are essential for a successful model-building process and for a high signal-to-noise-ratio. We apply our method to the well-known MIT-BIH ECG data set and present first results. We obtain a good recall of anomalies while having a very low false alarm rate (FPR) in a fully unsupervised procedure. We compare also with other anomaly detectors (NuPic, ADVec) from the state-of-the-art.

1 Introduction

Anomaly detection in time series is of increasing importance in many application areas, e.g. health care [6, 4], sensor networks [20, 14] or predictive maintenance [7]. Anomaly detection is a very active research area (for an overview see [3]), yet it is not easy to come up with a general definition of an anomaly. The notion of an anomaly greatly depends on the application area and on characteristics of the time series in question. To learn the characteristics of nominal and anomalous behavior from data it is often necessary to apply sophisticated methods from natural computing (evolutionary neural networks [11] and immune systems [26]). While some anomalies are simple to detect (e.g. a sudden spike in a relatively constant signal may be detected by simple threshold heuristics), other anomalies are subtle and more complex to detect. This is especially the case for periodic or quasi-periodic time series where the anomaly may be a time-shifted peak, a peak with a different form or other patterns which only emerge from a long-range analysis of the signal.

Electrocardiography (ECG) time series constitute a prominent real-world example of quasi-periodic signals. Anomaly detection in ECG readings plays an important role in health care and medical diagnosis. There exist well-maintained databases, e.g. the MIT-BIH database [9], where a large body of data is annotated with numerous types of anomalies which are characterized and verified by medical experts. Automated anomaly detection in such ECG data is still a challenging topic, because the deviations from nominal behavior are often subtle and require long-range analysis. Furthermore, there are considerable signal variations from patient to patient or even within an ECG time series.

Long short-term memory (LSTM) networks [13], which are a special form of recurrent neural networks (RNN) and thus belong to the class of deep learning methods, have proven to be particularly useful in learning sequences with long-range dependencies. They avoid the vanishing gradient problem [12] and are more stable and better scalable [10] than other RNN architectures. LSTMs have been successfully advanced the state-of-the-art in many application areas like language modeling and translation, acoustic modeling of speech, analysis of audio data, handwriting recognition and others [10]. We will use stacked LSTMs as the building block for our ECG time series prediction.

It is the purpose of the present paper to investigate whether anomaly detection in quasi-periodic ECG time series can be trained in an *unsupervised* manner, hence, without the usage of the anomaly class labels. We will describe in Sec. 2 an LSTM prediction model which is trained to predict over multiple horizons and is applied to time series containing nominal and also rare anomalous data. We observe multidimensional error vectors (one vector for each point in time) and fit a multivariate Gaussian distribution to them. Based on the Mahalanobis distance we can assign to each point in time a probability of being anomalous. Sec. 3 describes our experimental setup and the MIT-BIH Arrhythmia Database used in our experiments. Sec. 4 presents and discusses our results, while Sec. 5 concludes.

1.1 Related Work

Anomaly detection in general has been done with methods from machine learning [3] and more precisely from natural computing: Han and Cho [11] and other works cited therein use evolutionary approaches in optimizing neural networks for the task of intrusion detection. Kieu et al. [15] use deep learning (LSTM, autoencoder) for anomaly detection. [27] uses a multi-resolution wavelet-based approach for unsupervised anomaly detection. Stibor et al. [26] describe an immune-system approach to anomaly detection: They tackle a problem prevalent in anomaly detection (and relevant also for the ECG case): Often only nominal data are available during training, nevertheless the model should later detect anomalies as well. This task, known as one-class classification or negative selection, is solved in [26] with an immune-system approach. In our case we have an unknown, small number of anomalies embedded in nominal data. We describe in Sec. 2.5 a statistical test to find anomalies in an unsupervised, threshold-free manner.

Much work is devoted to anomaly detection in ECG readings: Several authors use multi-resolution wavelet-based techniques [23, 27]. A novelty-search approach on ECG data is taken in [18] in order to perform unsupervised anomaly classification. Sivaraks et al. [25] use motif discovery for robust anomaly detection.

The works of Malhotra [19] and Chauhan [4] are closest to our present approach. They describe nicely the general idea of LSTM based prediction and their application to ECG, motor sensor or power-consumption time series. But the big drawback of [19, 4] is that they need a manual and supervised separation into up to six data sets: training, validation, test sets which are further subdivided into nominal & anomalous subsets. This means that for a real-world application the ECG data for a new person would need to undergo an expert anomaly classification prior to any training. This will be highly unpractical in most application scenarios. Our method instead aims at using the whole body of data for a person and train the LSTMs, without the necessity to have supervised anomaly information.

2 Methods

2.1 LSTM for Time Series Prediction

The learning task is formulated as a time series forecasting problem. Hence, we attempt to train a model which is able to predict future values in the time series. The intuition behind this approach is that the usual quasi-periodic patterns in the ECG time series should be predictable with only small errors, while abnormal behavior should lead to large deviations in the predictions. Although the presented methodology is only applied to ECG data in this paper, it is sufficiently general to be applied to other (predictable) time series as well.

