FACHHOCHSCHULE NORDWESTSCHWEIZ

MASTER THESIS PROPOSAL

Deep Learning for Anomaly Detection

Author:
Kevin SANER

Supervisor:
Prof. Dr. Thomas HANNE

A thesis proposal submitted in fulfillment of the requirements for the degree of Master of Science in Business Information Systems

in the

School of Business

May 29, 2021

Declaration of Authorship

I, Kevin SANER, declare that this thesis titled, "Deep Learning for Anomaly Detection" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Your brain does not manufacture thoughts. Your thoughts shape neural networks."

Deepak Chopra

FACHHOCHSCHULE NORDWESTSCHWEIZ

Abstract

School of Business

Master of Science in Business Information Systems

Deep Learning for Anomaly Detection

by Kevin SANER

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

Acknowledgements

The acknowledgments and the people to thank go here.

Contents

List of Figures

List of Tables

List of Abbreviations

DNN	**D**eep **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
GRU	**G**ated **R**ecurrent **U**nit Network
LSTM	**L**ong **S**hort **T**erm **M**emory
ROC	**R**eceiver **O**perating **C**haracteristics
AUC	**A**rea **U**nder the **C**urve

For/Dedicated to/To my…

Chapter 1

Introduction

With the rise of the Internet of Things (IoT) and ever more sensors, gadgets and smart devices like smartwatches for fall detection or blood pressure monitoring, or fridges for temperature protective control in use, the amount of available data steadily increases (Alansari et al., 2018). Simultaneously, the possibilities to use the data to draw conclusions increases. This data is generally used to draw conclusions such as failure of a system or a medical issue, such as a heart attack. These events typically occur very rarely (Hauskrecht et al., 2007). However, when the number of instances of each class is approximately equal, most machine learning algorithms function best. Problems occur when the number of instances of one class greatly exceeds the number of instances of the other. This issue is very popular in practice, and it can be observed in a variety of fields such as fraud detection, medical diagnosis, oil spillage detection, facial recognition, and so on (Thabtah et al., 2020). The task of identifying the rare item, event or observation is often referred to as anomaly detection. Typically, the anomalous item translates to problems such as bank fraud or medical problems. Often, the anomaly does not adhere to the common statistical definition of an outlier. Therefore, many outlier detection methods (in particular unsupervised methods) fail on such data (Hodge and Austin, 2004)

A special disciplin in anomaly detection is to find the anomaly in a time series. The anomaly detection problem for time series is usually formulated as finding outlier data points relative to some standard or usual signal. Time series anomaly detection plays a critical role in automated monitoring systems. It is an increasingly important topic today, because of its wider application in the context of the Internet of Things (IoT), especially in industrial environments. The most popular techniques to find the anomalies are:

- Statistical Methods
- Support Vector Machines
- Clustering
- Density-based Techniques
- Neural Networks

Currently Neural Networks are regarded the cutting-edge research. Although first approaches to use artificial neural networks exist since 1969. They are popular in research for only about 15 to 20 years. Neural networks are well-suited for assisting people in solving complex problems in real-world situations. They can learn and model nonlinear and complex relationships between inputs and outputs; make generalizations and inferences; uncover hidden relationships, patterns, and predictions;

and model highly volatile data (such as financial time series data) and variances required to predict rare events (such as fraud detection). Since neural networks do not require time intensive feature engineering, but instead learn the features themselves a lot of time can be saved setting up a model compared to traditional machine learning approaches. At the moment, there is a lot of research focussing on neural networks, which is why scientist generally expect further improvements on this kind of technology. This promising outlook is why this research paper also purely focusses on Neural Networks for anomaly detection. Especially Recurrent Neural Networks and Convolutional Neural Networks are investigated and compared.

1.1 Definitions

Following, the most important terms in the context of anomaly detection using neural networks are elaborated and defined.

1.1.1 Univariate, Bivariate and Multivariate Data

Time series data investigation poses a special disciplin. Generally anomalies in time series are harder to detect for traditional statistical models, since the possibility of long term dependencies exist. Time series data comes in different forms. It is distinguished between univariate, bivariate and multivariate data. Univariate involves the analysis of a single variable while bivariate and multivariate analysis examines two or more variables. One significant difference between time-series and other datasets is that the observations are dependent not only on the components d, but also on the time feature n. Thus, time-series analysis and the statistical methods employed are largely distinct from methods employed for random variables, which assume independence and constant variance of the random variables. Time-series are important to data analysts in a variety of fields such as the economy, healthcare and medical research, trading, engineering, and geophysics. These data are used for forecasting and detecting anomalies.

Univariate Data

There is only one variable in this type of data. Because the information only deals with one variable that changes, univariate data analysis is the simplest type of analysis. It is not concerned with causes or relationships, and the primary goal of the analysis is to describe the data and identify patterns.

Bivariate Data

This type of data involves two different variables. This type of data analysis is concerned with causes and relationships, and the goal is to determine the relationship between the two variables.

Multivariate Data

Multivariate data is defined as data that contains three or more variables. It's similar to bivariate, but there are more dependent variables meaning there is not only one variable that influences the observed beahavior (independent variable) but several. The methods for analyzing this data are determined by the objectives to be met. Regression analysis, path analysis, factor analysis, and multivariate analysis of variance are some of the techniques. Data collected from several sensors installed in a car is an example of a multivariate time-series.

1.1.2 Neural Networks

An Artificial Neural Network (ANN) with several layers between the input and output layers, is known as a Deep Neural Network (DNN). Neural networks come in a variety of shapes and sizes, but they all have the same basic components: neurons, weights and functions. These components work in a similar way as human brains and can be trained just like any other machine learning algorithm.

Neuron

Artificial neurons represent the smallest building blocks of neural networks. A neuron usually receives separately weighted inputs which it sums. The sum is then passed through the activation function to calculate the output of the neuron. When training a neural network, the input weights are adjusted by the optimizer function to improve accuracy of the given task e.g. classification.

Layer

In neural networks three different kind of layers are distinguished. There are input, output and hidden layers. A layer can be described as a collection of neurons. All layers between the input and output layer are called hidden layers. In the input layer data is fed into the neural network. The output of the hidden layer is calculated by taking the weighted sums of input and passing it through the activation function. Typically, a more complex problem requires more hidden layers to accurately calculate the output. In the output layer the final result e.g. a classification is produced. Figure 1.1 how a simple Neural Network with just one hidden layer could look like.

