Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Computer Science and Engineering

DIPLOMA THESIS ASSIGNMENT

Student: **Marcel Német**

Study programme: Open Informatics
Specialisation: Artificial Intelligence

Title of Diploma Thesis: **Industrial Control System Security Analytics**

Guidelines:

There is a growing demand for security solutions for Industrial Control Systems (ICS). Currently only passive, non-intrusive solutions are accepted by ICS operators – they are afraid of unintended, negative side effects of changes to infrastructure.
The goal is to develop platform for the analysis of ICS network traffic captures. The data to be analyzed is collected in an Industrial Cyber Security Lab.
Tasks:

1. Familiarize with the problem of analyzing ICS protocol data.
2. Identify existing analytics modules that can be used for the analysis of ICS protocols. Develop own analytics modules.
3. Design and implement a platform that allows a user to interactively select the best analytics module or combination of modules for a given task.
4. Test the platform with ICS network data.
5. Assess usability of the analytics platform and the quality of the analytics solution.

Bibliography/Sources:

Rafael Ramos Regis Barbosa - Anomaly Detection in SCADA Systems A Network Based Approach, PhD Thesis, University of Twente, 2014
Abdulmohsen Almalawi, Xinghuo Yu, Zahir Tari, Adil Fahad, Ibrahim Khalil - An Unsupervised Anomaly-Based Detection Approach for Integrity Attacks on SCADA Systems. Computers & Security, 2014
Yang Zhang, Junliang Chen - Wide-area SCADA system with distributed security framework. Journal of Communications and Networks, 2012
Markus Stolze, René Pawlitzek, Andreas Wespi - Visual Problem-Solving Support for New Event Triage in Centralized Network Security Monitoring: Challenges, Tools and Benefits. IMF, 2003
Hervé Debar, Andreas Wespi - Aggregation and Correlation of Intrusion-Detection Alerts. Recent Advances in Intrusion Detection, 2001
Hervé Debar, Marc Dacier, Andreas Wespi - Towards a taxonomy of intrusion-detection systems. Computer Networks, 1999

Diploma Thesis Supervisor: Dr. Andreas Wespi

Valid until the end of the winter semester of academic year 2016/2017

doc. Ing. Filip Železný, Ph.D.
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, October 8, 2015

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science and Engineering

Master Thesis

Industrial Control System Security Analytics

Marcel Német

Supervisor: Dr. Andreas Wespi

Study Programme: Open Informatics

Field of Study: Artificial Intelligence

January 9, 2017

Acknowledgements

I would like to acknowledge the guidance, valuable advice and support provided by Andreas Wespi, Anton Beitler and Marc Ph. Stoecklin from IBM Research.

Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used as stated in document "Metodický pokyn o dodržování etických principů při přípravě vysokoškolských závěrečných prací".

In Zurich on January 9, 2017 .....

Abstract

Industrial Control Systems (ICS) are important for functioning of many critical facilities such as power plants, water treatment facilities or gas pipelines. Although security of such systems deserves attention, application of thorough security intelligence approaches to ICS is not a standard practice. Examples such as the *slammer* worm infection at US Davis-Besse nuclear plant, or the *Struxnet* ICS attack on nuclear centrifuges in Iran show the significance of the security threats in ICS. New security methods capable of better ICS protection are needed to prevent potential damages. ICS operators are afraid of system disruptions and require that the security measures are unobtrusive to the system. Taking concerns of operators in mind, analysis of passively collected network data and detection of intrusions in the collected data is an acceptable method for achieving improved ICS security. Behavior based anomaly detection algorithms for ICS are a viable solution. Such algorithms need to be configured properly to perform well. This thesis proposes an *assistant platform* for interactive configuration, evaluation and comparison of anomaly detection modules. The result is a functioning product that applies techniques of parameter configuration, data labeling, algorithm results evaluation as well as navigating and filtering of the results in an interactive way. The proposed solution allows users to select anomaly detection module and parameter sets that fit their labeling of anomalies and preferences for balancing precision and recall. The thesis discusses features and design of such a platform for an ICS environment. The thesis also presents results of a user testing conducted with five participants in which the users work with the platform to compare performance of anomaly detection modules developed by IBM Research on a data collected in an Industrial Cyber Security Lab.

Keywords: Industrial Control System, SCADA, Anomaly Detection

Abstrakt

Priemyselné kontrolné systémy majú dôležitú úlohu v mnohých nenahraditeľných systémoch ako sú napríklad elektrárne, čističky vôd alebo ropovody. Bezpečnosť týchto systémov si nepochybne zaslúži pozornosť, no nasadzovanie pokročilých bezpečnostných metód nie je bežnou praxou. Príklady ako útok "slammer" červom na jadrovú elektráreň David-Besse alebo "Struxnet" útok na jadrové centrifúgy v Iráne dokazujú vážnosť bezpečnostných hrozieb v kontrolných systémoch. Na predídenie potenciálnych škôd sú potrebné nové bezpečnostné metódy. Operátori kontrolných systémov sa obávajú narušení zabehnutých systémov. Akceptujú iba metódy, ktoré neohrozia systém. Detekcia útokov v pasívne zachytených dátach je akceptovanou možnosťou. Behaviorálne systémy na detekciu anomálií sú možným riešením. Aby plnili svoju úlohu, takéto algoritmy musia byť správne nakonfigurované. Táto práca prezentuje asistenčnú platformu ktorá umožňuje interaktívnu konfigurácie, ohodnotenie a porovnanie výkonu modulov na detekciu anomálií. Asistenčnú platformu sme otestovali s modulmi vyvinutými v IBM Research na dátach zachytených v priemyselnom bezpečnostnom laboratóriu vytvorenom nadnárodnou spoločnosťou na generovanie a distribúciu energie. Kvalita asistenčnej platformy bola ohodnotená na základe testovania s užívateľmi.

Názov práce v Českom jazyku: Bezpečnostní analýza průmyslového kontrolního systému

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Industrial Control Systems (ICS) are important for functioning of many critical facilities. Common types of ICS include Supervisory Control and Data Acquisition (SCADA) systems, Process Control Systems (PCS) and Distributed Control systems (DCS) [1]. Power plants, water treatment facilities, dams, oil refineries, gas pipelines, agricultural sites and other infrastructures use SCADA, PCS and DCS to monitor, manage and control physical processes.

Although security of such systems deserves attention, application of thorough security intelligence approaches to ICS is not a standard practice. The two main forms of protection that SCADA vendors and operators use when protecting SCADA systems are reliant on *air gap* (a physical isolation of SCADA network from other networks) and *security through obscurity* (concealment of information about the SCADA devices) [2]. With change of trends in industries, the mentioned forms of protection cease to be sufficient. Industries are interconnecting their SCADA systems with intranet and internet networks. The air gap should be replaced by a logical gap (a firewall) to maintain security [3] [4]. However, this is not always the case. The United States Industrial Control System Cyber Emergency Response Team listed approximately 7200 ICS devices directly reachable from the internet in their 2012 report [5]. The other trend and potential security liability is the use of Low Cost Commercial Off-The-Shelf (COTS) devices by operators [6]. Potential attackers can obtain and study COTS devices. Therefore, operators can neither rely on the security through obscurity.

The Ponemon Institute conducted a survey in 2011 with experienced IT security practitioners from utilities and energy companies [7]. Only 9 percent of 291 questioned specialists believe that their organization's security initiative is very effective in providing actionable intelligence (e.g. real-time alerts and threat analysis) about potential and actual exploits on their systems. Examples such as the *slammer* worm infection at US Davis-Besse nuclear plant [8], or the

Struxnet ICS attack on nuclear centrifuges in Iran [9] show the significance of the security threats in ICS.

Considering the above mentioned, new security methods capable of better ICS protection are needed to prevent potential damages. ICS operators are afraid of system disruptions and require that the security measures are unobtrusive to the system. Taking concerns of operators in mind, analysis of passively collected network data and detection of intrusions in the collected data is an acceptable method for achieving improved ICS security.

As a result of collaboration of IBM with a power generation and distribution company, we were able to explore data from datasets collected in an Industrial Cyber Security Lab, the first of its kind. The Industrial Cyber Security Lab created by the power generation and distribution company allows interaction with SCADA systems and contains all hardware and software components of a real hydroelectric power plant. Open Platform Communications (OPC) standard is used in many SCADA systems, including the Industrial Cyber Security Lab, to ensure the interoperability among devices from multiple vendors.

Various signature based and behavior based anomaly detection approaches [10, 11] for safeguarding SCADA systems have been explored in the past [12]. Some approaches concentrate on Modbus protocol [13, 14, 15, 16] or whitelisting network traffic aggregated over period of time (Netflow) [17].

In [18], we compared performance of traditional IT monitoring mechanisms with in-depth analysis of OPC packets on three intrusion scenarios. The results show that traditional methods are not sufficient and in-depth OPC packet analysis is required to recognize attacks in all presented scenarios. Answering the need for further OPC analysis, IBM had developed an OPC packet inspector and an analysis and forensics platform for the exploration of OPC event traces. The platform provides unique insights about a running OPC network environment and allows detecting types of anomalies that would have been missed when using Netflow only. Among other features the platform runs behavior-based anomaly detection algorithm modules specifically developed for detecting anomalies in time series from OPC protocol data. One of the modules uses a Windowed Growing Neural Gas [19] algorithm to detect anomalies. Another module uses technique based on sliding window regression forecasting using exponential smoothing implemented on the 'R' [20] statistical computing platform.

Behavior-based systems require tuning in order to be effective in the deployment environment. Due to the lack of good test data from ICS and SCADA systems, it is difficult to estimate

how existing or newly developed anomaly detection algorithms and parameters will perform when deployed on site.

An interactive system that would help operators label anomalies, as well as evaluate and compare performance of anomaly detection modules based on provided labeling would help to better understand capabilities of anomaly detection modules and select appropriate module and parameter set to analyze behaviors of devices in the ICS system. Integrating such an assistant system into the analysis and forensics platform would help ICS operators and security consultants tune detection modules more quickly.

1.2 Aim and Hypothesis

The aim of this thesis is to develop an interactive system that assists the human operator in tuning behavior-based security systems. In the rest of this thesis I refer to such a system as *assistant platform.* Users of such an assistant platform should be able to select anomaly detection modules and parameter sets that they wish to test and compare. They should be able to specify which data is to be used for training and test of the algorithms and to provide expertise on which behavior patterns should be detected as anomaly and which not.

The implemented assistant platform should allow users to significantly reduce time and efforts needed to shortlist anomaly detection modules and parameter sets that provide results similar to an anomaly annotation that they create and better understand how different anomaly modules and parameter sets compare.

Hyper-parameter optimization methods for tuning parameters of machine learning algorithms based on an objective function (e.g. area under a ROC curve) exist [21]. However, the focus of the thesis is rather to create a broader system that will provide more interactivity via understandable user-interface and give users options to explore and compare results of anomaly detection modules.

1.3 Structure

The thesis is structured as follows. Chapter 2 provides a background about SCADA systems, existing analysis and forensics platform and presents methods for evaluating anomaly detection algorithms. Chapter 3 lists partial problems that need to be solved and requirements for

the solutions. Chapter 4 presents a design of an assistant platform. Chapter 5 presents components and user interface of the implemented platform. Chapter 6 discusses set-up and results of evaluation with testing. Chapter 7 concludes this thesis.

Chapter 2

Background

2.1 SCADA systems

Supervisory Control and Data Acquisition (SCADA) systems are a type of Industrial Control Systems (ICS). The architectures of SCADA systems vary across facilities but some common components can be identified. Field devices (sensors or actuators) measure or control physical properties. Examples of field devices are valve or water level sensors. Remote Terminal Units (RTUs) provide an interface to control and read values from field devices. Small embedded devices called Programmable Logic Controllers (PLCs) are often used instead of RTUs. In power systems, PLCs can be referred to as Intelligent Electronic Devices (IEDs) [2]. Master Terminal Unit (MTU) polls the RTUs repeatedly to collect measured data. Human-Machine interfaces provide operators with access to the data collected by the MTU. Field devices together with RTUs are referred to as a field network, while MTU and HMIs reside in the control room (control network). Figure 2.1 shows an example of a simple SCADA architecture.

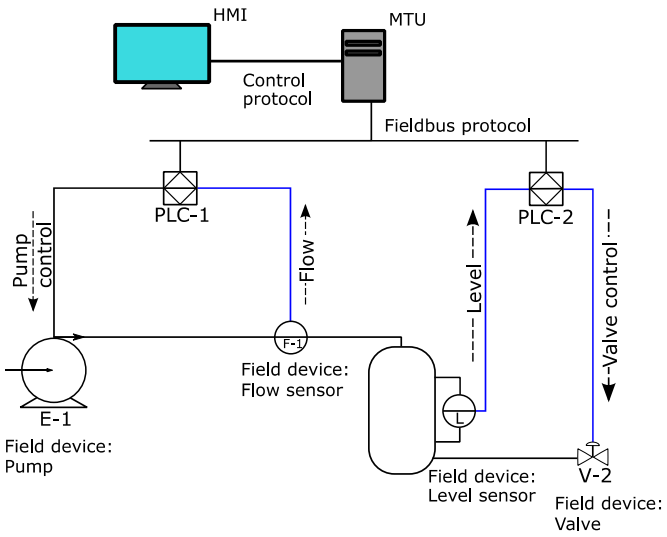

Figure 2.1: Example of SCADA system architecture

Variety of the communication protocols are employed in the SCADA systems. RTUs and PLCs exchange messages with MTU using so called fieldbus protocols. Fieldbus protocols can be SCADA-vendor specific (e.g. RP-570 [22] or Profibus [23]) or open-standard – e.g. Modbus (originally proprietary but made into open-standard) [24], Distributed Network Protocol 3 (DNP3) [25] or IEC 60870.