Data Preparation Consider a d-dimensional time series of length T. In a first step, it is often recommendable to scale or normalize the individual dimensions of the time series. In our setup, each dimension of each ECG signal are scaled into the range $[-1,1]$. The training and test samples are generated by extracting sub-sequences of suitable length from the original time series. This is done by sliding a window of length W with a lag of 1 over the time series and collecting the windowed data in a tensor $\mathscr{D} \in \mathbb{R}^{T' \times W \times d_{in}}$, where T' is the number of sub-sequences. Usually, one would select all d dimensions of the time series, so that $d_{in} = d$. Then, the first 80% of the samples of $\mathscr{D}$ are selected to form the training data $\mathbf{X}_{\text{train}}$. The remaining 20% are used as a test set $\mathbf{X}_{\text{test}}$ to later compute an unbiased error estimate. While the inputs are d_{in}-dimensional, the output-targets for each time step have the dimension m, since one can select for one time series multiple (m) prediction horizons. Technically, it is also possible to predict several time series dimensions with $d_{out} \leq d$ simultaneously in one model, however, we found in our experiments that the results do not improve in this case (for the investigated ECG time series data). The targets $\vec{y}_t \in \mathbb{R}^m$ are future values of the selected signal at times $t + h_i$ for $i \in \{1, ..., m\}$, where the horizons are specified in $H = (h_1, h_2, \ldots, h_m)$. Since we follow a many-to-many time series prediction approach, where the algorithm performs a prediction at each instance of time t, the tensor containing the target signals has the shape $\mathbb{R}^{T' \times W \times m}$ with $T' = T - W - \max(H) + 1$. T' is the same for the input- and output tensor. As before, the first 80% of the targets are used for training ($\mathbf{Y}_{\text{train}}$) and the remaining targets for the test set ($\mathbf{Y}_{\text{test}}$).

Model Architecture and Training A stacked LSTM architecture [13] with $L = 2$ layers is used to learn the prediction task. Each layer consists of $u = 64$ units. A dense output layer with m units and a linear activation generates the predictions for the specified prediction horizons in H. The net is trained with the sub-sequences of length W taken in mini-batches of size B from the training inputs $\mathbf{X}_{\text{train}}$ and targets $\mathbf{Y}_{\text{train}}$. 10% of the training data are held out for the validation set. The LSTM model is trained by using the Adam optimizer [16] to minimize the mean-squared-error (MSE) loss. Other loss functions, such as log-cosh (logarithm of the hyperbolic cosine) and MAE (mean absolute error) were tested as well and produced similar results for our data. Early stopping is applied to prevent overfitting of the model and to reduce the overall time required for the training process. For this purpose the MSE on the validation set is tracked. For most of the investigated time series, 10-20 epochs are sufficient to reach a minimum of the validation error.

2.2 Modelling the Residuals with a multivariate Gaussian Distribution

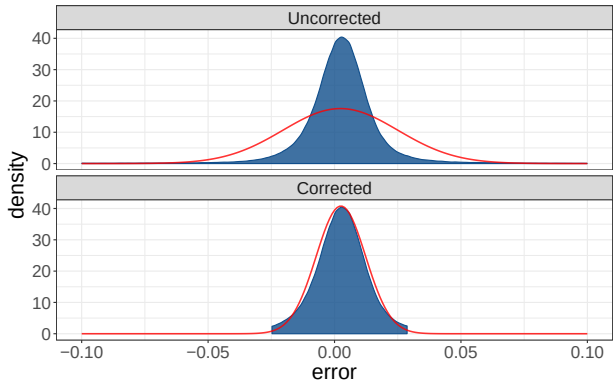

Figure 1: Gaussian fit without and with the removal of outliers in the tails of the error distribution. Exemplarily, this is shown for one dimension of the overall error distribution. The blue curve shows the empiric error distribution. The red curve depicts the estimated Gaussian.

Computing the Prediction Errors for the full Time Series After the LSTM prediction model is trained, the whole time series $\mathbf{X} \in \mathbb{R}^{T \times d_{in}}$ of length T is passed through the model and a tensor $\hat{\mathbf{Y}}$ of shape $\mathbb{R}^{T \times m}$ is predicted. Then, the prediction errors $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ are calculated, where $\mathbf{Y}$ contains the true target values for the horizons in H. Now, each row i in the matrix $\mathbf{E}$ represents an error vector $\vec{e}_i \in \mathbb{R}^m$.

Removing Outliers from the Prediction Errors We noticed in our initial experiments that it was not possible to find good Gaussian fits for the individual dimensions of the prediction errors in $\mathbf{E}$. An example for this is shown in Figure 1, upper part. This was due to the fact that the tails of the error distributions contained many outliers, which significantly distorted the estimated fit. Hence, we decided to remove the outliers in the tails of each dimension (only during the Gaussian modeling phase). We could find good solutions by discarding the upper and lower 3% quantile in each dimension. In our experiments we observed that this approach usually removes slightly more than 20% of the data records.