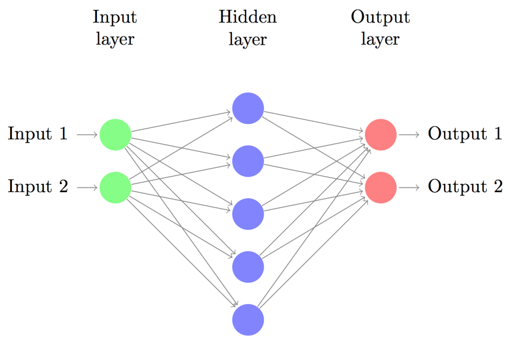


FIGURE 1.1: Input, Hidden and Output Layers (Denny Britz, 2015)

Optimizer Function

Supervised vs. Unsupervised Learning

Supervised learning refers to a task, where a machine learning algorithm learns on data that is labelled. A time series that contains an anomaly that is labelled as such would be an example of such a data set. In contrast, when a machine learning algorithm is applied to an unlabelled data set, it is called unsupervised learning. The goal of unsupervised learning is to find hidden patterns (clusters) in the data set.

1.2 Background

Following, it is explained how different kinds of neural networks work and what they are used for.

1.2.1 Neural Networks for Anomaly Detection

Out of the three most popular neural network architectures, convolutional neural networks (CNN), recurrent neural networks (RNN) and deep neural networks (DNN), RNNs are typically used for anomaly detection in time series [e.g (Malhotra et al., 2015), (**Verner**)]. RNNs have built-in memory and are therefore able to anticipate the next value in a time series based on current and past data. Classic or vanilla RNNs can theoretically keep track of arbitrary long-term dependencies in input sequences. There, however, is a computational issue: when using back-propagation to train a vanilla RNN, the back-propagated gradients can "vanish" or "explode" due to the computations involved in the process, which use finite-precision numbers. Because Long Short Term Memory (LSTM) units allow gradients to flow unchanged, RNNs using LSTM unit or Gated Recurrent units (GRU) partially solve the vanishing gradient problem and therefore drastically improve accuracy (Hochreiter and Schmidhuber, 1997).

Specially to mention in this context are LSTM (Long-Short Term Memory) and GRU (Gated Recurrent Units). Both achieved outstanding performance when used for tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems) (Chung et al., 2014)

LSTM

LSTM was first proposed in 1997 by Schmidhuber and Hochreiter (Hochreiter and Schmidhuber, 1997). The initial version to the LSTM unit consisted of a cells, input and output gates. In 1999, the LSTM architecture was improved by introducing a forget gate and therefore allowing the LSTM to reset its own state (Gers, Schmidhuber, and Cummins, 2000). LSTM is used in a supervised training approach, that means it tries to predict a predefined state taking the past and the current state. If the predicted state differs from the expected state, the weights of the different gates are adjusted using an optimizer algorithm such as gradient descent. Figure 1.2 shows how the gates and the cell are arranged. The cell represents the memory of the LSTM. In simple words, the LSTM works as follows to predict a new value:

1. Forget Gate: Obsolete information is removed from the cell state.
2. Input Gate: New information is added to the cell state
3. Output Gate: The new information and the cell state are added to make the new prediction.
4. The new cell state is propagated to the next LSTM unit

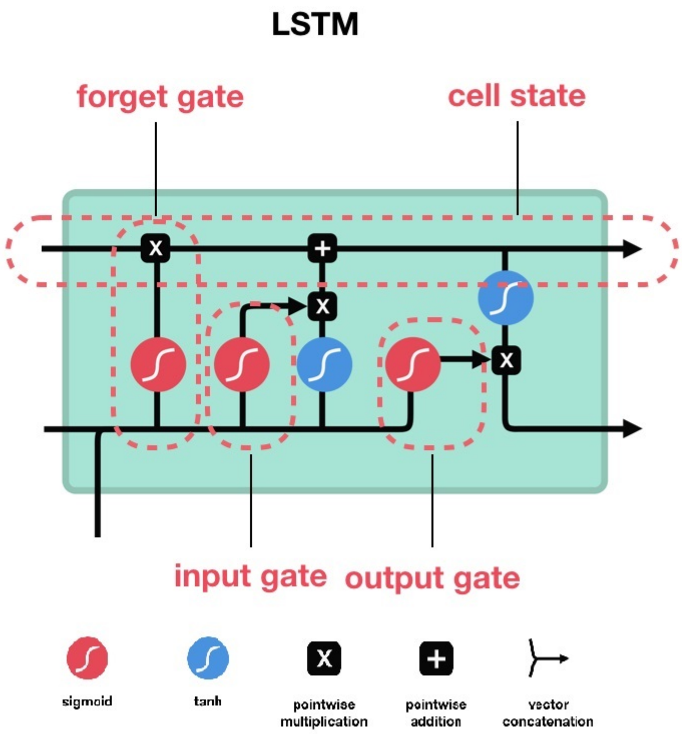


FIGURE 1.2: Gates and Cell of LSTM (Michael Phi, 2018)

CNN

In contrast to RNNs Convolutional Neural Networks are generally used for image classification. CNNs work as feature extractors and are able to recognize patterns . CNNs use layers that are not fully connected, to reduce complexity (compare to 1.2.1). In a CNN, a set number of neurons forms a filter. These filters or kernels are the actual feature extractors. A filter may represent a line or pattern (see Figure 1.3) (LeCun et al., 1998).

FIGURE 1.3: Example of a Filter used in CNN (Rich Stureborg, 2019)

To detect whether, a feature is occurrent in a picture, the filter is gradually moved over the picture in so called strides. In every step (stride) the dot-product between the filter and the part of the picture is calculated. The results of the operations are stored in activation maps. The greater the dot-product the more alike are the filter and the section of the image. Training the network hereby refers to determining the shapes of these filters. Other typical features of a CNN are the pooling layers. The pooling layers reduce the amount of computation necessary. The most commonly used pooling technique is max-pooling and works as shown in Figure 1.4.

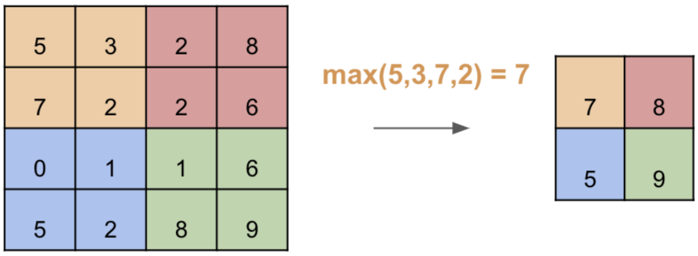


FIGURE 1.4: Example of max-pooling (Rich Stureborg, 2019)

The idea of max-pooling is to only keep the maximum value of an activation map. In the orange region 7 represents the maximum value, so it is kept while the other values are discarded (Rich Stureborg, 2019).

In 2019, Wen and Keys proposed to use CNN also for anomaly detection in time series since it shares many common aspects with image segmentation. A univariate time series is therefore viewed as a one-dimensional image.

1.2.2 Transfer Learning

The reuse of a previously trained model on a new problem is known as transfer learning. It is currently very popular in deep learning because it can train deep neural networks with a relatively small amount of data. This is particularly useful in the field of data science, as most real-world problems do not provide millions of labeled data points to train complex models. In transfer learning, the knowledge of an already trained machine learning model is applied to a different but related problem. For example, if a classifier was trained to predict if an image contains a backpack, the model's experience could be applied to recognize other objects such as sunglasses (Niklas Donges, 2020).