Open Platform Communications (OPC) protocol is widely used in SCADA systems to ensure seamless flow of information among devices from multiple vendors [26]. OPC was first released in 1996 under the name OLE for Process Control, OLE standing for Object Linking and Embedding, but was renamed in 2011. OPC abstracts vendor specific fieldbus protocols (e.g. Modbus or Profibus) into a standardized interface. HMI/SCADA systems can then send generic read and write requests to OPC servers, which take care of converting them to the vendor specific requests.

2.2 Environment and Data

As a result of the collaboration of IBM Research Zurich with a power generation and distribution company, we have access to industrial environments where we can interact with SCADA systems and capture the data from such systems. A citation from [18] explains that the available environments are: *"(1) an ICS simulation (ICSSIM) environment consisting of a setup of HMI/SCADA, process control, and RTU systems in a setup based on virtual machines and (2) a full-scale cyber security testing laboratory (CYBERLAB) consisting of all hardware and software components of a real hydroelectric power plant.*"

I use data captured in the mentioned Industrial Cyber Security Lab (CYBERLAB) environment as a basis for development and testing of the assistant platform. The dataset is a result of a full network packet capture that IBM obtained using the tcpdump [27]. It is further processed by IBM's software to extract OPC event traces from raw network packet captures. The OPC event traces can be represented as a time series of values written to field devices or read from field devices. The assistant platform as well as the anomaly detection modules included in IBM's analysis and forensics platform are designed to work with the time series data from the OPC event traces.

An important characteristic of the collected data is that the times when the time series values are recorded are not evenly spaced. In other words, it is not to be assumed that the time difference between two consecutive values in the time series is always the same. Taking this in mind, *time series* can be defined as:

Definition 2.1 (Time Series) A time series is a sequence of data point values measured at certain times and ordered by time. It is denoted as $\boldsymbol{X} = \{\{\boldsymbol{x_1}, \boldsymbol{y_1}\}, \{\boldsymbol{x_2}, \boldsymbol{y_2}\}, \ldots, \{\boldsymbol{x_n}, \boldsymbol{y_n}\}\}$, where a value $\boldsymbol{y_i}$ (a real number) was recorded at a time $\boldsymbol{x_i}$.

In the explored *CYBERLAB* environment, as well as in most other SCADA systems, the normal behavior differs for individual field devices and across industrial facilities. Due to a low amount of openly available test data, precise characterization of anomalies in SCADA systems does on exist. Hence, signature based systems are outnumbered by unsupervised anomaly detection systems. Thus, the assistant platform will rely on the expertise of the ICS operators and security consultants and allow them to label the data based on the experience with their SCADA system. Considering their annotation of the data, the assistant platform can evaluate the performance of the anomaly detection modules.

2.3 Current System and Anomaly Detection Modules

2.3.1 Existing Platform

The IBM's analysis and forensics platform currently contains various modules. Two anomaly detection modules and an OPC Explorer module are of importance for this project.

OPC Explorer module provides an API (Application Programming Interface) to query time series data extracted from the OPC packets for a desired field device. It also provides a web user interface to explore data recorded from devices. Figure 2.2 shows the interface of OPC Explorer Web UI.

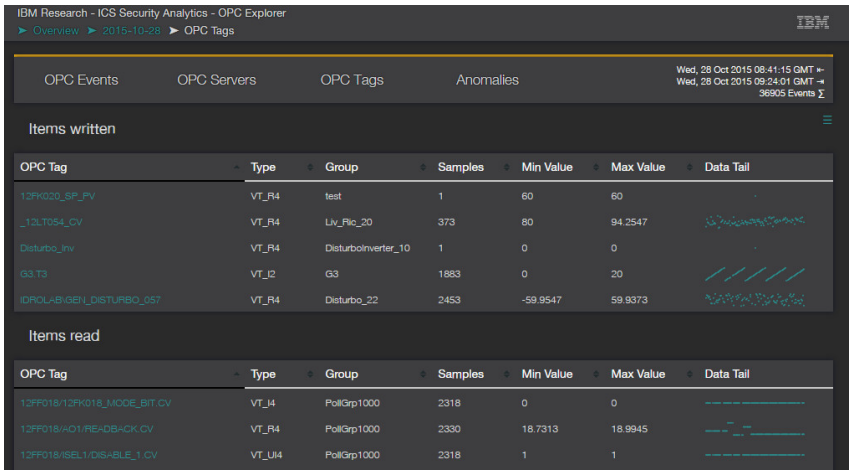

Figure 2.2: User interface of OPC Explorer

During the training phase, the anomaly detection modules learn normal behavior of the system using the training data. After this so called training interval the algorithms are able to compute anomaly likelihood scores for the previously unseen data. The output of anomaly detection algorithms is standardized so they can be compared among each other.

I refer to the output of anomaly detection modules as to scores:

Definition 2.2 (Scores) Scores are a sequence of values, each reporting anomaly likelihood for a time interval. It can be denoted as $\boldsymbol{s} = \{\{\boldsymbol{b_1}, \boldsymbol{e_1}, \boldsymbol{y_1}\}, \{\boldsymbol{b_2}, \boldsymbol{e_2}, \boldsymbol{y_2}\}, \ldots, \{\boldsymbol{b_n}, \boldsymbol{e_n}, \boldsymbol{y_n}\}\}$, where a real number $\boldsymbol{y_i}$ represents reported anomaly likelihood recorded for a time interval beginning at time $\boldsymbol{b_i}$ and ending at time $\boldsymbol{e_i}$.

Anomaly detection modules can be executed by submitting a job that contains: time interval that should be used for training of the normal behavior, time interval that should be analyzed (test interval), identifier of a device from which the analyzed time series comes from and set of algorithm parameters.

The algorithm modules download the data for the training and test from the OPC Explorer API. The following subsections describe the anomaly detection modules.

2.3.2 Windowed Growing Neural Gas

Windowed Growing Neural Gas (Wgng) [19] is a variant of the Growing Neural Gas (GNG) algorithm [28] that uses a sliding window over time to generate frames to be analyzed. The GNG is an alternative to Self-Organizing Maps (SOM) [28] but does not need to be provided a number of neurons in advance.

The Wgng splits temporal streams of data (e.g. time series) to produce frames. Based on a distance function, frames are assigned to neurons of GNG. The algorithm creates and deletes neurons to accurately represent commonly seen frames. Table 2.1 lists parameters and constraints for the anomaly detection module based on the Wgng algorithm.

Parameter	Name	Description	Constraint
Splitter parameters			
w	Window Size	Size of the sliding window (in time units)	$0 < w < a$
h	Window Hop	Size of the window hop (in time units)	$w/2 \leq h \leq w$
Neural network parameters			
a	Maximum Edge Age	Maximum history length (maximum age of edges) in term of time units.	a*250*60 > w
m	Maximum Neuron Number	The maximum number of natural neurons to spawn.	$m \geq 3$
k	Distance threshold	Threshold above which non-natural neurons will be spawned, in terms of factor of noise standard deviation.	$k > 1$
t_1	Neuron Memory	Number of historical frames to keep for a neuron (seeds).	
t_2	Edge Memory	Number of historical frames to keep for an edge (hist).	$t_2 > 1$
Alpha	Spawn Error Reduction	Reduction factor of error when spawning a natural neuron	$0 <$ alpha < 1
Emc	Error Minimum Count	Error Minimum Count after which neurons are considered as having a good definition of their error standard deviation	emc > 1
Periodicity checker parameters			
Beta	Agility	Defining the importance of the present over the past when updating the mean and variance.	$0 <$ beta < 1
P	Periodicity Threshold	Threshold on the Gaussian kernel under which period anomalies are returned.	$0 \leq p < 1$
pmc	Periodic Minimum Count	Periodic Minimum Count after which neurons occurrences will be checked for periodicity	pmc > 1

Table 2.1: Parameters and constraints of the Wgng module

2.3.3 A-node

The A-node anomaly detection module uses a technique based on sliding window regression forecasting. It uses an exponential smoothing implemented on the 'R' [20] statistical computing platform. The algorithm first extracts the sequence of inter-arrival times and treats it as a separate time series. Both time series are segmented (split) into windows. Two metrics are calculated per each window: mean and standard deviation. This gives rise to a total of four sequences of training data. The same is done for the sample data (the newest data chunk in the time series which is being analyzed) set which consists of a single window.

Two anomaly detection algorithms are then applied on each of the four sequences. The first algorithm is an outlier detection algorithm: For both metrics the expected value is calculated based on the training data. The metrics of the sample data set are compared to the expected value.

The second algorithm is the change point detection: An ETS forecasting method from the R [20] forecasting package is applied on the training sequences. The results of forecasting function are upper and lower bounds of the forecast for each given confidence level. The actual value of the sample data set is compared to the given bounds. This creates a total of eight anomaly scores which are treated as a vector whose length is the resulting anomaly score.

Parameter	Name	Description	Constraint
w	Window Size	Size of the sliding window (in time units)	
h	Window Hop	Size of the window hop (in time units)	
t	Maximum Training intervals		
p_1	Primary Confidence	Primary confidence level for prediction.	$p_1 < p_2$
p_2	Secondary Confidence	Secondary confidence level for prediction	$p_2 > p_1$

Table 2.2: Parameters and constraints of the A-node module

2.4 Evaluating Anomaly Detection Algorithms

This section discusses methods for evaluating anomaly detection algorithm outputs.

2.4.1 Anomaly Detection Algorithms Output Types

As mentioned in Section 2.3.1, the output of the anomaly detection modules in the current platform are scores. The other common type of output that anomaly detection algorithms might use are labels. In contrast to scores, labels classify data points only as *anomalous* or *benign.*

It is possible to convert one format to the other. Scores can be converted to labels by selecting a threshold value. Every score that reports a value equal or greater than a threshold represents an anomaly. Labels can be converted to numerical values by representing benign behavior with zero and anomalous behavior with 1.

2.4.2 Evaluation Metrics for Anomaly Detection Algorithms

When using anomaly detection algorithms with scores output, one must select a threshold which determines what is marked as an anomaly and what is still a normal behavior. Usually this leads to a tradeoff between the number of detected anomalies and number of false positives (normal behavior labeled as anomaly). By setting a low threshold, more anomalies will be detected but normal behavior might be marked as an anomaly more often. Pushing threshold higher means less false positives but also an increased possibility of missing some anomalies.

Commonly used evaluation metrics for anomaly detection algorithms are Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves [29].

ROC curves display false positive rates (FPR) on the horizontal axis and true positive rates (TPR) on vertical axis. These rates are defined as:

Definition 2.3 (False positive rate)

$$FPR = \frac{FP}{N}$$

where FP stands for number of false positives (normal behavior marked as anomaly) and N stands for negatives (total number of normal behavior data points).

Definition 2.4 (True positive rate)

$$TPR = \frac{TP}{P}$$

where TP stands for number of true positives (correctly marked anomalies) and P stands for positives (total number of data points marked as anomaly).

Precision recall curves display recall on horizontal axis and precision on vertical axis. These metrics are defined as follows:

Definition 2.5 (Precision)

$$Prec = \frac{TP}{TP + FP}$$

Definition 2.6 (Recall) Recall is just a different name for true positive rate:

$$Rec = \frac{TP}{P}$$

ROC and PR curves can graphically represent the quality of the algorithm output and allow us to compare outputs of multiple algorithms and thresholds in one picture. An example of both curves is shown in Figure 2.3.

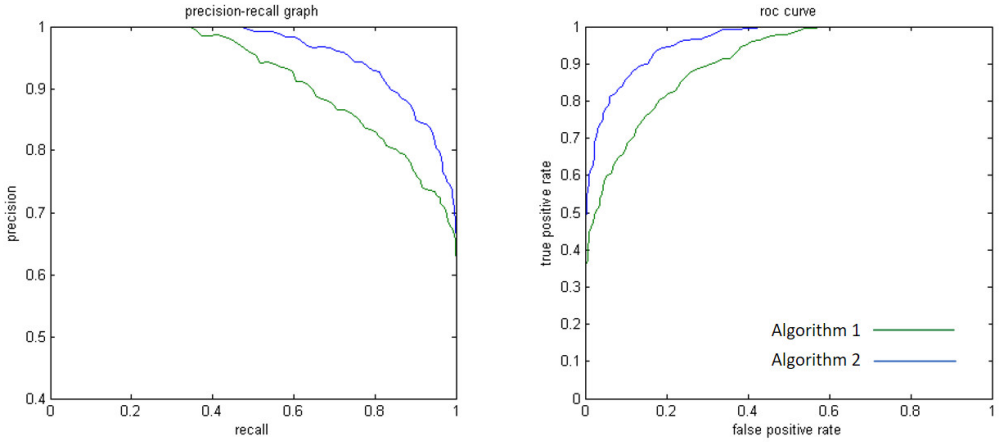

Figure 2.3: Precision and recall curves

2.5 Algorithm Parameter Tuning

Area of research known as *hyper-parameter* optimization focuses on selection of the best parameters for machine learning algorithms. Hyper-parameters are parameters that are not directly learnt within machine learning algorithms. Instead they need to be provided to algorithms as arguments. Several techniques for hyper-parameter tuning are documented [21], [30], [31], focusing on selection of the best parameters, best algorithm or best algorithm and parameters together. The hyper-parameter optimization is an automated method and selects the parameters based on a well-defined objective function.