Estimating a multivariate Gaussian Distribution After removing the outliers from the prediction errors $\mathbf{E}$, the errors are roughly Gaussian distributed and the parameters of a multivariate Gaussian distribution can be estimated. For this purpose the covariance matrix Σ and mean vector $\vec{\mu}$ are computed for the cleaned matrix $\mathbf{E}$. Then, the squared Mahalanobis distance

$$M(\vec{e}_i) = (\vec{e}_i - \vec{\mu})^T \Sigma^{-1} (\vec{e}_i - \vec{\mu}) \tag{1}$$

to the mean vector $\vec{\mu}$ is determined for each error vector $\vec{e}_i$ in $\mathbf{E}$.

2.3 Anomaly Detection

For most data points in the time series the corresponding Mahalanobis distance will be comparably small, since they are located close to the mean of the distribution. On the other side, unusual patterns in the error vectors $\vec{e}_i$ – such as large errors in one or more dimensions – will result in large values in the Mahalanobis distance. Therefore, the Mahalanobis distance can be used as an indicator for anomalous behavior in the time series signal. In our LSTM-AD algorithm, points with a Mahalanobis distance larger than a specified anomaly threshold will be flagged as anomalous. Figure 2 shows exemplarily the Mahalanobis distance over time for a selected ECG signal. Depending on the choice of threshold, more or less points will be classified as anomalous. If the threshold is set too small, the algorithm will likely produce many false detections. If the threshold is chosen too large, many anomalies might be missed. Ideally, a threshold can be found which allows to identify all anomalies without any false detections. However, in practice, one usually has to trade off true and false detections and select the threshold according to the own requirements.

2.4 Window-Based Error Correction

Initially, we could not obtain very good results when running our algorithm on several example ECG time series. The Mahalanobis distance signal was rather noisy and could not be used to distinguish between nominal and abnormal patterns in the data. In Figure 2 (top) this problem is visualized for one example time series. Further investigation showed that the predictions of the LSTM network were good in general, but not sufficiently accurate near the heart beat peaks where the prediction had been slightly shifted (up to 10 time steps) forwards or backwards compared to the real signal. We could identify that the quasi-periodic character of most ECG signals (small but ubiquitous frequency changes in the heart beat) is the main source of this problem. The following solution for this problem is proposed: In order to address the variability in the frequency of the signal, small corrections in the predictions of the individual horizons will be permitted. For each output dimension $k \in \{1, \ldots, m\}$ the target values $y_{t,k}$ are compared to the neighbored predictions $\hat{y} \in \hat{Y}_{t,k}^{(win)}$ and the prediction with the smallest absolute error is effectively taken:

$$\hat{y}_{t,k} \leftarrow \underset{\hat{y} \in \hat{Y}_{t,k}^{(win)}}{\arg\min} |y_{t,k} - \hat{y}| \tag{2}$$

$$\hat{Y}_{t,k}^{(win)} = [\hat{y}_{t-c_k,k}, \ldots, \hat{y}_{t,k}, \ldots, \hat{y}_{t+c_k,k}] \tag{3}$$

We found that reasonable results can be achieved with window parameters c_k up to a length of 10, depending on the prediction horizon h_k:

$$c_k = \min(h_k, 10). \tag{4}$$

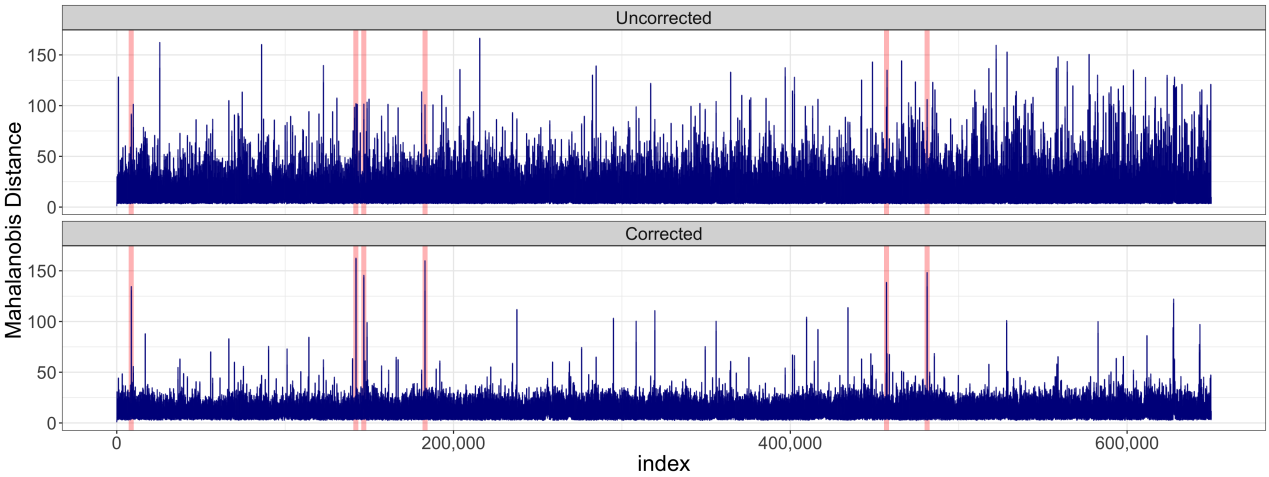

Figure 2: Mahalanobis distance over an ECG time series, before and after the window-based error correction method is applied.