1.3 Problem Statement

Defining a ground truth is one of the most difficult aspects of time series anomaly detection. Determining when anomalous behavior begins and ends in time series is a difficult task, as even human experts are likely to disagree in their assessments. Furthermore, there is the question of what constitutes a useful detection when detecting anomalies in time series. In the past, RNN have successfully been used for anomaly detection (e.g. [(Malhotra et al., 2015); (Fan et al., 2016)]. Therefore, designs for various use cases are well researched. RNN are well suited for the task, however, take a long time to train due to the complexity of how a single unit is designed (see 1.2.1). In comparison CNN are not as complex and therefore, generally take less time to train. However, CNNs are generally used for image recognition and were only very recently used for anomaly detection in time series. It is therefore mostly unknown what designs are applicable for successful anomaly detection in time series data. While RNN are able to deal with multivariate data by design, a classical CNN requires design changes to be able to deal with multivariate data.

Wen and Keys (2019) proposed to use a special kind of U-net, an improved version of a Fully Convolutional Neural Network (Ronneberger, Fischer, and Brox, 2015). Further, a CNN is not capable to analyze streaming data so it relies on segmentation of the data. These data segments are called snapshots. In order to not miss any data points, the frequency of taking these snapshots should be at least as high as the length of snapshot so that every time point is evaluated by the model at least once. However, for better performance it might prove beneficial to use a higher frequency which means every point is evaluated various times by the model (Wen and Keyes, 2019). The proposed design change and the fact that every point is evaluated multiple times, increases complexity and evaluation time and therefore counteract the architectural advantage of CNN compared to a RNN. When designing a neural network many parameters have to be chosen, this applies to both mentioned types of Neural Networks. For example, when designing a CNN, the number of layers, the activation function(s) of a single neuron and the optimizer function have to be chosen. Additionally, when using CNN for time series data the length and frequency of the snapshots have to be determined. Similarly, when designing a RNN also the number of layers and the optimizer function have to be determined. Because the basic building blocks of both networks types are very different it is difficult to fairly compare the complexity of two architecture approaches. Another important parameter which applies to both network types is the number of epochs for which the networks are trained. Through the epochs the training time is determined. In order to compare the two types of neural networks, two networks of similar complexity have to be designed. With equal training time the performance of both can be compared and evaluated. A RNN is therefore only set up as benchmark while the main goal of this research project is to clarify whether CNNs are really useful and propose an advantage over RNNs when applied on time series data for anomaly detection.

1.4 Thesis Statement

Convolutional Neural Networks are superior to Recurrent Neural Networks when looking for anomalies in time series data regarding training time and complexity.

1.4.1 Subquestions

- How does a CNN for univariate and multivariate data need to be designed for successful anomaly detection in time series data?
- What advantages and disadvantages arise when using a CNN compared to a RNN for anomaly detection in univariate and multivariate time series?
- What parameter settings are crucial for a fair performance comparison between RNN and CNN?
- Optional: How does transfer learning affect the performance of CNN compared to RNN in anomaly detection in time series?

1.4.2 Research Objectives

Following the research objectives of this paper are defined.

1. Determine what design changes a CNN requires to detect anomalies in time series data.

2. Determine how the CNN should be designed for the comparison with a RNN
3. State the advantages and disadvantages of the chosen CNN architecture.
4. Define parameters which allow a fair comparison of CNN and RNN

1.4.3 Limitations

Recently there have been approaches that combine CNN and RNN into a hybrid network for tasks such as handwriting recognition or video-based emotion recognition (Dutta et al., 2018) (Fan et al., 2016). However, this paper only compares pure CNN and RNN, and does investigate a hybrid approach.

1.4.4 Significance

Until now, time series data was almost only approached with RNNs. This paper should answer the question whether CNNs propose a valid alternative and even propose some advantages over RNNs. The paper will answer the fundamental question whether research should channel efforts to further investigate CNNs for anomaly detection in time series data or whether no benefits can be discovered and research is better to focus on other areas.

1.4.5 Chapter Overview

Chapter 2

Related Literature

2.1 Anomaly or Outlier?

Generally, there is no agreement on how to distinguish between anomalies and outliers. The following often used citation proves equality of the term outliers and anomalies.

"*Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature.*" - Aggarwal (2013)

By others, outliers are regarded as corruption in data, while anomalies are abnormal points with a particular pattern. In the context of this paper, only the term anomaly is used to refer to such irregular behaviour. It is hereby important to provide a clear definition for the concept of anomaly. This is critical since different meanings of abnormalities necessitate different detection methods. As a result, it is important to identify the key characteristics of anomalies and to use the description to highlight the boundaries. Following, two of the most common definitions of anomalies:

"*Anomalies are patterns in data that do not conform to a well defined notion of normal behavior.*" - Chandola et al. (2009)

Ord (1996), defines anomalies as follows:

"*An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.*"

Anomalies have two major features, according to both of these definitions:

- The anomalies' distribution deviates significantly from the data's overall distribution.
- Standard data points make up the vast majority of the dataset. The anomalies make up a very small portion of the overall dataset.

The development of anomaly detection methods is dependent on these two factors. The second property, in particular, prevents the employment of common classification methods that depend on balanced datasets.

2.1.1 Types of Anomalies

Anomalies come in a variety of shapes and sizes. Anomalies can be divided into three categories:

1. **Point Anomalies** - A point anomaly occurs when a single point deviates dramatically from the rest of the data. A point anomaly in a time seris is, for example, a temperature peak in an otherwise, over time, steady temperature.

2. **Collective Anomalies** - Individual points may or may not be anomalous, but a series of points may be. A time series example could look as follows: A bank customer withdraws $500 from her account per weekday. Although withdrawing $500 every now and then is common for the consumer, a series of withdrawals is unusual.

3. **Contextual Anomalies** - Some points can appear natural in one context but be identified as anomalous in another: In Germany, a daily temperature of 35 degrees Celsius is considered natural in the summer, but the same temperature in the winter is considered unusual.

Knowing ahead of time what kind of anomaly the data might contain helps the data analyst choose the best detection process. Some methods for detecting point anomalies fail to detect collective or contextual anomalies entirely (Braei and Wagner, 2020).

2.2 Anomaly Detection on Univariate Time Series

First, anomaly detection on univariate time series using deep learning approaches is investigated to gain insight into the used architecture and the overall training and detection process.