In contrast to the hyper-parameter optimization methods, focus of this thesis is to allow users of the system enter their expert knowledge about expected behavior, help them understand the behaviors of the ICS devices and how anomaly detection modules can be applied. The thesis should explore possibilities for a design of a semi-automated platform that offers common and effective features for evaluating and comparing anomaly detection algorithms in an accessible way for ICS operators. Such a platform can be extended with more advanced methods based on the needs of the users.

Chapter 3

Problem Specification

Sections in this chapter discuss required features of the platform and particular requirements that the features must meet.

3.1 Configuration of Algorithm Arguments

The platform should allow users to select anomaly modules and parameters which they want to test and execute the analysis. Since both of the algorithms A-node and Wgng are to be configured using only numerical parameters, the platform needs to support numerical parameters. Algorithms have two types of parameter constraints: 1) Minimum and maximum for each parameter. 2) Mutual constraints between parameters. The platform needs to check whether a value of a parameter is within allowed range, verify the adherence to mutual constraints of the parameters and execute only the valid parameter sets. Apart from parameters, algorithms require training and test time intervals to be specified. Algorithm use values that were recorded within the train interval to learn parameters of a normal behavior. Time series values measured within the test interval are analyzed by algorithms and they return anomaly likelihood scores as a result. The platform needs to allow user to specify train and test intervals.

3.2 Data Labeling

Since the data is not annotated (it is not specified what parts of data belong to normal or anomalous behavior), platform needs to allow users to annotate data. Such an annotation is not to be used as training data for anomaly detection modules. The algorithms train only using the normal behavior of the system which is specified by the training interval. The annotation is used to evaluate whether algorithms can recognize specific type of anomalies. When users label the time series, the way of annotating should not force the user to annotate the whole time series. Instead, users should be able to choose parts that they want to annotate.

3.3 Scores Evaluation

The platform should evaluate scores produced by anomaly detection modules based on the annotation provided by the user. As defined in Section 2.3.1, scores produced by anomaly detection modules are series of numerical values; each value corresponds to a time interval. Scores represent likelihood that an anomaly occurred in a given time interval. The individual time intervals of scores can be of any length and can overlap. The values of scores can be any real numbers. The range of values can differ for each anomaly detection modules but also for the same anomaly detection module if it uses other parameter settings. Figure 3.1 visually shows how scores might look using a bar chart. The height of the bars is the anomaly likelihood score reported by algorithm for the time interval that corresponds to width of the bars. Some anomaly detection algorithms produce only labels, anomalous or benign. If such algorithms need to be evaluated, the anomalous/benign labels would first need to be converted to numbers (e.g. to 1 and 0 respectively). The result of evaluation should be the number of false/true positives/negatives and precision/recall for each possible threshold that can be applied to individual scores.

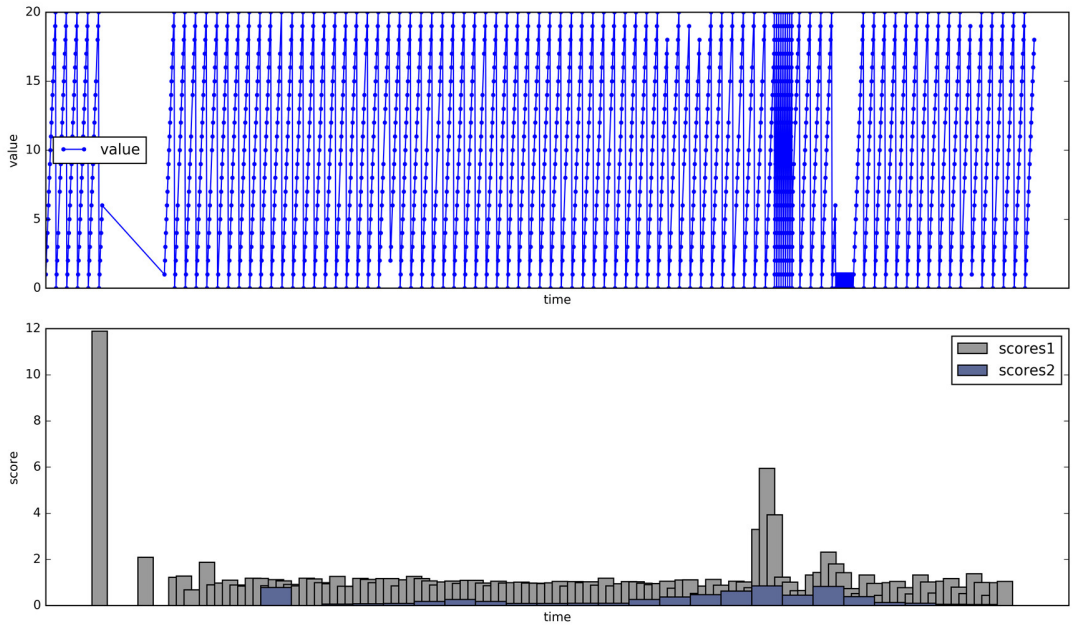

Figure 3.1: Example of scores produced by algorithms(bottom) for given time series (top)

3.4 Comparing evaluations

The platform needs to enable users to compare the evaluations for scores produced by various anomaly detection modules, parameter sets, training intervals, thresholds and anomaly annotations. The platform should allow users to sort the evaluations based on precision/recall and shortlist the anomaly detection modules and parameter sets that earned the best evaluations in regards to the anomaly annotation provided by users.

3.5 Implementation Requirements

The designed solution should provide good usability and offer interactive elements that will help users understand the data in a visual way. The system is to be integrated in the current IBM platform and the user interface style should be coherent with the interface of the existing platform.

Chapter 4

Solution Approach

This chapter presents proposed solution for an assistant. It proposes features and user interface elements to meet the goals outlined in Chapter 3.

I split the proposed solution into four main functional components: 1) configurator assistant, 2) results explorer, 3) evaluator and 4) evaluation explorer. The following sections describe the functional components in detail.

4.1 Configurator Assistant

The configurator assistant groups features which are necessary for configuring anomaly detection modules and executing the jobs. The features are: 1) displaying time series values, 2) selecting training and test intervals, 3) selection of parameter values, 4) generating combinations of parameter values, 5) validating combinations of parameter values, 6) executing anomaly detection modules.

The devices in the SCADA networks have different behaviors. Hence the configuration of algorithms individually for each device can result in better results of anomaly detection. For this reason, the proposed solution addresses configuration of algorithm modules and parameters for each device individually.

The configurator assistant user interface should display captured values of a device to allow user to explore the collected data.

The anomaly detection modules require training and test interval arguments to run. A user interface should contain element for configuring such intervals. The proposed solution allows users to select the intervals using sliders that mark up the selected interval in the captured values plot. Multiple pairs of training and test intervals can be added to test how selecting different training intervals affect performance of algorithms. Figure 4.1 shows the designed UI element. The light blue area of the slider is used to select data for training interval and the

dark blue area selects the test interval. Multiple pairs of training and test can be added and the added pairs show to the right from the slider.

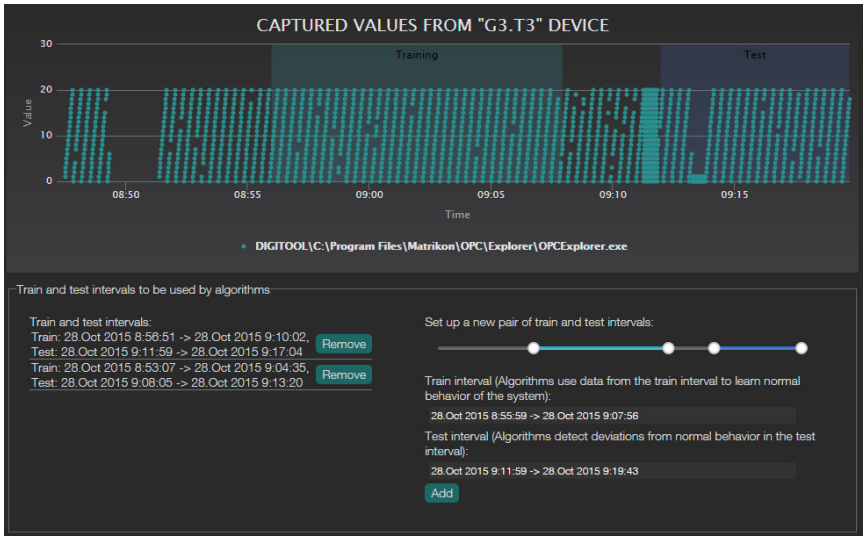

Figure 4.1: UI Element for Setting up training and test intervals

An important feature of the configurator assistant is the selection of parameters that should be tested. In the proposed solution users can input preferred values one by one or as a range of values with a step (e.g. start = 100, end = 200, step = 25) which is converted to single values when executing the algorithms (e.g. to 100, 125, 150, 175, 200). The proposed platform then generates a Cartesian product of input values for individual parameters. The system should prevent inputting values that breach the minimum-maximum constraints for individual parameters. Further, the system needs to checks mutual constraints for the generated parameter sets. Information about a number of valid and invalid parameter combinations provides instant feedback to the user. Figure 4.2 shows a proposed user interface element for configuring parameters. In the figure, the values of the "Width" – "w" parameter are being edited. The configuration interface element displays descriptions of parameters and the constraints. If user tries to input value outside the allowed range, an error message is shown.

Each valid parameter set combined with training interval, test interval and device identifier are sent to the anomaly detection modules to calculate the anomaly scores within the training interval.

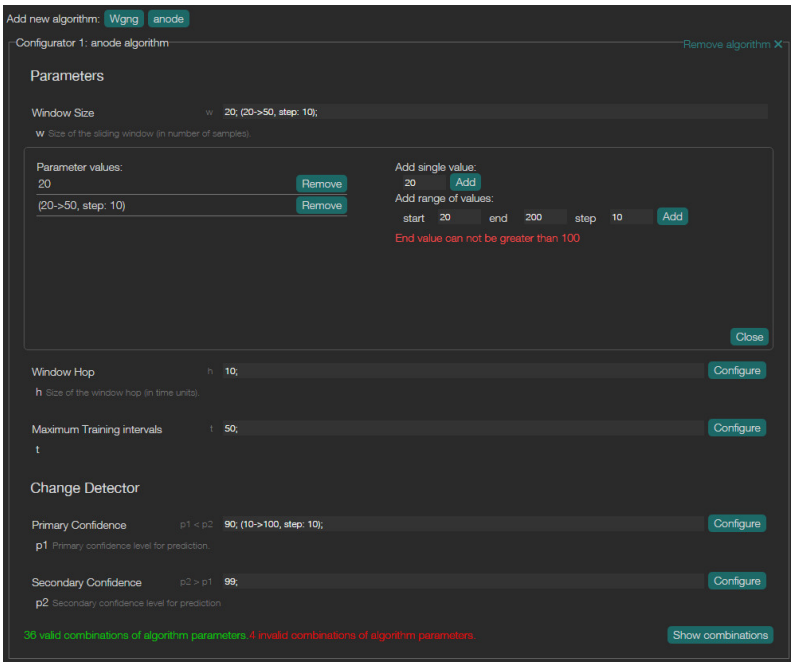

Figure 4.2: Proposed UI for configuring parameters

4.2 Results Explorer

In order for users to view and compare results of the anomaly detection modules analysis (scores) the solution should contain following features: 1) archiving of computed results, 2) presenting the results and 3) visualization of results

Once an anomaly detection module computed the results for a given set of arguments, such results, together with the original arguments should be persisted. Storing the computed scores together with the arguments enables working with the results in the future and compare them to other results. The list of the computed scores, together with the original arguments should be presented to the users enabling them to view the scores in a visual form. The scores should be displayed aligned with the time series interval which they report on. As shown in the Figure 3.1, scores can be represented well with a bar chart where width of the bar is the interval the algorithm reports on and the height of the bar is the reported value. Such a representation, however, quickly becomes hard to read, since the bars in the chart overlap. A simpler way of visualizing the scores, as a line chart, allows to view multiple scores at once. Figure 4.3 proposes a user interface element to compare results of multiple scores, aligned with time series.