The window-based error correction is applied right after the LSTM prediction of $\hat{\mathbf{Y}}$ and before the prediction errors $\mathbf{E}$ are computed in Sec. 2.2.

Although this approach corrects the predictions of the LSTM network explicitly with the true target values $\mathbf{Y}$ of the time series, it is not supervised in the sense that no anomaly labels are presented to the algorithm at any time of the training and correction process. As will be shown in Sec. 4, we could significantly improve the performance of LSTM-AD utilizing this correction step.

2.5 Threshold-free Anomaly Detection

For determining which points are anomalous, we extend Rosner's outlier test [24] to a multivariate scenario. This means that we want to test the hypothesis:

H_0 : There are no anomalies in the data.

H_a : There are up to ν percent of anomalies in the data.

We assume that the residuals of each output k with $k \in \{1,...,m\}$ are approximately normal distributed. Based on these residuals we iteratively calculate a multivariate normal distribution with mean $\vec{\mu}$ and covariance matrix Σ. In a next step we calculate the differences of an observed vector compared to this underlying multivariate normal distribution. We do so by calculating the squared Mahalanobis distance according to Eq. (1), which can be assumed to be χ^2-distributed with m degrees of freedom.

The proposed algorithm selects the highest value of the Mahalanobis distance $M_j = \max_i M(\vec{e}_i)$ and evaluates whether this value is above the $1-\alpha$ quantile of the χ^2_m-distribution. We refer to this quantile for the ease of notation as $\chi^2_m(1-\alpha)$. If so, the observation is considered to be an anomaly. A row window around the anomaly index is then deleted from the data set and the algorithm proceeds with calculating the multivariate normal distribution based on the reduced data set. The procedure stops when at most ν% of the data were labeled as an anomaly or if $M_j \leq \chi^2_m(1-\alpha)$.

The procedure is summarized in Algorithm 1.

Algorithm 1 Extended Rosner test

procedure TEST($\nu, \alpha, \mathscr{D}$) ▷ ν: maximal percentage of anomalies; $\mathscr{D} \in \mathbb{R}^{n\times m}$: set of residuals, $1-\alpha$ confidence level
  $N \leftarrow \lceil \nu \cdot n \rceil$
  for $j = 1,2,\ldots,N$ **do**
    $\vec{\mu} \leftarrow$ MEAN($\mathscr{D}$), $\Sigma \leftarrow$ COV($\mathscr{D}$)
    $M_j \leftarrow \max_i M(\vec{e}_i)$ ▷ M acc. to Eq. (1) for all $\vec{e}_i \in \mathscr{D}$
    if $M_j \geq \chi^2_m(1-\alpha)$ **then**
      $\mathscr{D} \leftarrow \mathscr{D} \setminus \{$a window around index of $M_j\}$
    end if
  end for
  Set q to the last index where $M_j > \chi^2_m(1-\alpha)$
  return $\{M_1, M_2, \ldots, M_q\}$ and corresp. row indices.
end procedure

3 Experimental Setup

3.1 The MIT-BIH Arrhythmia Database

For our experiments we use the MIT-BIH Arrhythmia database [9, 21, 22], which contains two-channel electrocardiogram (ECG) signals of 48 patients of the Beth Israel Hospital (BIH) Arrhythmia Laboratory. Each recording has a length of approximately half an hour, which corresponds to 650 000 data points each. The two channels recorded are the modified limb lead II (MLII) and a modified lower lead V5. For our LSTM-AD algorithm, both the MLII and V5 signal will be used as inputs of the LSTM model and only the MLII signal is predicted for the specified horizons.[1] The individual signals have a

[1] We also performed experiments where both signals (MLII and V5) are predicted, but could not observe any noticeable improvement.

Table 1: Anomaly types in the 13 ECG signals considered for the experiments. The descriptions are taken from [21]. The second column shows the overall number of the various anomaly types for the 13 considered ECG signals.

Code	#	Description
A	44	Atrial premature beat
V	53	Premature ventricular contraction
!	17	Isolated QRS-like artifact
a	6	Aberrated atrial premature beat
F	2	Fusion of ventricular and normal beat
x	8	Non-conducted P-wave (blocked APC)

quasi-periodic behavior, with differing heart-beat patterns for each subject.

There are a multitude of different events in all ECG time series, which were labelled by human experts. The whole list of heart beat annotations can be viewed at [21]. For this initial investigation which presents first results of our unsupervised approach, we decide to limit ourselves to all time series with 50 or fewer events. Overall, 13 time series will be considered for our experiments. The selected time series contain 130 anomalous events from 6 anomaly classes, which are listed in Table 1. Since only one point is labelled for each anomaly, we place an anomaly window of length 600 around each event, which roughly corresponds to the length of one heart beat before and after the labeled point. A more detailed database description can be found in [22].