2.2.1 LSTM

Because of their ability to retain long term memory, Long Short Term Memory (LSTM) networks have been shown to be especially useful for learning sequences containing longer term patterns. Malhotra et al. (2015) model normal behavior with a predictor and then use the prediction errors to classify abnormal behavior. This is especially useful in real-world anomaly detection scenarios where instances of normal behavior are plentiful but instances of anomalous behavior are rare. For prediction, multiple time steps into the future are forecasted to ensure that the networks capture the temporal structure of the chain. As a result, each point in the series has multiple corresponding expected values made at various points in the past, resulting in multiple error values. The likelihood of normal behavior on the test data is calculated using the probability distribution of the errors produced when predicting on normal data. For this approach Malhotra et al. use a deep LSTM neural network. The proposed network architecture with two hidden layers is successfully applied on different univariate time series such Electrocardiograms (ECG), a valve time series and a power demand dataset.

Especially to mention is the training approach used in both papers. To train the LSTM based Recurrent Network the data was divided into four sets. a non-anomalous training set, a non-anomalous validation set, a mixture of anomalous and non-anomalous validation set and test set, also consisting of a anomalous and non-anomalous sequences. This approach was chosen to deal with the rare class problem, which is typical in anomaly detection. Since non-anomalous data is plentiful, the model is trained in an unsupervised fashion, predicting the future but not able to classify if anomalous behaviour was present. The goal of this approach was to establish a baseline of non-anomalous behaviour with the training set. The learned model was then validated and test with anomalous and non-anomalous data to evaluate its performance.

2.2.2 CNN

The use of CNNs for time-series analysis has received interest in recent years. Munir et al. (2019) forecast time-series and identify anomalies based on the prediction error using a CNN architecture called deep-learning based anomaly detection method (DeepAnT). DeepAnT uses an architecture that is divided into two modules. The first module is called "Time Series Predictor". The "Time Series Predictor" consists of a CNN. As the name of the module indicates, the CNN is responsible for predicting the future time stamps of a given time series, whereas the "Anomaly Detector" module is responsible for tagging given data point as anomalous. Figure 2.1 represents the architecture of the CNN-based predictor module. It consists of two convolutional layers, each followed by a max-pooling layer. As the last layer, however, a fully connected layer, where all neurons are connected to all neurons of the previous layer, is used. The last is reponsible for predicting the next time step. The number of output nodes, hereby, corresponds to the number of predicted time steps. One output node means only the next time step into the future is predicted, whereas three output nodes would imply a sequence of three data points are predicted.

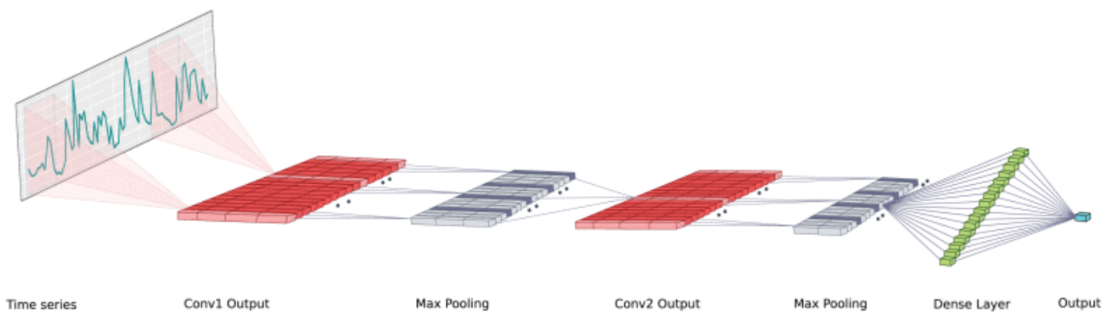


FIGURE 2.1: CNN Architecture for Time Series (Munir et al., 2019)

Once the next time steps are predicted, the values are then passed to the "Anomaly Detector". The detector module calculates the Euclidian distance between the predicted and the actual data point. This measure of discrepancy is used as anomaly score. A high anomaly score indicates a significant anomaly at a given time step. For this module, a threshold based on the time series form must be specified. Most anomaly detection algorithms actually require such a threshold.

2.2.3 Comparison

In an extensive study Braei and Wagner (2020) compared 20 different anomaly detection methods. The anomaly detection methods were divided into the three following categories.

- Statistical Methods
- Classical Machine Learning Methods
- Deep Learning Based Methods (Neural Networks)

Comparison of AUC

The methods were applied on data containing point, collective and contextual anomalies. In order to compare the different approaches, AUC-Values [1] were used. The results showed that the statistical methods generally performed best on point and collective anomalies while deep learning approaches performed rather poorly. On a dataset that contained contextual anomalies, the situation reflected the exact opposite. Deep learning approaches clearly outperformed statistical methods. It was observed that deep learning approaches kept their ability to generalize while statistical methods overfitted on the data (Braei and Wagner, 2020).

Computation Time

The second parameter that was used to compare the different categories was training and inference time. Inference refers to the time used to classify the test data. Compared to statistical methods and classical machine learning, deep learning approaches once again performed rather poorly regarding training time. Looking at inference time deep learning approches generally perform well, outperforming the other two categories. However, there are huge difference within the deep learning approaches. While CNNs have a very low inference time, and outperform most other algorithms, LSTMs have the highest inference time of all tested algorithms (Braei and Wagner, 2020).

2.3 Anomaly Detection on Multivariate Time Series

In this chapter, anomaly detection on multivariate time series is investigated. Once again, CNNs and LSTMs are examined on the state of the art application possibilities for anomaly detection, but also on time series forecasting.

2.3.1 LSTM

First, relevant works that use LSTMs are investigated. LSTMs will be used to establish a benchmark.

Anomaly Detection using LSTM

In his dissertation, Alexander Verner (2019), compared different versions of LSTM Neural Networks against traditional machine learning algorithms such as Support Vector Machines (SVM) and Random Forests. The various algorithms were applied to data of sensors that measure blood glucose levels. To measure the blood glucose levels multiple sensors are used for an accurate result, thus a multivariate time series is the outcome. The data set used in Verner's work did not contain anomalies by default, but the anomalies were embedded into the data set artificially. Since the data was also labelled after the type of anomaly it contained, the learning approach can be classified as supervised. Thus, the goal of the study was to correctly classify which anomaly was present. The first approach used a deep or stacked LSTM, with 10 units in the first layer and 35 in the second. The resulting accuracy of only 0.4 was, however, rather dissappointing. This poor performance was explained with an

[1] An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. The AUC measures the entire two-dimensional area underneath the entire ROC curve. See section 2.5.1

occuring information loss within the deep network. Next, a single layer architecture, with 100 units was proposed. With just a 10% increase the resulting accuracy was still unacceptable. In a third attempt, the neural network was enriched with an embedding layer. Using this technique, the neural network was able to detect the relationship between the frequencies of measurements that were closely positioned in the time-series. It achieved an accuracy of 98%. The final architecture then used bidirectional LSTMs (BLSTMs). According to Graves (**Graves2005**), BLSTM outperform LSTM on supervised labelling tasks. A BLSTM can extract even more information from raw data by taking into account the relationship of each measurement to both previous (past) and subsequent (future) measurements in the input time-series. This enhancement resulted in a 99% accuracy. This impressive accuracy, however, comes at a cost. LSTMs are very time consuming and resource intensive to train. Verner used an Amazon Web Service computing instance with 24 CPUs, 4 GPUs and 128 GB of RAM but the training of an LSTM still took about 10 hours.