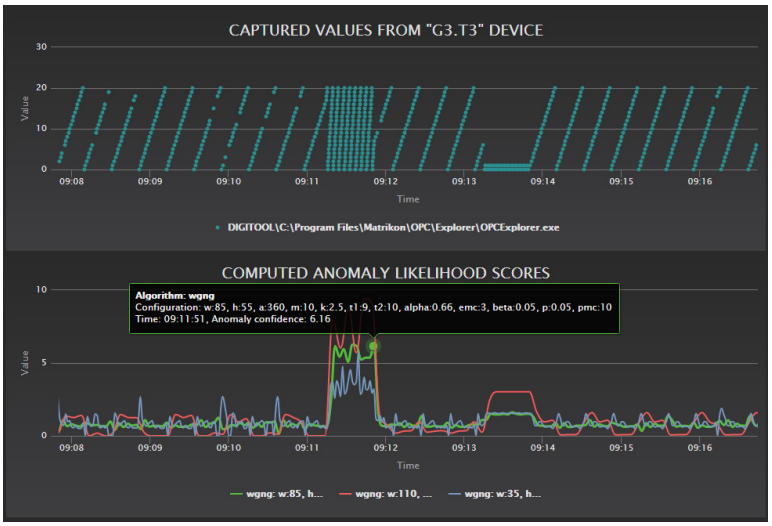

Figure 4.3: Visualization of scores, aligned with time series

4.3 Evaluator

One of the specified goals of the platform is to evaluate algorithms. In this section I propose a method for evaluating calculated scores, based on anomaly labels provided by users. An anomaly is an abnormal behavior of an Industrial Control System that operators of the ICS need to pay attention to. I propose a following definition of anomaly:

Definition 4.1 (Anomaly) An anomaly is a time interval $\boldsymbol{a} = \{\boldsymbol{a_s}, \boldsymbol{a_e}\}$ where $\boldsymbol{a_s}$ is the time when the anomaly started, $\boldsymbol{a_e}$ is the time when the anomaly ended and $\boldsymbol{a_s} \leq \boldsymbol{a_e}$.

Such a representation of anomaly is not dependent on the data points in the time series or any underlying data structures. It enables users to label anomalies within time when no data points appear (e.g. outage of the system). Start and end time of the anomaly can be the same, hence, an anomaly can represent a moment in time too. Since this format of representing an anomaly uses only time intervals, ICS experts can use it to markup irregular behavior that they observed in the real world independent of time series values.

If a time series is long, requiring users to study the whole time series and label it properly would be a tedious task. Instead, I propose a following method: users can select a time interval of interest and label anomalies that occur within such a time interval. I refer to such an interval as an evaluation range. It is defined as follows:

Definition 4.2 (Evaluation range) An evaluation range is a time interval $\boldsymbol{e} = \{\boldsymbol{e_s}, \boldsymbol{e_e}\}$ where $\boldsymbol{e_s}$ is the start time of the range and $\boldsymbol{e_e}$ is the end time of the evaluation range $\boldsymbol{e_s} \leq \boldsymbol{e_e}$. Parts of an evaluation range where user marks no anomaly are considered normal behavior.

A proposed interface element that allows users to set up anomaly labels together with evaluation range is presented in Figure 4.4. Using sliders, users can select a new anomaly interval and add it to a list of anomalies. Setting evaluation range, they assert that this range is annotated as they intent and can be used as a reference to evaluate results calculated by algorithms.

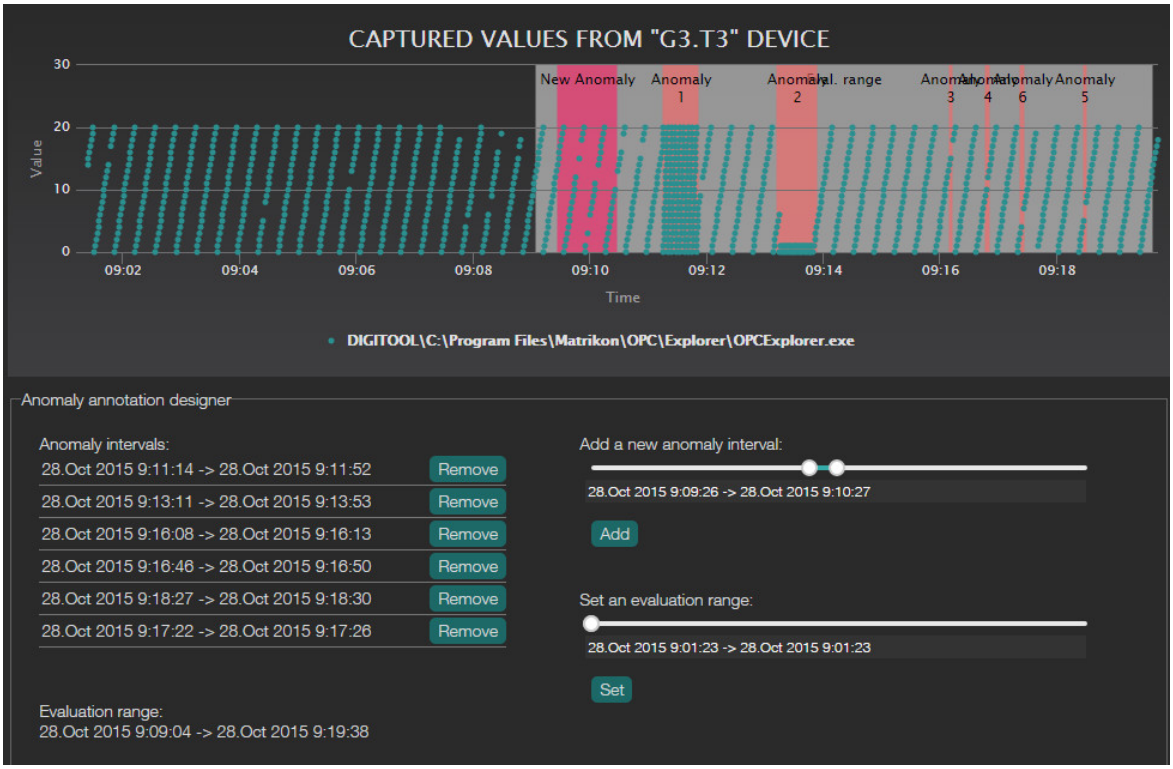

Figure 4.4: UI element for annotating anomalies

Based on an anomaly labeling and an evaluation range, scores can be evaluated. The goal is to compare how well scores produced by algorithms match the labeling provided by users. I aim to solve an evaluation problem, defined as follows:

Definition 4.3 (Evaluation problem) An instance of the evaluation problem is

$$\boldsymbol{E} = (\boldsymbol{s}, \boldsymbol{e}, \boldsymbol{t}, \boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n})$$

where $\boldsymbol{s}$ are scores, $\boldsymbol{e}$ is an evaluation range, $\boldsymbol{t}$ is a threshold and $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}$ are anomalies.

Definition 4.4 (Threshold) A threshold is a real number denoted as $\boldsymbol{t}$.

Instance of a solution of an evaluation problem is a set of following metrics: set of true positives, set of false positives, set of true negatives, set of false negatives, precision and recall (denoted respectively: $TP_t, FP_t, TN_t, FN_t, Prec_t, Rec_t$. Thus the solution is denoted as

$$L = \{TP_t, FP_t, TN_t, FN_t, Prec_t, Rec_t\}.$$

To calculate the metrics, I propose a following method:

There is no direct matching between scores and anomaly annotations. Both scores and anomalies are represented as time intervals. In order to create matching between scores and anomaly annotations I split individual scores s_i from s to three disjoint subsets. The split is based on whether time interval of s_i intersect with an anomaly interval a_i or an evaluation range e.

Definition 4.5 (Outer scores) Outer scores are a subset of scores from s . Their time intervals do not overlap with the evaluation range e. We denote them as s_{out}.

Definition 4.6 (Benign scores) Benign scores are scores that overlap with an evaluation range but do not overlap with any of the anomalies a_i. We denote them as s_b.

Definition 4.7 (Anomalous scores) Anomalous scores are scores that intersect with the evaluation range and at the same time they intersect with one or more anomalies a_i. We denote them s_a.

Figure 4.5 illustrates splitting of scores based on existence of intersection with anomaly intervals (marked as gray bands). Evaluation range spans the whole figure area, so there are no elements in s_{out}.

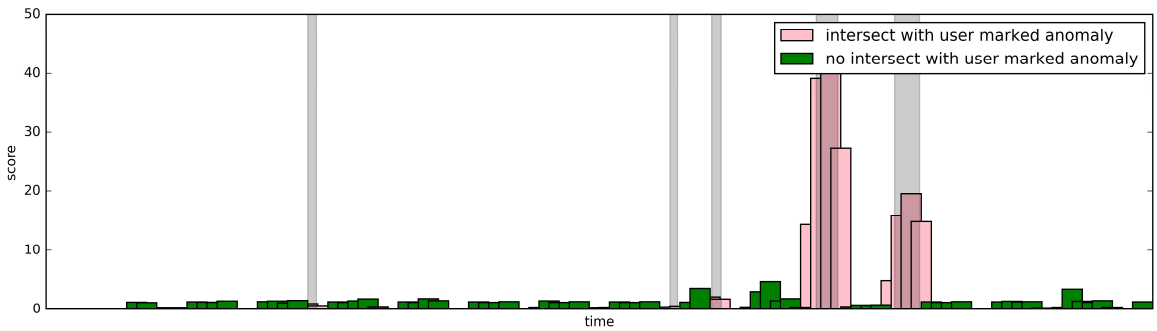

Figure 4.5: Scores classification based on intersection with anomaly

Further we can split scores s into two disjoint subsets based on a selected threshold value:

Definition 4.8 (Positive scores) Let $s_i = \{b_i, e_i, y_i\}$, $s_i \in s$. If a value of score y_i is greater or equal to the threshold, t then set of positive scores s_{pt} contains s_i.

Definition 4.9 (Negative scores) Let $s_i = \{b_i, e_i, y_i\}$, $s_i \in s$. If a value of score y_i is lower than t, then set of positive scores s_{nt} contains s_i.

In an ideal situation all scores that intersect with anomalies marked by user would have greater values than the scores which do not intersect with anomalies. To accomplish this in the presented figure, all red scores would have to be taller than green ones. This would mean that a threshold exists such that scores can be split to perfectly match expectations of the user ($s_a \subset s_{pt}$ and $s_b \subset s_{nt}$).

The five defined sets have following properties:

$$s_{out} \cup s_b \cup s_a = s$$

$$s_{out} \cap s_b =$$

$$s_{out} \cap s_a =$$

$$s_b \cap s_a =$$

$$s_{pt} \cup s_{nt} = s$$

$$s_{pt} \cap s_{nt} =$$

Comparing the sets resulting from split by user annotation (s_{out}, s_a, s_b) and sets resulting from split by threshold t (s_{pt}, s_{nt}), true/false positives/negatives sets are defined as follows:

True positive scores for threshold t are $TP_t = s_{pt} \cap s_a$

False positive scores for threshold t are $FP_t = s_{pt} \cap s_b$

True negative scores for threshold t are $TN_t = s_{nt} \cap s_b$

False negative scores for threshold t are $FP_t = s_{nt} \cap s_a$

Figure 4.6 and Figure 4.7 demonstrate how different thresholds affect the classification into sets. Based on size of the sets, we can compute a precision for scores s and threshold t as

$$Prec_t = \frac{|TP_{tij}|}{|TP_t| + |FP_t|}$$

We can compute a recall for scores s and threshold t as

$$Rec_t = \frac{|TP_{ijt}|}{|TP_{ijt}| + |FN_{tij}|}$$

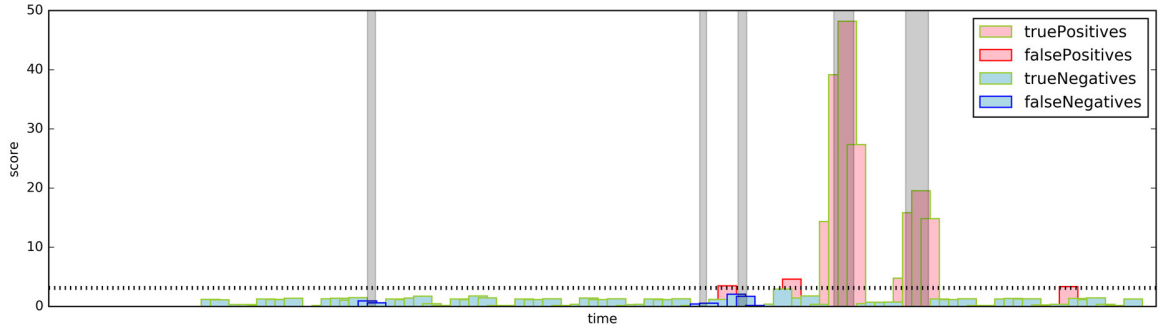

Figure 4.6: classification of scores based on threshold and user labeling

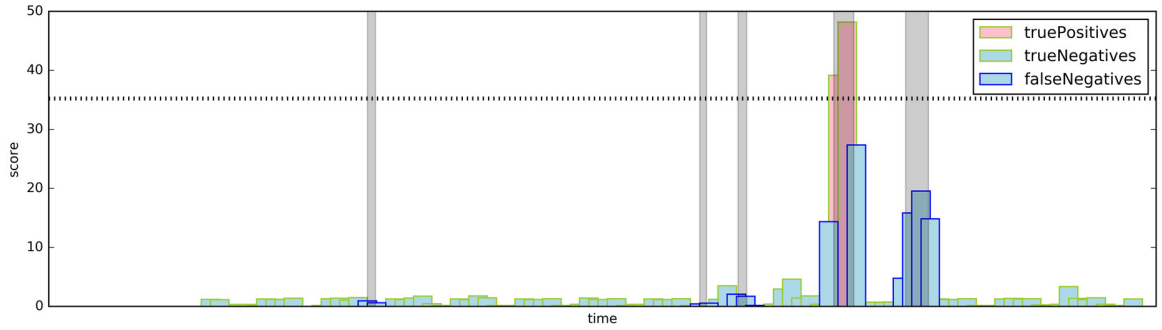

Figure 4.7: classification of scores based on threshold and user labeling, greater threshold

Figure 4.7 shows that the proposed split to sets might be not desirable. In the example from the figure, the current split marks the two bars adjacent to the second anomaly from the end of time series as false negatives. However, the algorithm did manage to detect the anomaly marked by the user. To fix this, an alternative way of marking true positives and false negatives is as follows: If there is at least one true positive score ($s_i \in s_{pt}$) that intersects with an anomaly a_k, then all other scores s_j that also intersect with an anomaly a_k will be considered true positives as well. A result of applying the false negatives fix is illustrated in Figure 4.8.