3.2 Parameterization of the Algorithms

All algorithms compared in this work require a set of parameters, which are – if not mentioned otherwise – fixed for all experiments. For each algorithm an anomaly threshold can be set, which specifies the sensitivity of the algorithm towards anomalies and which trades off false detections (false positives) and missed anomalies (false negatives). This threshold is usually set according to the requirements of the anomaly detection task – allowing either a higher precision or a higher recall.

LSTM-AD We implemented our proposed algorithm using the Keras framework [5] with a TensorFlow [1] backend. The parameters of the algorithm are summarized in Table 2. Most of the parameters are related to the stacked LSTM network. We did not systematically tune the parameters.

ADVec Twitter's ADVec algorithm [28] is a time series anomaly detection algorithm, which is based on the generalized ESD test and other robust statistical approaches. There are mainly two parameters which have to be provided: The first parameter α represents the level of statistical significance with which to accept or reject anomalies. Although we did not tune the parameter extensively, we found the $\alpha = 0.05$ to deliver the best results. The second parameter maxanoms specifies the maximum number of anomalies that the algorithm will detect as a percentage of the data. This parameter is used as anomaly threshold.

Parameter	**value**	**Parameter**	**value**
H	$(1,3,\ldots,47,49)$	$\mathscr{L}$	MSE
L	3	u	$(64,64)$
B	2048	W	80
optimizer	ADAM	α_{init}	0.001
d_{in}	2	d_{out}	1 (MLII)

Table 2: Summary of the the parameters used for the LSTM anomaly detector.

NuPic Numenta's anomaly detection algorithm [8] has a large set of parameters which have to be set. Although the parameters can be tuned with an internal swarming tool [2], we decided to use the standard parameter settings recommended in [17], since the time-expensive tuning process is not feasible for the data considered in this work. NuPic outputs an anomaly likelihood for each time series point in the interval [0,1], which is suitably thresholded to control the sensitivity of the algorithm.

3.3 Algorithm Evaluation

In order to evaluate and compare the performance of our proposed LSTM-AD algorithm and the two other algorithms, several common performance quantities are used in this paper. Similarly to ordinary classification tasks, a confusion matrix can be constructed for each time series, containing the number of true-positives (TP), false-positives (FP), false-negatives and true-negatives (TN). TP indicates the number of correctly identified anomalies, whereby only one detection within the anomaly window (Sec. 3.1) is counted. All false detections outside any anomaly window are considered as false-positives (FP) and each anomaly window missed by an algorithm will be counted as a false negative (FN). All other points are considered as true-negatives (TN). From the confusion matrix, the additional well-known metrics precision p, recall r (true-positive rate, TPR) can be derived, as well as the false-positive rate (FPR), the positive likelihood ratio (PLR), and the F_1-score, which are defined as:

$$FPR = \frac{FP}{FP+TN}\,, \quad PLR = \frac{TPR}{FPR}\,, \quad F_1 = \frac{2p\cdot r}{p+r}. \quad (5)$$

4 Results & Analysis

Firstly, we confirm that our LSTM models do not overfit, since the training and test set errors are for all time series nearly the same. The median of all such training and test errors is $2.91\cdot 10^{-3}$ and $3.43\cdot 10^{-3}$, respectively. The first 80% of each time series is used as training set and the remaining 20% is used subsequently for the test set.

The anomaly results for the ECG readings are summarized in Table 3 and Table 4. In Table 3, the anomaly

threshold for the Mahalanobis distance was tuned individually to maximize the F_1-score for each ECG signal. As seen in the table, the Mahalanobis distance is generally a good indicator for separating nominal from anomalous behavior in the heartbeat signals, if a suitable threshold is known. For all time series a recall value of 0.5 or larger can be observed and with one exception, also the F_1-score exceeds the value 0.5. On average, a F_1-score of approximately 0.81 can be achieved for all time series. Note that all FPRs are smaller than $3 \cdot 10^{-5}$.

Table 3: Results for all ECG time series with less than 50 anomalies (in total 13). For these results, the anomaly threshold is chosen for each time series individually, so that the F_1-score is maximized.

No.	threshold	TP	FN	FP	Prec	Rec	F_1	FPR $*10^5$	PLR $/10^5$
1	49.31	17	17	14	0.55	0.50	0.52	2.15	0.23
2	71.31	6	0	4	0.60	1.00	0.75	0.62	1.62
4	11.25	1	1	0	1.00	0.50	0.67	0.00	Inf
9	27.12	15	13	3	0.83	0.54	0.65	0.46	1.16
10	14.96	33	7	6	0.85	0.82	0.84	0.92	0.89
11	60.75	1	0	0	1.00	1.00	1.00	0.00	Inf
12	59.18	2	0	0	1.00	1.00	1.00	0.00	Inf
13	116.93	6	0	0	1.00	1.00	1.00	0.00	Inf
15	99.74	5	0	5	0.50	1.00	0.67	0.77	1.30
17	40.02	1	0	0	1.00	1.00	1.00	0.00	Inf
20	75.40	1	0	0	1.00	1.00	1.00	0.00	Inf
21	30.30	1	0	0	1.00	1.00	1.00	0.00	Inf
22	121.42	3	0	7	0.30	1.00	0.46	1.08	0.93
mean	–	7	2	3	0.82	0.87	0.81	0.46	Inf
Σ	–	92	38	39	0.70	0.71	0.70	0.46	1.53

Since the anomaly threshold is used to trade off false-positive and false-negatives (precision and recall), one can vary the threshold in a certain range and collect the results for different values. This is done in Figure 3, which also shows the results for different thresholds for ADVec and NuPic. It has to be noted that ADVec accounts only for fixed-length seasonalities, it is not built for quasi-periodic signals as they occur in ECG readings, so it is quite understandable that it has only low performance here.