Time Series Forecasting using LSTM

LSTMs are normally used to forecast time series such as for example stock prices or air temperature. Following the architecture of a use case applied to a weather time series is described and investigated. The data set contains 14 different variables including air temperature, atmospheric pressure, wind direction etc. The goal in this case is to predict the air temperature. In a first approch, a neural network using Gated Recurrent Units was proposed, in order to keep the required computation effort low. The used NN had 32 GRUs and one linear neuron in the output layer. This simple architecture was already able to outperform the established baseline, which used only used todays temperature to predict the same temperature for the next day. Since the model seemed to underfit, in a second attempt another layer with GRUs was added, to increase network capacity. The second layer was able to improve the results although not significantly. Since, the network does still not overfit the complexity could be further increased at the cost of computation time. However, a different approach was chosen. Similar as in section 2.3.1 a bidirectional approach was considered. The data was therfore reversed and thus processed in antichronological order using the same network architecture as in the previous attempt. The result strongly underperformed, the established baseline indicating that a bidirectional approach would not be able to improve the results achieved with a onedirectional approach. For the chosen example, it is important that data is processed in chronological order (Chollet and Allaire, 2018).

The above described typical use case of an LSTM is relevant because it could be applied in an unsupervised learning approach. Similar as in section 2.2.2, with a traditional LSTM a base could be learned to predict various time steps in the future. Again applying the Euclidean distance, a probability for an occuring anomaly could be calculated.

2.3.2 CNN

In this section CNNs for anomaly detection but also for classification of time series are investigated.

Classification of Time Series Data using CNN

In their research project Zheng et al. (2014) tried to beat the state of the art classification algorithm for time series, which is the k-Nearest Neighbor algorithm(k-NN). k-NN has been empirically shown to be extremely difficult to beat. The typical problem of the k-NN algorithm, however, is its computation time. Zheng et al. proposed to use their own developed architecture, which is called Multi-Channel Deep CNN (MC-DCNN). Each channel, hereby represents a CNN with convolutional and pooling layers. Typically channels in CNN are used to extract features from the different spectra of pictures. A colored picture for example consists of three channels, red, green and blue. Each channel now works as feature extractor on just one color. This feature of CNN is now used in time series classification. Every channel learns features independently using a single dimension of the multivariate time series as input. Another difference to image classification is that multivariate time series classification uses multiple 1D subsequences rather than 2D image pixels as data. Because CNN only learn features, no classification can be done. In order to classify, a CNN architecture is combined with a Multilayer Perceptron (MLP) that uses fully connected layers. Figure 2.2 shows the just described architecture. On the left, the different channels are shown, where every channel takes on its own univariate time series. With the denoted feature maps and pooling layers the features of the time series are learned. In the MLP on the right, finally, the classification of the time series is done. It is important to note that this architecture does not predict the next time steps in the series but instead a given time series is directly classified. The classification is hereby for example used to classify the physical activity depending on the heartrate and is not used for anomaly detection. In order to use the proposed architecture for anomaly detection however, only the output layer has to be changed.

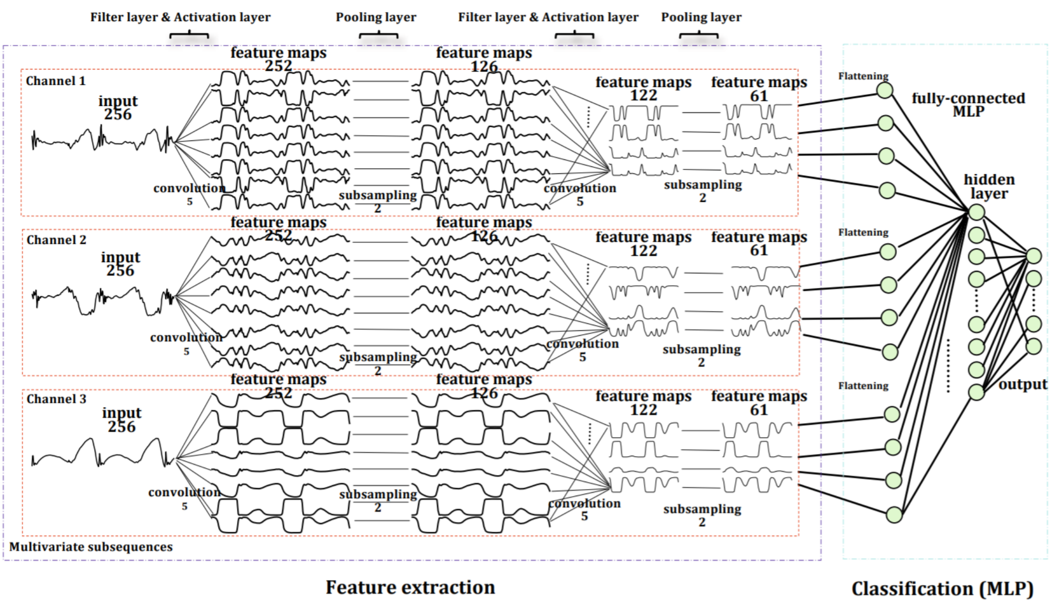


FIGURE 2.2: Architecture of the MC-DCNN (Zheng et al., 2014)

Zheng et al. state that the used architecture was superior to the k-NN algorithm regarding accuracy. Further, the experiments shows that deeper architecture are able to learn more robust high-level features. Further, the MC-DCNN architecture performs much faster than the k-NN algorithm, especially when a large dataset is present.

U-Nets for Anomaly Detection

Wen and Keyes (2019) use a special type CNN architecture to detect anomalies. The architecture used is called U-Net. A U-Net consists of so called encoding and decoding layers. The encoding layers work like a standard CNN whereas the decoding layers are used for upsampling. Upsampling refers to restoring the previously condensed feature maps to its original size. In short, the encoding layers extract the most important features of the time series and decoding layers are using these features to assemble a new time series of the same dimensions as the original one. This encoder-decoder-architecture is often referred to as autoencoder architecture. The main weakness of autoencoders is that in the encoding part through downsampling information is permanently lost. To prevent this information loss, U-Nets introduce so called skip channels also called skip connections. A skip connection, as the name implies, is one that connects an earlier part of the network to a later part of the network and transfers data. The idea is simple: skip channels bring back missing knowledge from some earlier layers so that the network can be properly contextualized. This architecture was proven successful when applied to segmentation of neuronal structures in electron microscopic images in the original paper (Çiçek et al., 2016). In order to handle multivariate time series U-Nets also make use multiple channels. Figure 2.3 shows the architecture of the above described U-Net. The left part of the picture shows the encoding layer with five convolutional layers. The encoding layer is followed by the decoding layer, in the right part of the picture. The decoding layer consists of 4 upsampling layers. This symmetric architecture is completed by the yellow lines, that represent the skip channels.