This adjustment has an impact on the usability of the platform. Without the fix, users should annotate anomalies very precisely and minimize the length of anomaly interval, to only label necessary time. With the fix users can label time interval that contains anomaly without knowing the exact time span of the anomaly. Then running the evaluations, they can identify

an algorithm that is capable of detecting an anomaly in the range, even if the location of the anomaly was not apparent based on the time series values.

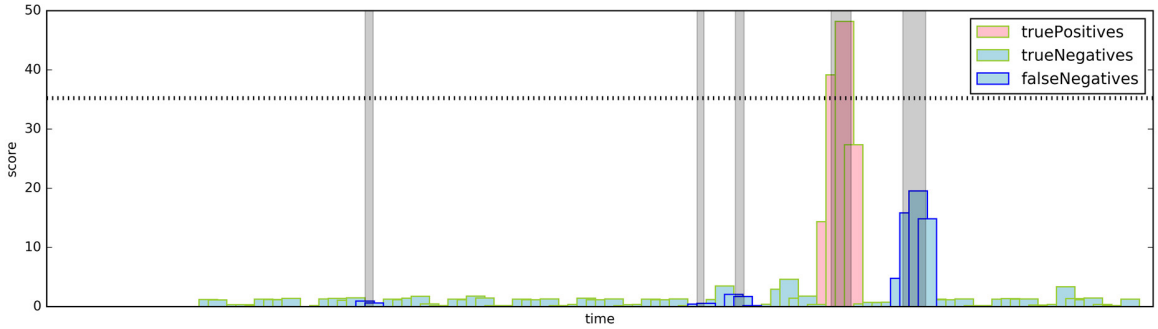

Figure 4.8: classification of scores based on threshold and user labeling, fix for false negatives

4.4 Evaluation Explorer

By applying the described method, an evaluation can be computed for every threshold of scores.

Precision recall curves for all thresholds of anomaly detection module with specific parameter set can be used as a basis for comparing parameter sets and algorithms.

Definition 4.10 (Precision-Recall Curve) A Precision-Recall curve is a set of tuples of precision and recall calculated for all possible thresholds $\boldsymbol{t_i}$ for scores $\boldsymbol{s}$. It is denoted as

$$\boldsymbol{PR} = \{(\boldsymbol{p_1}, \boldsymbol{r_1}), (\boldsymbol{p_2}, \boldsymbol{r_2}), \ldots, (\boldsymbol{p_n}, \boldsymbol{r_n})\}$$

A visual way to compare precision recall curves can help users quickly understand a relation between algorithms and parameters. Figure 4.9 shows a proposed way to quickly – only by moving the mouse cursor – compare threshold settings for multiple algorithms setups. The highlighted point in the figure represents one of many possible thresholds which can be selected. In the "Captured anomaly likelihood scores" plot, user can see scores that produced given precision recall curve and a threshold associated with the precision and recall. Additionally, over the threshold line, number of true positives and true negatives is given.

Algorithm configurations which result in poorly performing precision recall curves can be filtered out in following ways:

- Filtering scores out by minimum acceptable recall and minimum acceptable precision

- Sorting the remaining scores by the best possible value of precision that meets the minimum acceptable recall or analogically, by the best possible value of the recall that meets the minimum acceptable precision.
- Filtering out all results which have a precision recall curve dominated by another precision recall curve

Definition 4.11 (Precision-Recall Curve is dominated) A Precision-Recall curve PR^b for scores s^b is dominated by a Precision-Recall curve PR^a for scores s^a if:

$$\forall p_i^a, r_i^a \exists p_i^b, r_i^b : p_i^a \geq p_i^b \wedge r_i^a \geq r_i^b \text{ and } \exists p_i^a, r_i^a, p_i^b, r_i^b : p_i^a > p_i^b \wedge r_i^a > r_i^b$$

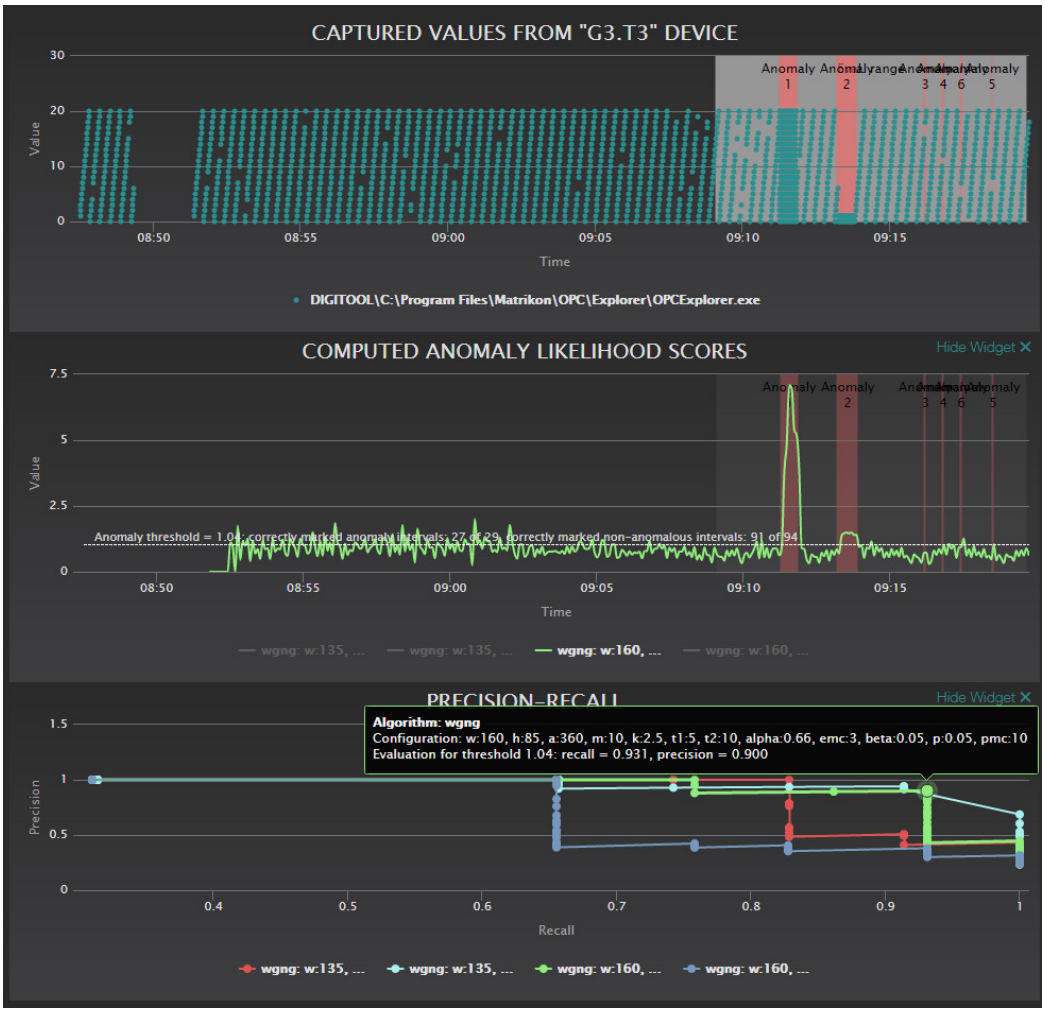

Figure 4.9: UI element for comparing thresholds of Precision-Recall curves

Chapter 5

Implementation

In the implementation part of this project I have created the system with features, as described in Chapter 4, that consists of four components: an assistant platform frontend, an assistant platform backend, a scores evaluator module and a database to store results of the computations. This chapter explains the architecture and implementation details of the system.

5.1 Architecture

This section describes individual components of the platform and their relationship with other components. The architecture is depicted in Figure 5.1.

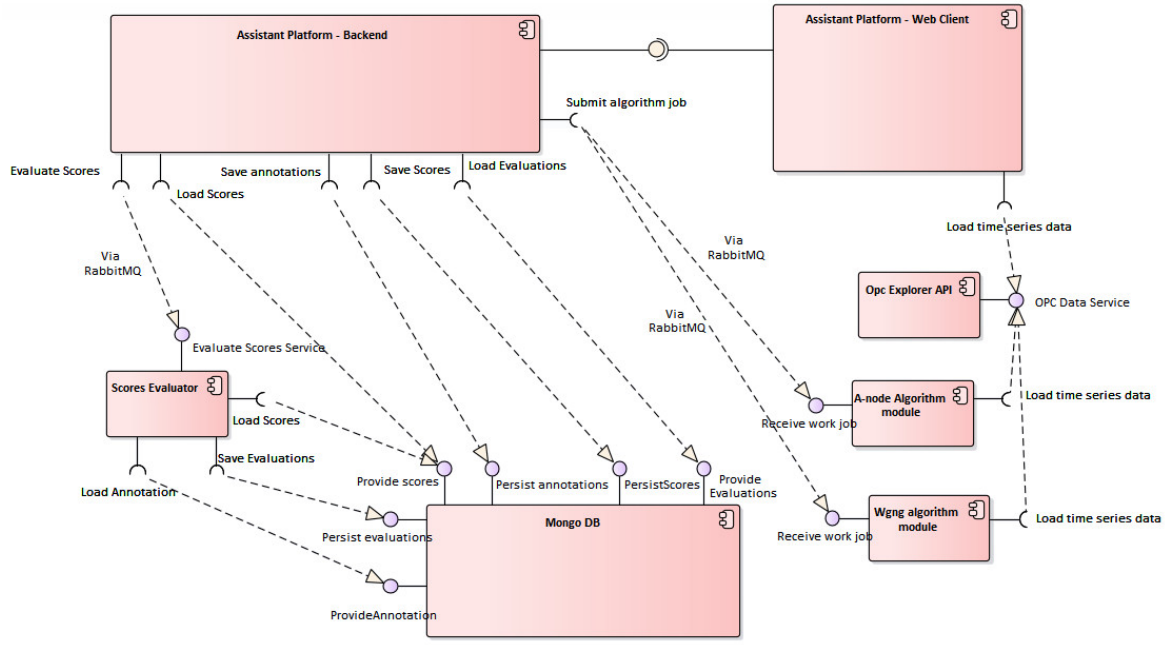

Figure 5.1: Platform Architecture

5.1.1 Assistant Platform – Frontend

The frontend of the assistant platform is developed with ReactJS [32] and ReduxJS [33] frameworks. These frameworks enable reusing of created user interface components and a unidirectional flow of information that keeps the complex UI coherent. A state of the webpage in the browser is fully dependent on the ReduxJS *store* variable that is modified in a single place – a *reducer* which processes *actions* fired by UI elements or by received socket messages. ReactJS uses *virtual DOM (document object model)* where the updates to UI are performed first. Only when a change is detected in the *virtual DOM,* it is propagated to a browser DOM. Combining ReactJS and ReduxJS allows the website to functions well as a single page application [34]. The main responsibilities of the component are:

- generating valid combinations of parameters for algorithms, based on user input
- combining them with selected training and test interval and an ID of the source field device (sensor or actuator where the time series being analyzed was recorded)
- generating a job for anomaly detection modules and sending it to an assistant platform backend
- receiving results of the jobs (scores) and updating the table of scores
- visualizing time series data
- creating anomaly annotations that combine multiple anomaly intervals and an evaluation range
- saving the anomaly annotation to the database using the backend as a middleman
- loading existing anomaly annotations and displaying a table of them
- executing evaluation of all scores in the database, comparing them to a selected anomaly annotation, using the backend as a middleman
- loading evaluations from the database via the backend
- displaying score values and precision recall curves

Webpack [35] and Babel [36] translate JSX [37] and ES6 (ECMAScript 2015) [38] expressions into widely accepted ES5 standard. The frontend communicates with two components: the OPC Explorer API and the assistant platform backend. The communication with OPC Explorer API is via HTTP (Hypertext Transfer Protocol) and is used to load time series values. The communication with the backend is via SocketIO [39] web socket. Many of the user interface components are manually written, some of them (sliders, tabs) are from other libraries.