Table 4 shows the results which are obtained when no prior knowledge about the anomaly threshold is assumed. Instead, the threshold is calculated automatically and incrementally by our approach from Sec. 2.5 which is inspired by Rosner's ESD test. The average precision and F_1 are lower than in Table 3 since the false positives (FP) are higher.

In Figure 4, two excerpts of time series No. 13 (left) and No. 10 (right) with the detections of our LSTM-AD algorithm, NuPic and ADVec are exemplarily shown. In both examples it can be seen that LSTM-AD detects all indicated anomalies, while NuPic and ADvec only detect two and one anomaly, respectively. Additionally, the other two algorithms produce several false positives.

The importance of our proposed window-based error correction method (Sec. 2.4) is illustrated in Figure 2 for ECG signal No. 13: If no window-based error correction is applied, the obtained Mahalanobis distance cannot be

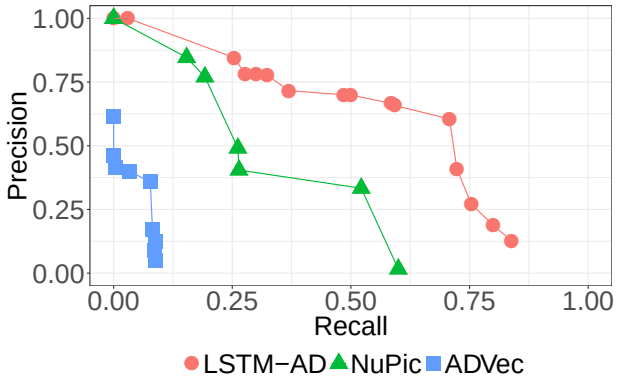

Figure 3: Precision-Recall plot for LSTM-AD, NuPic and ADVec. Precision and recall are computed over the sum of TP, FP, FN of the 13 ECG time series. For all three algorithms, each point is generated by scaling the individual best thresholds up and down by a common factor. For LSTM-AD, the best thresholds are reported in Table 3.

Table 4: Results for the considered 13 ECG signals for the threshold-free method based on Rosner's ESD test.

ECG No.	threshold	TP	FN	FP	Prec	Rec	F_1	FPR $*10^5$	PLR $/10^5$
1	11.30	7	27	3	0.70	0.21	0.32	0.46	0.45
2	11.30	5	1	4	0.56	0.83	0.67	0.62	1.35
4	11.30	2	0	4	0.33	1.00	0.50	0.62	1.62
9	11.30	10	18	0	1.00	0.36	0.53	0.00	Inf
10	11.30	6	34	4	0.60	0.15	0.24	0.62	0.24
11	11.30	1	0	8	0.11	1.00	0.20	1.23	0.81
12	11.30	2	0	5	0.29	1.00	0.44	0.77	1.30
13	11.30	6	0	4	0.60	1.00	0.75	0.62	1.62
15	11.30	4	1	5	0.44	0.80	0.57	0.77	1.04
17	11.30	1	0	7	0.12	1.00	0.22	1.08	0.93
20	11.30	1	0	5	0.17	1.00	0.29	0.77	1.30
21	11.30	1	0	8	0.11	1.00	0.20	1.23	0.81
22	11.30	3	0	7	0.30	1.00	0.46	1.08	0.93
mean	11.30	3	6	4	0.41	0.80	0.41	0.76	Inf
Σ	146.92	49	81	64	0.43	0.38	0.40	0.76	0.50

suitably used to distinguish between nominal and anomalous patterns in the displayed ECG data. Only after applying our approach, a better signal-to-noise ratio is established which allows to perfectly separate the anomalies from the nominal points. For most of the investigated ECG readings we found that the window-based error correction significantly improves the signal-to-noise ratio in the Mahalanobis distance. Table 5 shows: The average F_1-score increases from F_1=0.50 when no window-based error correction is applied to F_1=0.81.

In Table 6, various measures are listed for the individual anomaly classes. The anomaly types a, F and x can all be detected by LSTM-AD. Also for the anomaly class V a high recall can be achieved. However, the two remaining types appear to be hard to detect for our algorithm.