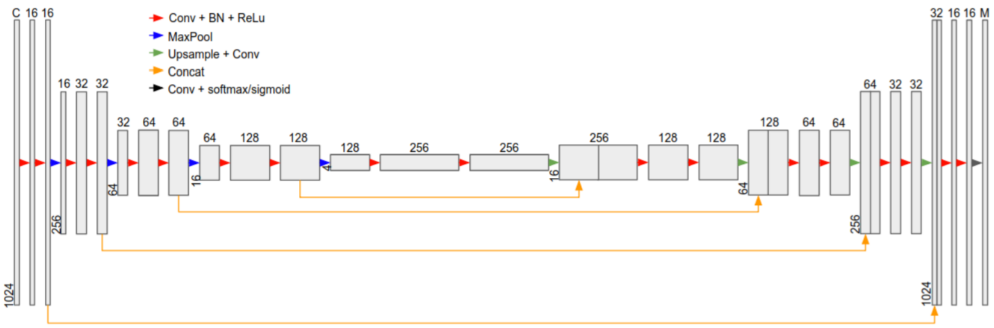


FIGURE 2.3: U-Net Architecture (Wen and Keyes, 2019)

Finally, the model tries to classify what kind of anomaly the multivariate time series contained e.g a seasonal anomaly (contextual) or a point anomaly. Depending on whether the anomaly classes are mutually exclusive or not either a sigmoid or softmax activation is used as last layer activation function.

The proposed U-Net was tested on four scenarios: a univariate task with sufficient data, a multivariate task with sufficient data, a univariate task with insufficient data and transfer learning, and a multivariate task with insufficient data and transfer learning.

For univariate task the dodgers loop sensor data was used. It involves a 28-week time series of traffic on a ramp near Dodger Stadium with a 5-minute frequency. The goal is to spot unusual traffic patterns caused by sporting events. Out of 39 events, only three could not be detected. The missing detection were attributed to missing values in the data set. Missing values apparently also were the reason for some false positives.

For the multivariate task the gasoil heating loop data set was used. It contains 48 simulated control sequences for a gasoil plant heating loop that was hacked at one stage. There are 19 variables in all time series. A multivariate U-Net with 19 channels was trained. Out the 18, in the test present attacks, only one was missed. However, also 3 false alarms (false positives) were reported.

The tasks that included transfer learning are investigated further in section 2.4.1.

2.4 Transfer Learning

The idea behind transfer learning is to transfer knowledge from one domain to another. For example, the knowledge of a Neural Network that classifies pictures of dogs and cats can be used to classify pictures of rabbits. This is possible because all these animals share some features such as eyes or ears. Transfer learning typically uses a, on a large data set, pre-trained network. Some layers are then retrained on the target data set, that typically is much smaller and might not consist of enough data to train a powerful Neural Network.

2.4.1 Transfer Learning with CNN

Since CNN work as feature extractors, the idea when applying transfer learning is to transfer these features to different domains. In the first few layers a CNN typically learns low level features like corners or edges. The features are typically adopted unchanged, meaning the learned weights of the neurons are frozen. The deeper layers of CNN assemble the low level features to high level features such as an ear. Depending on how related the domain of the pretrained and the new network are, deep layers should be retrained to form other features. If a CNN that is trained to recognize animals should be used to classify furniture, the features possibly can not be adopted one-to-one. Instead, the pretrained weights of the neurons are used as initial weights compared to randomly initialised weights when training a CNN from scratch. Using this approach the CNN is fine tuned to new domain using its previous knowlegde. The only part of the neural network that is trained from scratch is the classification layer at the end. This part typically consists of the fully connected layers and the output layer. These parts need to adopt to the new classes (Chollet and Allaire, 2018).

Transfer Learning in Anomaly Detection

Wen and Keyes (2019) differentiated between tranfer learning on univariate and multivariate tasks. For univariate tasks, the architecture, a U-Net, shown in figure 2.3 was used. Only the output layer was changed to match the number of classes in the target task. The pretraining was carried out with various data sets containing different kinds of anomalies such as additive outliers, anomalous temporary changes of volatility, and violations of cyclic patterns as these are the most common anomalies found in univariate time series. In theory, the features learned on these anomalies should be relevant for general anomaly detection. Because features are extracted from both anomalous and nominal segments, the nominal behavior of pretraining series should be diverse. To pretrain the U-Net, synthetic time series were generated. These time series were enriched with the previously mentioned anomalies to create labelled data set for pretraining. Figure 2.4 shows

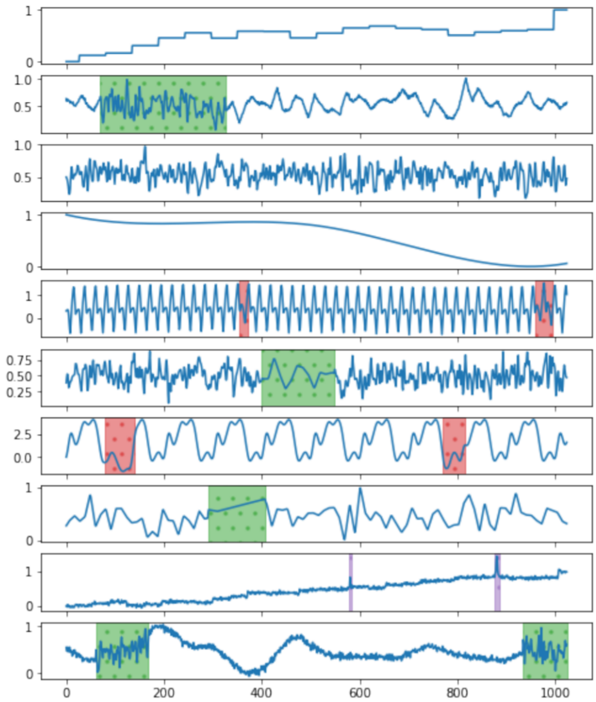


FIGURE 2.4: Synthetic Time Series (Wen and Keyes, 2019)

Important the similarity between pretraining data set and target data set.

2.4.2 Transfer Learning with RNN

Although not used as often, transfer learning can also be used when traning RNN. Similar to CNN, RNNs require large labelled data sets for training. Additionally, training RNNs is computation-expensive due the sequential nature of computations and requires significant tuning efforts. To apply transer learning, RNNs are typically trained on a related source domain, meaning the resulting weigths of the neurons figure as a base. The resulting model is then fine tuned on the target domain. When applying transfer learning using RNN, better accuracy and reduced training efforts on the target domain are the main goals .