5.1.2 Assistant Platform – Backend

The backend is implemented with NodeJS. The main responsibility of the backend is to receive messages from the frontend. Based on the received messages, it sends queries to the database or submits jobs to *anomaly detection modules* or *scores evaluator module*. The backend communicates with Mongo DB [40] via the HTTP API. Communication with anomaly modules and scores evaluation module is over RabbitMQ message queue [41]. The backend loads archived algorithm job descriptions together with the results of jobs (scores), anomaly annotations and evaluations of the scores from the database and returns them to the frontend. It constructs *find*, *aggregation* and *map-reduce* queries for Mongo DB to retrieve specific views of data, including the query for the set of non-dominated precision-recall scores and precision recall curves filtered by minimum value of precision/recall as described in Section 4.4. Thanks to expressive query language of Mongo DB, backend needs to do little extra data processing.

5.1.3 Scores Evaluator Module

The scores evaluator module uses Python [42] with Pandas [43] and NumPy [44] libraries to evaluate scores comparing them to the anomaly annotation created by users. The computation module itself that needs the annotation and scores data as arguments is wrapped with a database loader wrapper. The wrapper loads the scores and annotation directly from the Mongo DB and saves results of evaluation back to Mongo DB. In this way the data does not have to me shuffled through the assistant platform backend. To fetch the data from the database, the scores evaluator module receives only IDs of documents to work with from the assistant platform backend. The module runs on the server as a Docker [45] container and pulls new jobs from RabbitMQ [41] message queue. Thanks to the Docker deployment, the module can be run scaled up by replicating the instances to speed up evaluating hundreds thousands of scores. The instances connect to message queue pool on the start and pull unprocessed jobs

5.1.4 Mongo DB

MongoDB fits great for the task of storing documents such as scores, evaluations and anomaly annotations. Documents can be nested in a natural structure. The jobs submitted to anomaly detection modules are archived in Mongo DB. When algorithms finish the job descriptions in the database are updated with the results of the jobs (scores). Scores are saved inside a job as a MongoDB *embedded document*. When scores are evaluated based on anomaly annotations

provided by users, evaluations for respective anomaly annotations are stored as embedded documents inside the score document. The anomaly annotations are stored in a separate Mongo DB database, since they do not need a link to former.

5.2 User Interface

This section presents a user interface of the implemented assistant platform, split into three tabs that separate functionality of the platform. The referenced figures can be found in Appendix C.

5.2.1 Configurator Tab

The field device table resides on the top of the webpage. It allows to see basic statistics about field devices and select a particular device. In Figure C.1 the configurator tab and G3.T3 field device is selected. The "Captured Values" plot presents the data from the device. Under the plot, there is a panel for setting up training and test intervals. These intervals are used by anomaly detection modules to learn normal behavior and analyze data. Even lower, there is a pane with configurator available for each anomaly detection module (A-node or Wgng). Using the configurators, user can generate large number of parameter combinations quickly. The configurator automatically checks the validity of combinations taking the constraints of an algorithms in mind. On the bottom of the page, a "Run" button is displayed. Clicking the button will instruct anomaly detection modules to analyze the data running using the generated parameter sets.

5.2.2 Results Tab

Results tab contains previously computed scored from anomaly detection modules. The scores are archived in Mongo DB database and can be explored using an interactive plot. The view is shown in Figure C.2.

5.2.3 Evaluator Tab

The evaluator tab allows users to create anomaly annotations and run evaluations of algorithms (Figure C.3). When evaluations are calculated, the evaluator tab allows users to explore calculated Precision-Recall curves and filter based on minimum precision or minimum recall.

The “Hide algorithm configurations with non-optimal Precision-Recall curves” filtering option in the table at the bottom of the screen in Figure C.4 hides scores algorithm configurations that have dominated Precision-Recall curves.

Chapter 6

Assessment and Evaluation

To evaluate the developed assistant platform, I have conducted a testing with users. This chapter describes the process of testing preparation, the results of the test.

6.1 Goal and Metrics

The goal is to invite users to test the developed software and to provide an evaluation of the quality of the solution. The user testing provides insights about how users perceive the assistant platform and how much guidance users require to use the system effectively. The testing can help to identify the most useful functions, opportunities to improve the user interface and inspire ideas for new features. Additionally, users with knowledge of the security domain can provide feedback and ideas for improvement.

6.2 Target Group

The tested software focuses on configuration, evaluation and comparison of anomaly detection modules for ICS. The ideal candidates for a user interface testing would be ICS operators and security consultants. Due to the limited access to such ideal candidates, I had to extend the target group. Since the problem that the assistant platform addresses is complex, I included individuals that are pursuing or have completed higher education, assuming that such users can understand the problem and adopt similar approach to address it.

6.3 Test Preparation - Surveys

In preparation for the testing, I have established several surveys and an informational guide for participants. This section explains the role of the prepared documents in the user testing process. Appendix C contains all documents.

6.3.1 Screening Survey

The purpose of the screening survey is to verify whether candidates meet pre-defined criteria for participation in the test. The participants should be pursuing higher education or should have completed it. They should be also able to understand mathematical plots, since the assistant platform contains several. Further requirements include that the participants feel comfortable with using advanced interactive websites and good command of English. Finally, I wanted to invite some users who understand machine learning and statistics and some users who have no previous experience with the above mentioned.

Table 6.1 shows results of the screening survey. Since, I was actively searching for the candidates that meet the criteria, all the candidates met the requirements. One participant did not have previous experience with statistics or machine learning.

Question	**Answer counts**
Do you currently pursue or have you previously completed a higher education degree (university/university of applied sciences/other post-secondary education)?	Yes: 5 No: 0 Cannot answer: 0
What is your experience reading mathematical plots (graphs)?	High: 5 Intermediate: 0 Basic: 0 Lower or none: 0
How comfortable do you feel using modern interactive websites (for example any of following: gmail.com, maps.google.com, google drive, drop box/box/iCloud or purchasing airplane tickets online)?	High: 5 Intermediate: 0 Basic: 0 Lower or none: 0
Do you have a work experience or have you completed a university course in statistics, machine learning, statistical learning or anomaly detection?	Yes: 4 No: 1 Cannot answer: 0
What is your command of English?	Very high: 5 High: 0 Intermediate: 0 Basic: 0 Lower or none: 0

Table 6.1: Results of the screening survey

6.3.2 Pre-Test Questionnaire

The pre-test questionnaire is answered by participants who meet the conditions set by the screening survey. The purpose of the questionnaire is to get more detailed information about individual participants. Having more information about participants can help to understand their approach to working with the software. The pre-test questionnaire contains questions about professional specialisation, degree of experience using software to solve machine learning tasks and degree of experience using anomaly detection software. The last part of the questionnaire is open ended and invites participants to list their computer skills. The complete questionnaire is included in Appendix C. Summaries of answers given by participants are included in Section 0, which evaluates the testing sessions.

6.3.3 Information Guide for Participants

Before the participants started working with the assistant platform they were provided with an information guide that explained fundamentals of anomaly detection and evaluation of anomaly detection modules. The guide explained what time series, anomalies, training and test intervals are, as well as how precision and recall curves can be used to compare algorithms. The copy of the information guide is provided in Appendix C.

6.3.4 Post-Test Questionnaire

Question ID	Question
A1	How do you think a presence of an assistant affected you?
A2	I considered the tasks
A3	How would you describe your experience working with the software?
A4	Do you have any suggestions for improving the software?
A5	Do you have any suggestions for new functionality of the software?
A6	Please evaluate following statement: Information guide provided before the testing helped me in completing the tasks.

Table 6.2: Post-test questionnaire questions – set A

After testing the platform, participants filled in a post-test questionnaire. Questions in the questionnaire focus on obtaining feedback about the assistant platform. Two sets of questions were included. The first set of questions enabled more open ended answers and multi choice selection. The first set of questions is listed in Table 6.1. The second one is a standardised set

of *Perceived Usefulness and Ease of Use* [46] questions where user had to mark a number on a scale from *Likely* to *Unlikely*. The questions are listed in Table 6.3.Complete questionnaire with answer options is in Appendix C.

Question ID	Question
B1	Using the system in my job would enable me to accomplish tasks more quickly
B2	Using the system would improve my job performance
B3	Using the system in my job would increase my productivity
B4	Using the system would enhance my effectiveness on the job
B5	Using the system would make it easier to do my job
B6	I would find the system useful in my job
B7	Learning to operate the system would be easy for me
B8	I would find it easy to get the system to do what I want it to do
B9	My interaction with the system would be clear and understandable
B10	I would find the system to be flexible to interact with
B11	It would be easy for me to become skillful at using the system
B12	I would find the system easy to use

Table 6.3: Post-test questionnaire questions – set B

6.4 Set-Up of the Test

The test sessions with participants took place on premises of IBM Research Zurich laboratory on December 19, 2016 from 13:30 to 18:30. Each of the five conducted sessions took approximately 50 minutes, including completing the tasks and filling in the questionnaires.

6.4.1 Roles

During the test, I was the only present person apart from the participant, acting as a moderator. As a moderator, I provided the participants with the necessary assistance and guided them through the test. I did not help participants with the tasks unless some exceptional situation occurred.

6.4.2 Environment Set-Up

The test took place in a quiet meeting room with a table and number of chairs. The questionnaires and tasks for participants were provided on paper. During the test, the participant was alone in the room with the moderator (me). The participants worked with the platform on a

laptop with a 14'' screen and pixel resolution of 1920x1080, US English keyboard and mouse with a scroll wheel. The operating system of the computer was Windows 10 [47]. The screen of the laptop was recorded with CamStudio software [48]. The laptop's integrated microphone array was used to record audio in the room.

6.4.3 Initial State of the Application

When participants start working with the computer, the Google Chrome web browser [49] is set to full screen mode so that the assistant platform web page fills the whole screen of the laptop. Figure 6.1 shows the initial state of the laptop screen and also a zoomed area of the device list.

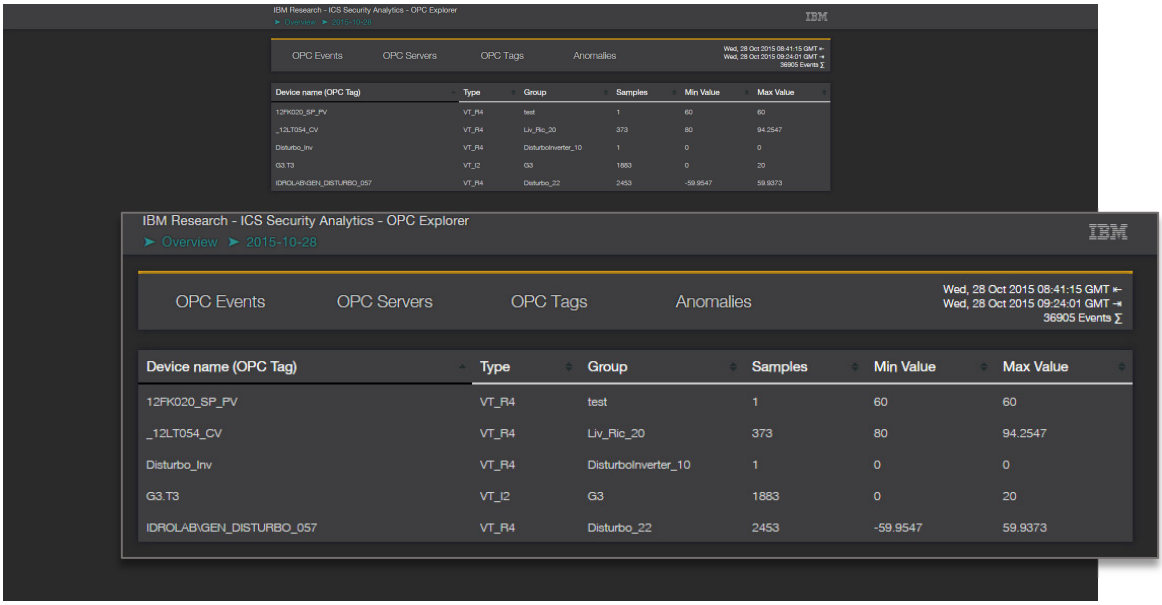

Figure 6.1: Initial state, screenshot and zoomed area

6.5 Tasks for Participants

Tasks 1-6 focus on selecting a desired field device, viewing captured time series data, setting up trading and test intervals, configuring and running Wgng [19] anomaly detection module. In tasks 8-14, participants should label the time series with anomalies, evaluate and compare outputs produced by Wgng. Tasks 15-21 focus on configuring the other available algorithm – A-node, evaluating its results and comparing outputs of both algorithms. In some tasks participants are requested to explain how they understand how features of the platform work

(e.g. to understand filtering features). Instruction in some tasks are very precise while some tasks are more open ended.

6.5.1 List of Tasks

Table 6.4 includes a complete list of tasks.

Task 1: Select G3.T3 device
Select G3.T3 device from a device list.