5 Conclusion & Future Work

We have presented a fully unsupervised method to detect anomalies in ECG readings. This method relies on an ac-

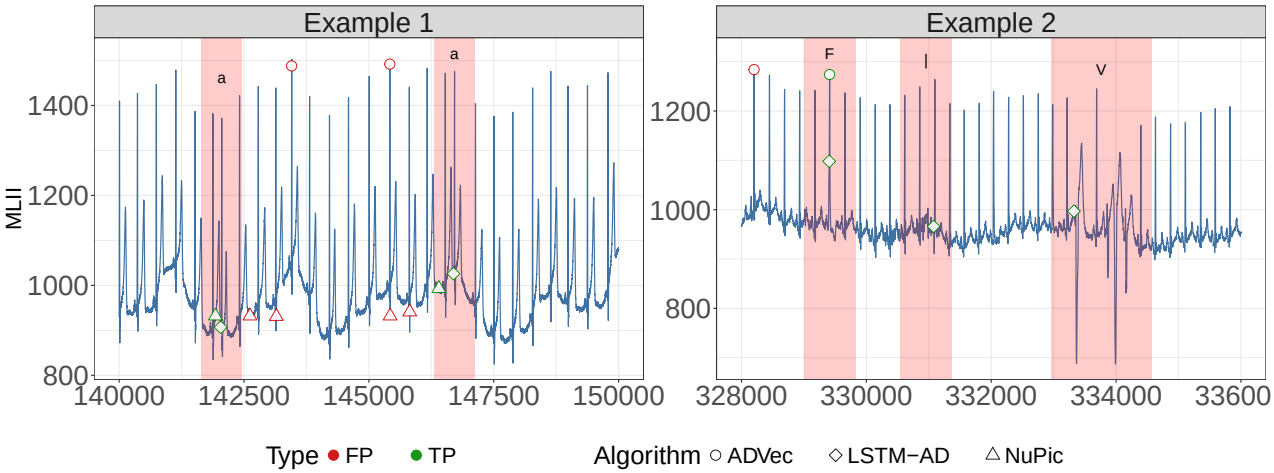

Figure 4: Subsets of two example time series taken from the MIT-ECG data with the anomalies detected by the algorithms LSTM-AD, NuPic and ADVec. The red rectangles in the plot indicate the true anomaly windows. True-positives are indicated by green colors while False-positives are colored red.

Table 5: Comparison of the results for all considered time series, with and without the window-based error correction, as described in Sec. 2.4. The last column F_1(Corr) is copied from Table 3. The remaining columns depict the quantities which are obtained, if no window-based error correction is applied (thresholds chosen such that F_1 is maximized).

ECG No.	threshold	TP	FN	FP	Prec	Rec	F_1	F_1(Corr)
1	20.60	12	22	23	0.34	0.35	0.35	0.52
2	5.83	2	4	2	0.50	0.33	0.40	0.75
4	7.03	1	1	0	1.00	0.50	0.67	0.67
9	16.83	13	15	10	0.57	0.46	0.51	0.65
10	16.21	28	12	2	0.93	0.70	0.80	0.84
11	40.60	1	0	0	1.00	1.00	1.00	1.00
12	28.48	1	1	1	0.50	0.50	0.50	1.00
13	93.83	5	1	130	0.04	0.83	0.07	1.00
15	87.97	2	3	5	0.29	0.40	0.33	0.67
17	35.90	1	0	38	0.03	1.00	0.05	1.00
20	25.10	1	0	1	0.50	1.00	0.67	1.00
21	32.31	1	0	0	1.00	1.00	1.00	1.00
22	77.33	2	1	37	0.05	0.67	0.10	0.46
mean	–	5	4	19	0.52	0.67	0.50	0.81
Σ	–	70	60	249	0.22	0.54	0.31	0.70

Table 6: Various metrics for 5 different anomaly classes. The threshold was tuned individually for each time series by attempting to maximize the F_1-score.

	TP	FN	Prec	Rec	F_1	FPR $*10^5$	PLR $/10^5$
A	23	21	0.65	0.52	0.58	0.14	3.64
V	44	9	0.73	0.83	0.78	0.19	4.36
|	9	8	0.63	0.53	0.58	0.06	8.49
a	6	0	0.79	1.00	0.88	0.02	53.44
F	2	0	0.65	1.00	0.79	0.01	80.16
x	8	0	0.73	1.00	0.85	0.03	29.15

curate LSTM predictor to learn the nominal behavior of the ECG for several prediction horizons. By learning the error distribution between predicted and perceived values, a multivariate normal error model for the nominal data is built. When applying the model, anomalous events have a high probability of being detected through an unusual high Mahalanobis distance.

Our method is unsupervised in the sense that no anomaly class labels are needed for training the algorithm. In fact, it is even not necessary that anomalous events are present at all in the training data, i.e. our algorithm can operate as a one-class classificator. We checked this by repeating the experiment leading to Table 3, but this time removing all data around anomalies during LSTM training. When using the trained model as anomaly detector on all data, it worked as accurate as in Table 3, the mean F_1-score being now $F_1 = 0.83$.

We achieve for the ECG readings these high precision, recall and F_1-values (on average higher than 80%, see Table 3), if we tune the final threshold for the Mahalanobis distance such that F_1 is maximized. Admittedly, this last step is not unsupervised, since we calculate the confusion matrix based on the true anomaly labels.