Another advantage of learning in such a manner is that the pre-trained network has already learned to extract a rich set of generic features that can then be applied to a wide range of other similar tasks

Transfer Learning for Time Series Analysis

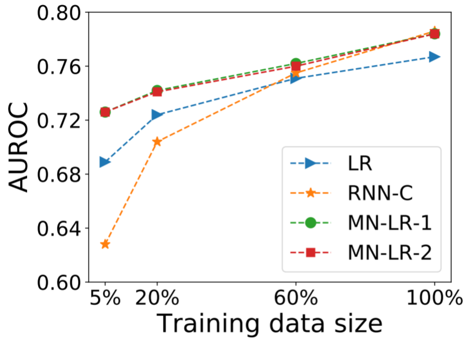
Gupta et al. (2018) used a RNN trained on patient phenotypes [2] to identify previously unseen phenotypes, but also for as well as for an unrelated task of in-hospital mortality. The idea was to train a RNN on physiological parameters to classify phenotypes. A phenotype could be the decrease in blood glucose levels. Detecting and classifying a phenotpye closely resembles the task of anomaly detection.

[2] A clinical phenotype is the presentation of a disease in a given individual

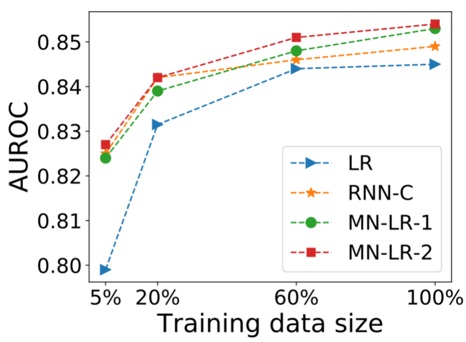
To pretrain, Gupta et al. used time series data such as glucose level or heart rate. The goal was to determine, whether a phenotype was present. Since more than one phenotype can be present, the task can be considered as multi-label classification problem. The neural network used consisted of two layers of GRUs. The output layer was designed to represent all possible phenotypes, as activation function the sigmoid function was used. This network was trained on the labelled data set.

The pretrained neural network was then applied to two different target tasks. The target tasks were, hereby, sligthly reduced in complexity. Instead of a multi-label classification, the problems were formulated as a binary classification problem, determining in case a), whether or not a phenotype was present and in case b) whether a patient would survive based on the time series observations after Intensive Care Unit (ICU) admission.

On the target domain, training with different size data sets was done. The pretrained network was tuned on the target task with at lowest just five percent of data. As a benchmark, to compare the obtained results a logistic regression (LR) and a RNN classifier (RNN-C) were used. The two benchmarks, hereby, only operate on the target domain data set. The performances were compared using AUC (see section 2.5.1) for all used algorithms per data set (Case a & Case b). Figure 2.5 shows the results of the experiments. The names MN-LR-1 and MN-LR-2 hereby refer to the pretrained networks with one respectively two GRU layers.



(a) Phenotyping



(b) In-hospital mortality prediction

FIGURE 2.5: AUC and ROC (AUROC) in Transfer Learning (Gupta et al., 2018)

From the ROC curves of case a) it can be seen that when the training set for the target task is small, the performance gains from transfer learning are greater. Interestingly, in case b) the benchmark RNN is able perform similarly as the two pretrained networks, indicating that better results are obtained the closer related source and target domain are.

2.5 Parameter Settings for fair comparison

In this section it is investigated how CNNs and RNNs can be compared fairly. Since the two's architectures fundamentally differ it is difficult to establish a baseline, where both networks are of similar complexity. In their work Braei and Wagner (2020) chose another approach. Instead of trying to build networks with similar complexity, they fine tuned each evaluated neural network to its full potential and compared their scores as well as their training and inference time. To compare the scores, Braei and Wagner introduced an anomaly threshold (similar as in section 2.2.2). To find out the optimal threshold, it is varied to optimize sensitivity (true positive rate) and specificity (true negative rate). Varying the threshold and drawing the false positives against the true positives results in the so called Receiver Operating Characteristics Curve (ROC-Curve), which is used to find the optimal threshold for a certain model. How the ROC-Curve is used to compared different is explained in the next section.

2.5.1 ROC and AUC

The receiver operating characteristic curve, ROC-Curve, and the associated metric area under the curve (AUC), together also calles AUROC, which is the area under the ROC-Curve, are two metrics that are frequently used. This metric is highly useful, especially for detecting anomalies. The ROC-Curve depicts the relationship between the true positive rate and the false positive rate at various threshold values. The variables needed for the calculations are:

- TP: True Positives
- FN: False Negatives
- FP: False Positives
- TN: True Negatives

The true positive rate is defined as follows:

$$TPR = TP/(TP + FN)$$

The false positive rate is defines as:

$$FPR = FP/(FP + TN)$$

Typically, lowering the classification threshold causes more items to be classified as positive, which increases both False Positives and True Positives. A typical ROC curve is depicted in figure 2.6.

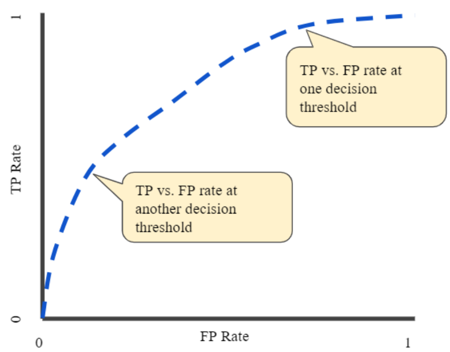


FIGURE 2.6: Example ROC-Curve (Google, 2020)

The area under the above shown curve, however, is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen. The AUC in anomaly detection expressed the likelihood that the measured algorithm assigns a random anomalous point in the time-series a higher anomaly score than a random normal point. As a result, AUC is a good way to compare different anomaly detection methods (Braei and Wagner, 2020).

2.5.2 Computation Time

Another approach to compare the performances is to compare the computation times. To compare the computation times Braei and Wagner (2020) looked at the average training and inference time of traditional machine learning models as well as statistical models and neural networks on univariate time series. Generally, neural networks are expected to invest most of their computation time in training and are able to outperform the traditional machine learning on inference time. However, the practical results showed that Recurrent Neural Networks (LSTM & GRU) not only needed a long time to train but also had a large inference time on certain data sets. With LSTM performing second best on the given data set, behind k-means clustering, it can be concluded that this comes with a tradeoff. On the same data set a CNN was trained. The CNN achieved an AUC-Value of 0.818, compared to 0,84 of the LSTM, which is not much worse than LSTM. Looking at the training and inference time, the CNN is far superior. The LSTM takes about 1000 seconds for training and also inference. The CNN, however, trains for 50 seconds, with an inference time of under one second. Making the CNN the better choice, when relying on small inference times.