Task 2: View captured data
Examine a plot of values captured from the G3.T3 device. Try to understand the plot.

Task 3: Set up training and test intervals
Set up the system to use the following time interval as a training interval (times that you select can differ slightly):

From 28.Oct 2015 8:51:20 to 28.Oct 2015 9:07:04

Set up the system to use the rest of the captured data as a test interval for algorithms (times that you select can differ slightly):

From 28.Oct 2015 9:07:04 to 28.Oct 2015 9:20:00

Task 4: Generate combinations of parameters for the "Wgng" algorithm
Configure Wgng algorithm to use following values of parameters (leave other parameters' default values):

Parameter	Values for parameter
Window size (w)	Remove the default value and add values between 35 and 300 with step 25: 35, 60, 85, 110, 135, 160, 185, 210, 235, 260, 285
Window hop (h)	Remove the default value and add values between 25 and 300 with step 15: 25, 40, 55, 70, 85, 100, 115, 130, 145, 160, 175, 190, 205, 220, 235, 250, 265, 280
Neuron Memory (t1)	0, 1, 3, 5, 7, 9
Other parameters	Leave default values

Task 5: Check valid combinations
Check which combinations of parameters are valid combinations.

Task 6: Execute the algorithm
Execute the algorithm with configured combinations.

Task 7: Examine results

Examine some results produced by algorithm. Try to understand what they mean.

Task 8: Switch to evaluator tab

Switch to evaluator tab and familiarize yourself with the current view.

Task 9: Annotate anomalies

Provide annotation of anomalies to the system. Set up a system so that it considers following time intervals as anomalies:

Anomaly #	**Interval**
1	from 28.Oct 2015 9:11:14 to 28.Oct 2015 9:11:51
2	from 28.Oct 2015 9:13:10 to 28.Oct 2015 9:13:53
3	from 28.Oct 2015 9:16:10 to 28.Oct 2015 9:16:13
4	from 28.Oct 2015 9:16:46 to 28.Oct 2015 9:16:50
5	from 28.Oct 2015 9:18:27 to 28.Oct 2015 9:18:30
6	from 28.Oct 2015 9:17:22 to 28.Oct 2015 9:17:26

Task 10: Set up evaluation range

Set up a system so that it runs evaluation in the following time interval (times that you select can differ slightly):

From 28.Oct 2015 9:09:04 to 28.Oct 2015 9:19:38.

Task 11: Run the evaluation

Save the anomaly annotation that you created and run evaluation of algorithm configurations.

Task 12: Explore evaluations for some of the configurations.

Explore few of the computed evaluations of the parameter value combinations.

Task 13: Find the best parameter configuration with precision at least 95% (P = 0.95)

Find which configuration would be best when it is required that the precision of the algorithm in identifying anomalies is at least 95%. In other words, the precision of the algorithm needs to be at least 95% and at the same time the recall should be as high as possible.

Task 14: Explore filtering options

Select "Hide algorithm configurations with non-optimal Precision-Recall curves" option. Try to understand what it does.

Task 15: Switch back to the configurator tab

Switch back to the view that allows you to configure algorithms.

Task 16: Remove the "Wgng" algorithm configurator

Remove the "Wgng" algorithm configurator widget from configurator tab.

Task 17: Generate combinations of parameters for the "A-node" algorithm

Configure A-node algorithm to run with following values of parameters and all their combinations (the other parameters should be left at a default value):

Parameter	Values for parameter
Window size (w)	Remove the default value and add values between 20 and 40 with step 5: 20, 25, 30, 35, 40
Window hop (h)	Remove the default value and add values between 10 and 20 with step 5: 10, 15, 20
Other parameters	Leave default values

Task 18: Execute the algorithm

Execute the algorithm with configured combinations.

Task 19: Switch to results tab and wait for the results to be computed

Switch to results tab and wait (about one minute after executing computation) until all results of the "A-node" algorithm are computed.

Task 20: Navigate to evaluator tab and evaluate new algorithm results

Navigate to evaluator tab and evaluate the newly created algorithm parameter combinations with the saved anomaly annotation that you created before.

Task 21: Compare "A-node" to "Wgng"

If not selected, select "Hide algorithm configurations with non-optimal Precision-Recall curves" option in the "Executed algorithm configurations" widget. Try to understand how results of A-node algorithm compare with "Wgng". Try setting minimum recall to R = 1. Reset minimum precision setting to 0. Try to understand how "A-node" compares to "Wgng" with minimum recall set to R = 1.

Table 6.4: List of tasks

6.5.2 Optimal Completion of Tasks

In this section we present the optimal way to complete the tasks.

6.5.2.1 Task 1: Select G3.T3 device

Clicking on the G3.T3 row in the displayed table completes the Task 1 (shown in Figure 6.2).

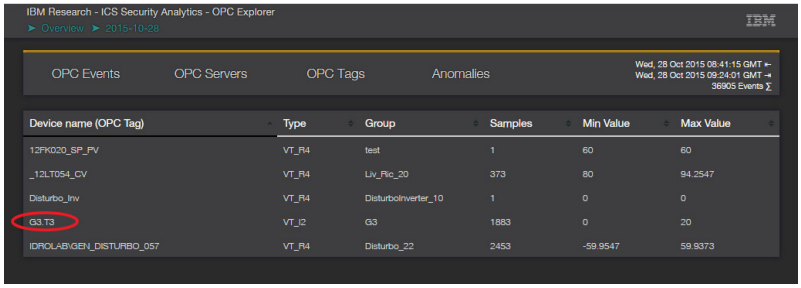

Figure 6.2: Task 1 completion

6.5.2.2 Task 2: View captured data

To complete Task 2, users should explore the displayed time series for the G3.T3 device. There is no preferred way. The purpose of this task is to learn how participants understand the time series plot. Figure 6.3 shows the user interface after completion of Task 1 and marks user elements that need to be used to complete Task 3 and Task 4.

6.5.2.3 Task 3: Set up training and test intervals

To complete Task 3, users should move knobs of the train and test interval sliders to set up training interval and test interval as instructed.

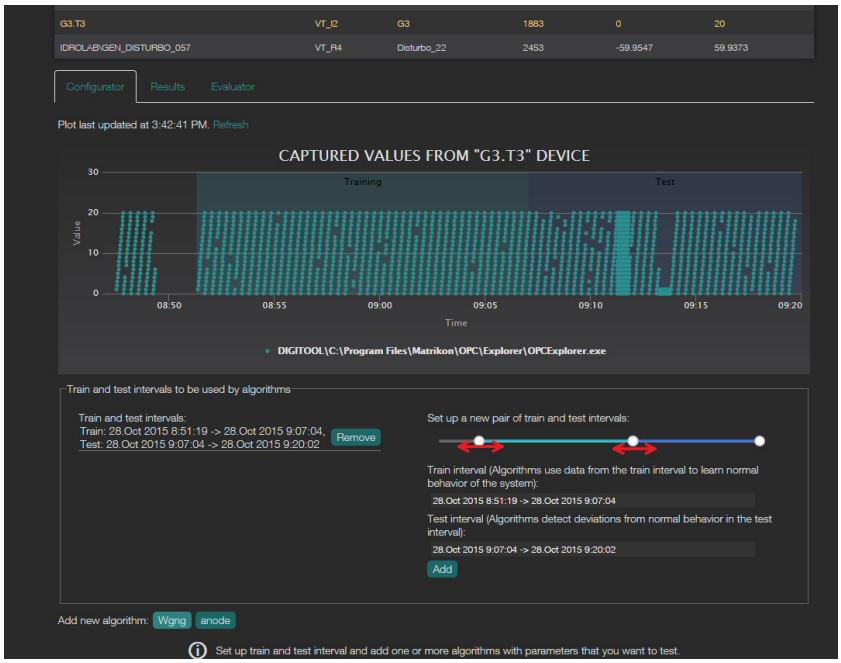

Figure 6.3: Completion of Tasks 2 and 3

6.5.2.4 Task 4: Generate combinations of parameters for the Wgng algorithm

To complete Task 4, participant should click on the "Wgng" button which is highlighted in. Consequently, the configuration interface for the Wgng algorithm appears. For each parameter of the Wgng algorithm that needs to be configured, participant should click on a configure button next the parameter and setup the parameter range. Setting up of the Window size parameter is shown in Figure 6.4.

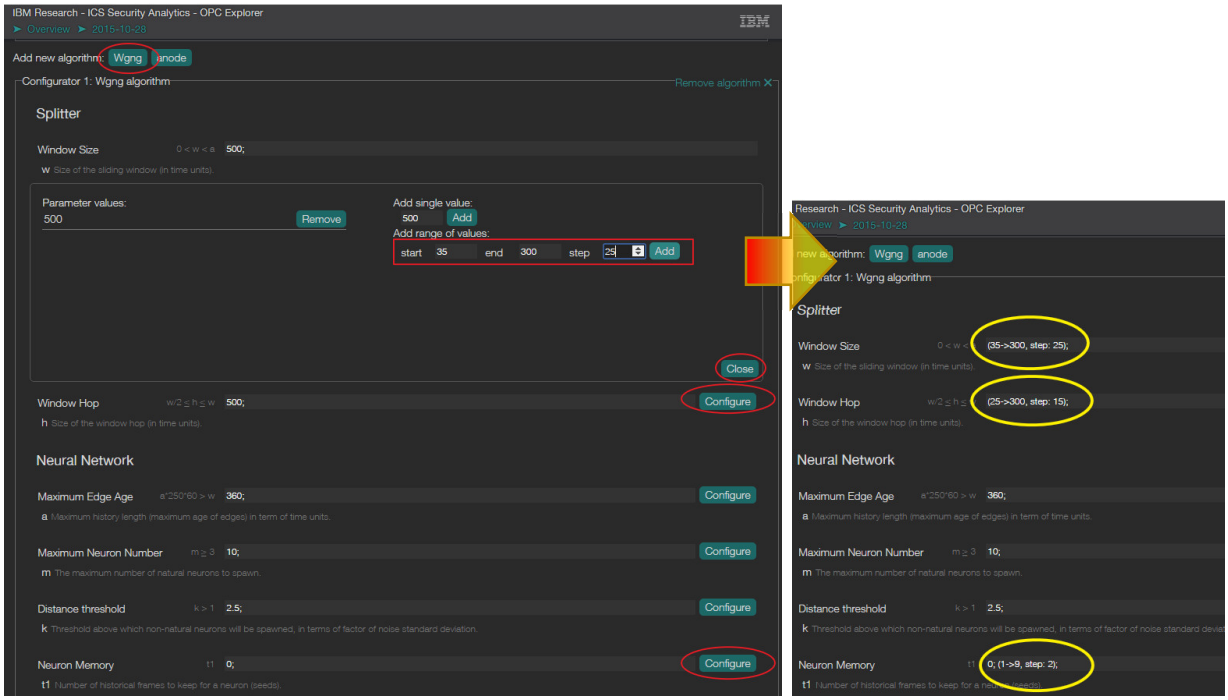

Figure 6.4: Completion of Task 4. Necessary actions (left) and result (right)

6.5.2.5 Task 5: Check valid combinations

To complete Task 5, participants should click on the *Show combinations* button. In the table that appears they should explore the valid and invalid combinations of parameters using the pagination menu. Figure 6.5 highlights the user interface elements that should be used.

6.5.2.6 Task 6: Execute the algorithm

To complete Task 6, participants should click on the button showing "Run 366 algorithm configurations". Confirmation message is shown as presented in Figure 6.6.

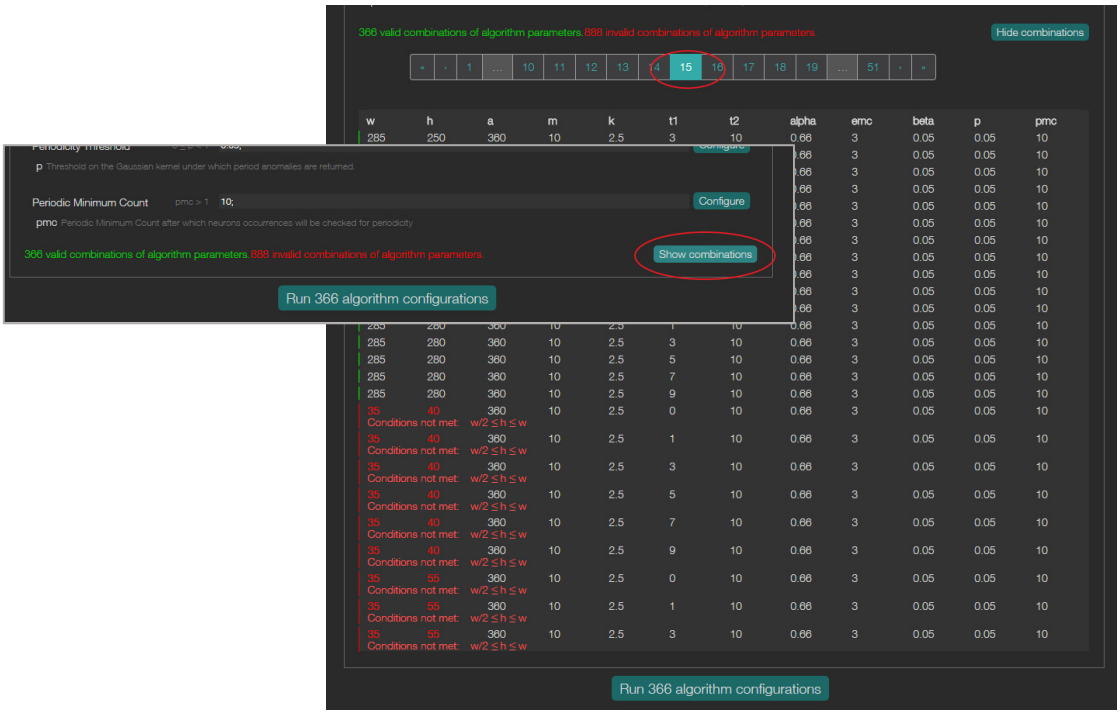

Figure 6.5: Completion of Task 5

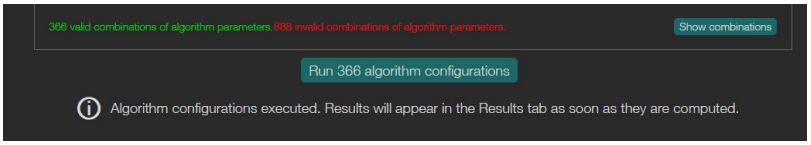

Figure 6.6: Completion of Task 6

6.5.2.7 Task 7: Examine results

To complete Task 7, participants should click on the "Results" tab, then use "Show Results" buttons in the table on the bottom of the screen and explore algorithm outputs (scores) that are displayed in the plot. Figure 6.7 highlights the relevant user interface elements.