The alternative unsupervised case based on Rosner's test (Sec. 2.5 and Table 4) is weaker in terms of precision and recall. This may be due to the fact that the current error data do not fulfill the assumption of being normally distributed and therefore also the assumption of a χ^2-distribution is violated. This results in the χ^2-criterion giving no useful thresholds.

It has to be noticed that the measure 'Mahalanobis distance' has the same discriminative power in both cases. It is only that the final threshold is not adjusted optimally for the individual time series in the alternative case. Viewed from the practitioner's perspective, it may be acceptable to start with a non-optimal threshold and adjust it in a human-in-the-loop approach. However, a fully unsupervised high-quality method would be nicer.

We have shown that the window-based error correction is essential to achieve a Mahalanobis distance graph where the anomaly cases clearly stand out (Fig. 2 and Table 5).

Our LSTM-AD algorithm outperformed two state-of-the-art anomaly detection algorithms (NuPic and ADVec) on the investigated ECG readings, achieving a higher precision and recall over a large range of anomaly thresholds.

In this work we have presented first results of an unsupervised anomaly detector suitable for ECG readings or other quasi-periodic signals. The results are encouraging, but there is still room for improvement. Possible future works include: 1) Improving the modeling step such that the nominal error distribution comes closer to a Gaussian shape and hence the nominal Mahalanobis distance closer to a χ^2-distribution. Then the unsupervised extended Rosner test can be expected to work better. 2) To do so, one has to address the problem of non-stationarity in the ECG readings, e. g. by applying suitable preprocessing steps to reduce the effect of signal quality changes. 3) Enrich the model by multi-resolution approaches to span larger prediction horizons on a coarser scale.

References

[1] M. Abadi et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[2] S. Ahmad. Running swarms. `http://nupic.docs.numenta.org/0.6.0/guide-swarming.html`, May 2017.

[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[4] S. Chauhan and L. Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–7. IEEE, 2015.

[5] F. Chollet et al. Keras. `https://keras.io`, 2015.

[6] M. C. Chuah and F. Fu. ECG anomaly detection via time series analysis. In *International Symposium on Parallel and Distributed Processing and Applications*, pages 123–135. Springer, 2007.

[7] M. C. Garcia, M. A. Sanz-Bobi, and J. del Pico. SIMAP: Intelligent system for predictive maintenance: Application to the health condition monitoring of a windturbine gearbox. *Computers in Industry*, 57(6):552–568, 2006.

[8] D. George and J. Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLoS Comput Biol*, 5(10):e1000532, 2009.

[9] A. L. Goldberger et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[10] K. Greff et al. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[11] S.-J. Han and S.-B. Cho. Evolutionary neural networks for anomaly detection based on the behavior of a program. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(3):559–570, June 2005.

[12] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] R. U. Islam, M. S. Hossain, and K. Andersson. A novel anomaly detection algorithm for sensor data under uncertainty. *Soft Computing*, 22(5):1623–1639, 2018.

[15] T. Kieu, B. Yang, and C. S. Jensen. Outlier detection for multidimensional time series using deep neural networks. In *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, pages 125–134. IEEE, 2018.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] A. Lavin and S. S. Ahmad. Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark. In *IEEE Conference on Machine Learning and Applications (ICMLA2015)*, 2015.

[18] A. P. Lemos, C. Tierra-Criollo, and W. Caminhas. ECG anomalies identification using a time series novelty detection technique. In *IV Latin American Congress on Biomedical Engineering 2007, Bioengineering Solutions for Latin America Health*, pages 65–68. Springer, 2007.

[19] P. Malhotra et al. Long short term memory networks for anomaly detection in time series. In *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 89–94, 2015.

[20] P. Malhotra et al. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

[21] G. B. Moody and R. G. Mark. PhysioNet: The MIT-BIH Arrhythmia Database. `https://www.physionet.org/physiobank/database/mitdb/`, 1992.

[22] G. B. Moody and R. G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[23] H. M. Rai, A. Trivedi, and S. Shukla. ECG signal processing for abnormalities detection using multi-resolution wavelet transform and artificial neural network classifier. *Measurement*, 46(9):3238–3246, 2013.

[24] B. Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

[25] H. Sivaraks and C. A. Ratanamahatana. Robust and accurate anomaly detection in ECG artifacts using time series motif discovery. *Computational and mathematical methods in medicine*, 2015.

[26] T. Stibor et al. Is negative selection appropriate for anomaly detection? In *Proceedings of the 7th annual conference on Genetic and evolutionary computation*, pages 321–328. ACM, 2005.

[27] M. Thill, W. Konen, and T. Bäck. Time series anomaly detection with discrete wavelet transforms and maximum likelihood estimation. In O. Valenzuela, I. Rojas, et al., editors, *Intern. Conference on Time Series (ITISE)*, 2017.

[28] O. Vallis, J. Hochenbaum, and A. Kejariwal. A novel tech-

nique for long-term anomaly detection in the cloud. In *6th USENIX Workshop on Hot Topics in Cloud Computing, Philadelphia, PA*, 2014.