2.6 Research Gap

Anomaly detection on time series data has been widely researched. There are comprehensive comparisons of statistical methods, traditional machine learning approaches and neural networks such as the works of Braei and Wagner (2020) or Verner (2019). Braei and Wager, focus thereby only on univariate time series. Complementary, Verner only investigated multivariate time series. In their work, Braei and Wagner could show that CNN and RNN are able to achieve similar performances, while CNNs are superior in computation time. Verner faced a similar drawback in his research, where the training of the RNN took approximately 10 hours.

Since CNNs are only very recently investigated for anomaly detection in time series, Verner did not include them in his research. The approaches presented by Zheng et al. (2014) and Wen and Keyes (2019), however, show promising results, when using CNNs for time series classification respectively anomaly detection.

As research gap it has been identified that there is currently no research on how well CNNs perform when used in the field of anomaly detection. The works mentioned above do not establish a state of the art baseline, to compare the achieved results. It is therefore currently not assessable, how useful CNNs are for anomaly detection.

Further, since the use of neural networks steadily increases, it can be anticipated that transfer learning approaches to save training time and improve accuracy will become more popular. Both papers, than compare the different approaches for anomaly detection do not take advantage of this approach.

Chapter 3

Research Methodology

3.1 Introduction

This section describes which research approach was chosen and why. Further, it elaborates, how the research approach is implemented and what work is done in the corresponding sections.

3.2 Research Design

As the research methodology design science research was chosen. Design science research typically focuses on the creation and performance of artifacts with the explicit goal of improving the artifact's functional performance. It is often applied to artifact categories such as algorithms, human/computer interfaces and design techniques and therefore optimally suits the purpose of this research project.

Design science research knows five process steps. These are:

- Awareness of the Problem
- Suggestion
- Development
- Evaluation
- Conclusion

Following, it is outlined what will be done in the different process steps and how it is going to help answering the research question.

3.2.1 Awareness of the Problem - Literature Review

3.2.2 Suggestion - Data Selection

3.2.3 Suggestion - Design of Experiments

3.2.4 Development - Experiments

3.2.5 Evaluation - Results Analysis

3.2.6 Conclusion - Recommendations

Bibliography

Aggarwal, Charu C. (2013). *Outlier analysis*. Vol. 9781461463, pp. 1–446. ISBN: 9781461463962. DOI: 10.1007/978-1-4614-6396-2.

Alansari, Zainab et al. (2018). "The rise of Internet of Things (IoT) in big healthcare data: Review and open research issues". In: *Advances in Intelligent Systems and Computing*. Vol. 564, pp. 675–685. ISBN: 9789811068744. DOI: 10.1007/978-981-10-6875-1_66.

Braei, Mohammad and Sebastian Wagner (2020). *Anomaly detection in univariate time-series: A survey on the state-of-the-art*. arXiv: 2004.00433.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3, pp. 1–58.

Chollet, François and Joseph J. Allaire (2018). *Deep Learning with R*. Manning Publications, pp. 1–360. ISBN: 9781617295546.

Chung, Junyoung et al. (Dec. 2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *NIPS 2014 Deep Learning and Representation Learning Workshop*. arXiv: 1412.3555. URL: http://arxiv.org/abs/1412.3555.

Çiçek, Özgün et al. (2016). "3D U-net: Learning dense volumetric segmentation from sparse annotation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9901 LNCS. DOI: 10.1007/978-3-319-46723-8_49.

Denny Britz (Sept. 2015). *Implementing a Neural Network from Scratch in Python – An Introduction*. URL: http://www.wildml.com/2015/09/implementing-a-neural-network-from-scratch/ (visited on 04/18/2021).

Dutta, Kartik et al. (2018). "Improving CNN-RNN hybrid networks for handwriting recognition". In: *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*. Vol. 2018-Augus, pp. 80–85. ISBN: 9781538658758. DOI: 10.1109/ICFHR-2018.2018.00023.

Fan, Yin et al. (2016). "Video-Based emotion recognition using CNN-RNN and C3D hybrid networks". In: *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445–450. ISBN: 9781450345569. DOI: 10.1145/2993148.2997632.

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins (2000). "Learning to forget: Continual prediction with LSTM". In: *Neural Computation* 12.10, pp. 2451–2471. ISSN: 08997667. DOI: 10.1162/089976600300015015.

Google (2020). *Classification: ROC Curve and AUC*. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Gupta, Priyanka et al. (2018). *Transfer learning for clinical time series analysis using recurrent neural networks*.

Hauskrecht, Milos et al. (2007). "Evidence-based anomaly detection in clinical domains." In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 319–323. ISSN: 15594076.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.

1997.9.8.1735. URL: https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

Hodge, Victoria J. and Jim Austin (2004). *A survey of outlier detection methodologies*. DOI: 10.1023/B:AIRE.0000045502.10941.a9.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2323. ISSN: 00189219. DOI: 10.1109/5.726791.

Malhotra, Pankaj et al. (2015). "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 2015 - Proceedings*, pp. 89–94. ISBN: 9782875870148.

Michael Phi (Sept. 2018). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. URL: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (visited on 04/18/2021).

Munir, Mohsin et al. (2019). "DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series". In: *IEEE Access* 7, pp. 1991–2005. ISSN: 21693536. DOI: 10.1109/ACCESS.2018.2886457.

Niklas Donges (Sept. 2020). *What is transfer learning? Exploring the popular deep learning approach*. URL: https://builtin.com/data-science/transfer-learning (visited on 04/18/2021).

Ord, Keith (1996). "Outliers in statistical data : V. Barnett and T. Lewis, 1994, 3rd edition, (John Wiley and Sons, Chichester), 584 pp., [UK pound]55.00, ISBN 0-471-93094-6". In: *International Journal of Forecasting* 12.1.

Rich Stureborg (Jan. 2019). *Conv Nets for dummies*. URL: https://towardsdatascience.com/conv-nets-for-dummies-a-bottom-up-approach-c1b754fb14d6 (visited on 04/18/2021).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9351, pp. 234–241. ISBN: 9783319245737. DOI: 10.1007/978-3-319-24574-4_28.

Thabtah, Fadi et al. (2020). "Data imbalance in classification: Experimental evaluation". In: *Information Sciences* 513, pp. 429–441. ISSN: 00200255. DOI: 10.1016/j.ins.2019.11.004.

Verner, Alexander (2019). "LSTM Networks for Detection and Classification of Anomalies in Raw Sensor Data". PhD thesis. Nova Southeastern University, pp. 1–131. URL: https://nsuworks.nova.edu/gscis_etd/1074..

Wen, Tailai and Roy Keyes (2019). *Time Series Anomaly Detection Using Convolutional Neural Networks and Transfer Learning*. arXiv: 1905.13628.

Zheng, Yi et al. (2014). "Time series classification using multi-channels deep convolutional neural networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8485 LNCS, pp. 298–310. ISBN: 9783319080093. DOI: 10.1007/978-3-319-08010-9_33.