6.5.2.8 Tasks 8, 9, 10, 11

To complete Tasks 8 – 11, participants should select the "Evaluator" tab, use the knobs on the "Add new anomaly interval" slider to select start and end time of anomalies they want to add. "Add" button adds the anomaly to the list on the left. Evaluation range should be added by moving the knobs on the "Set an evaluation range" slider and then clicking on the "Set" button. "Save anomaly annotation and evaluate all algorithm configurations" button will run the evaluation. Figure 6.8 shows the steps to be taken.

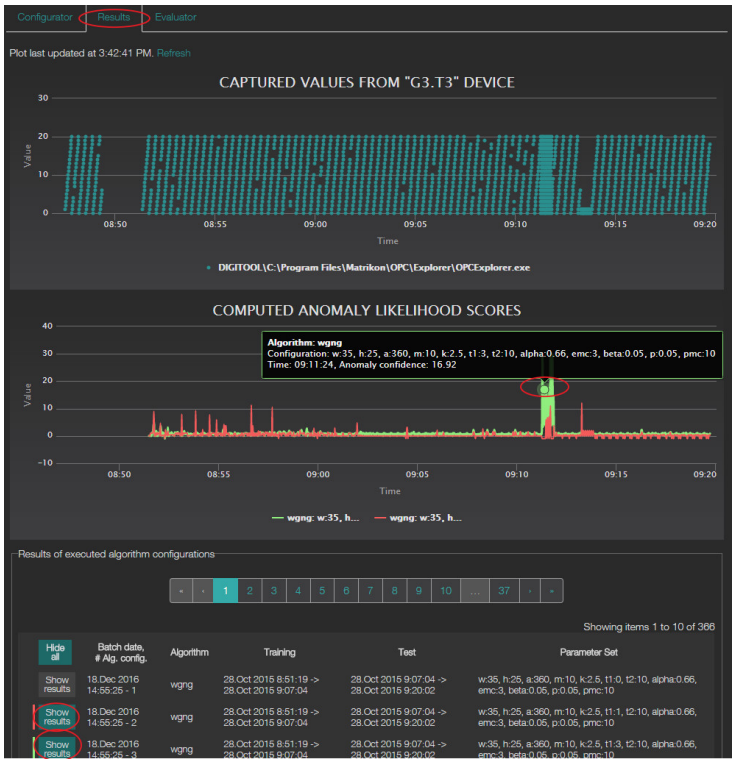

Figure 6.7: Completion of Task 7

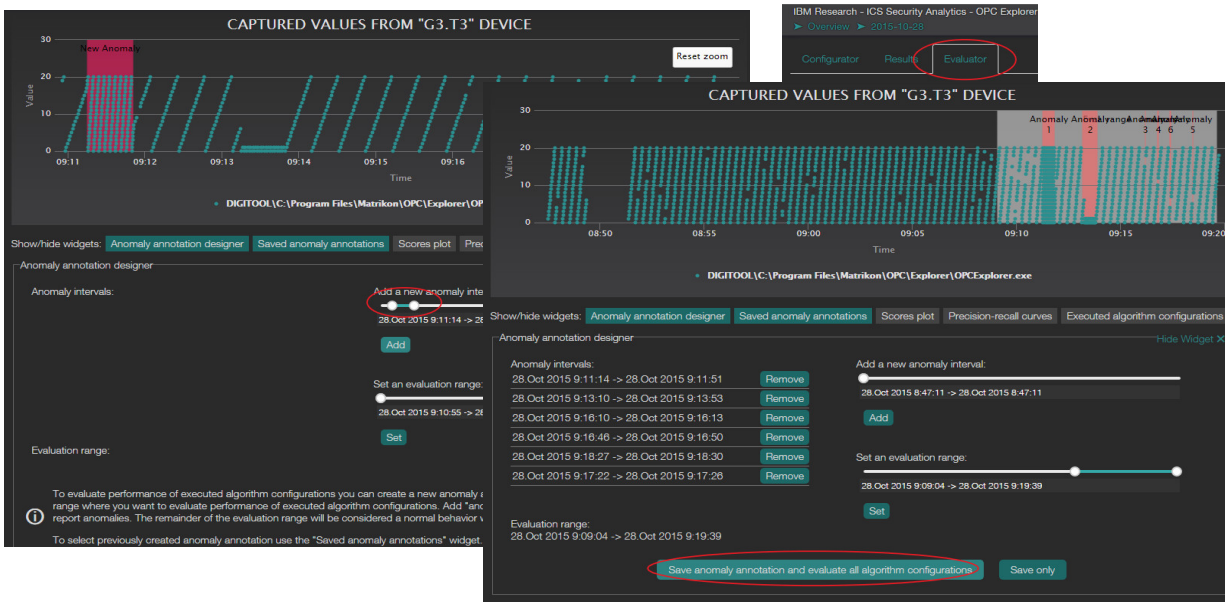

Figure 6.8: Completion of Tasks 8,9,10 and 11

6.5.2.9 Tasks 12 and 13

To complete tasks 12 and 13, participants need to click on the “Show results” button next to any of the configurations in the table that is displayed after the evaluation of algorithms was executed. Plots of scores and Precision recall are displayed as shown in Figure 6.9.

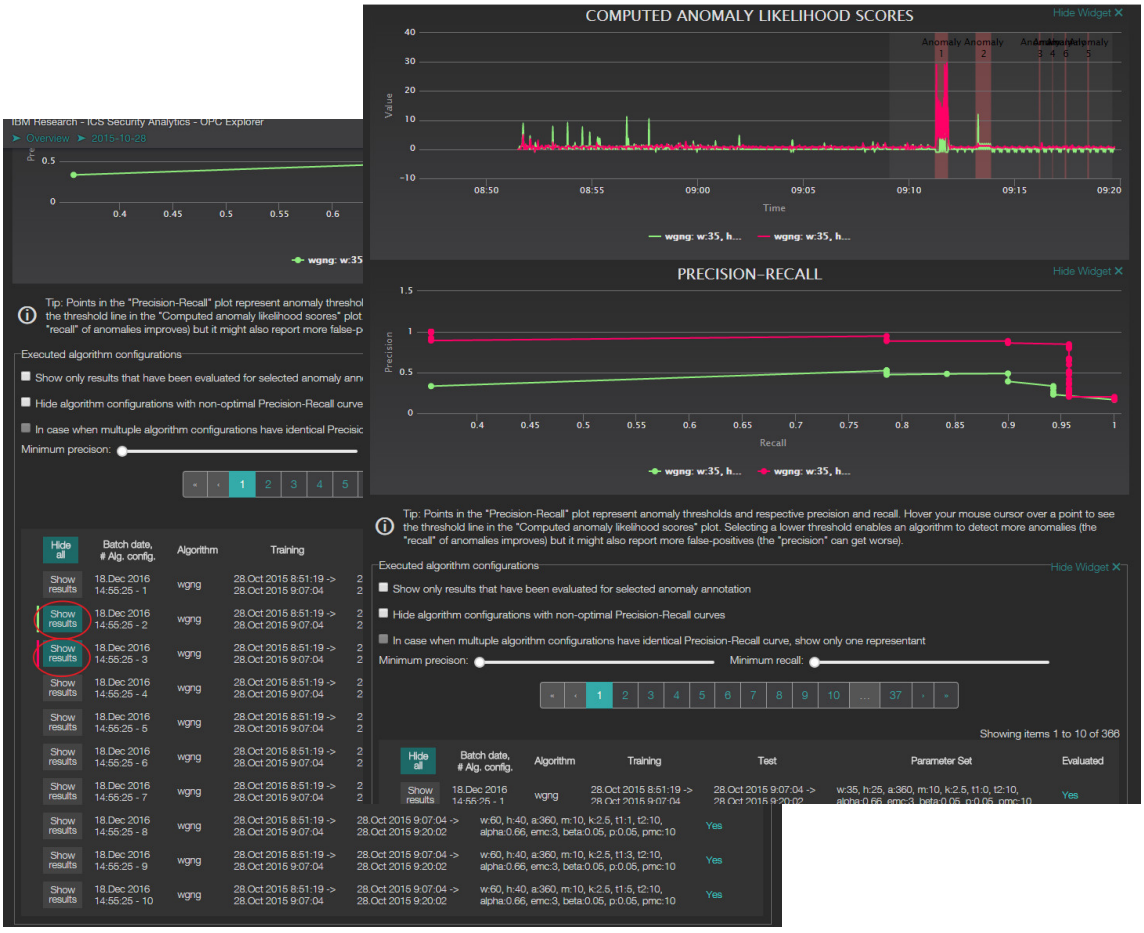

Figure 6.9: Completion of Tasks 12 and 13

6.5.2.10 Tasks 14 and 15

To complete tasks 12 and 13, participants should use the “minimum precision” slider and optionally the “Hide algorithm configurations with non-optimal Precision-Recall curves” to find the best configuration that meets the minimum precision. Participants should display the precision recall curve for the best solution. Figure 6.10 shows the user interface elements that should be used.

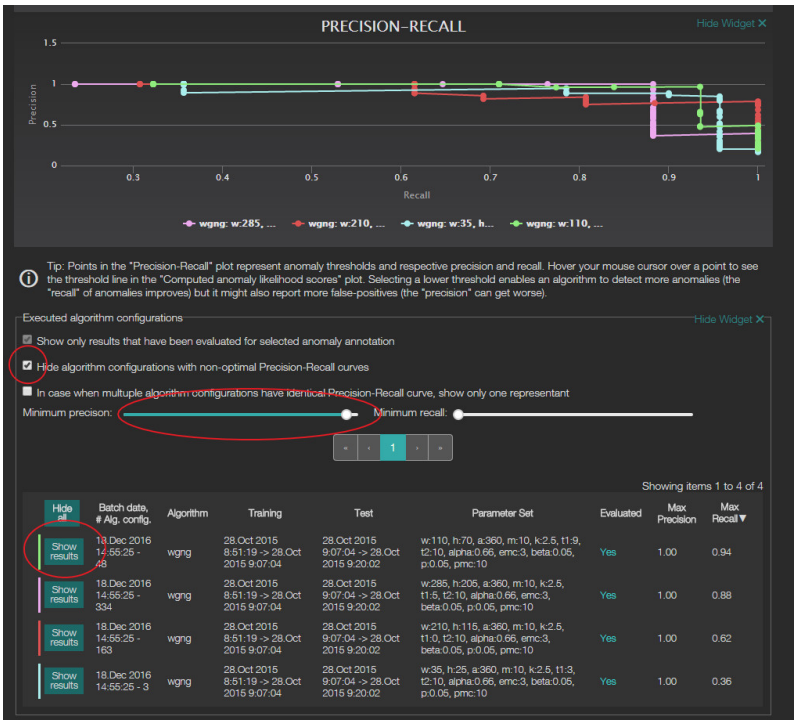

Figure 6.10: Completion of tasks 14 and 15

6.5.2.11 Tasks 16 to 20

Completion of the tasks 16, 17, 18, 19 and 20 where user configures A-node algorithm is similar to configuration of the Wgng algorithm.

6.5.2.12 Task 21: Compare "A-node" to "Wgng"

To complete Task 21, user should set the minimum recall slider to maximum value and explore the precision recall curve and scores plot of the top row in the Executed algorithm configurations table as shown on Figure 6.11.

6.6 Testing Conditions

6.6.1 Participant Group Characterization

Five people participated in the user testing. All the participants were affiliated with IBM Research Zurich either as employees or as interns. Three of the participants had some or good knowledge about machine learning and anomaly detection. Two participants did not have knowledge in the domain.

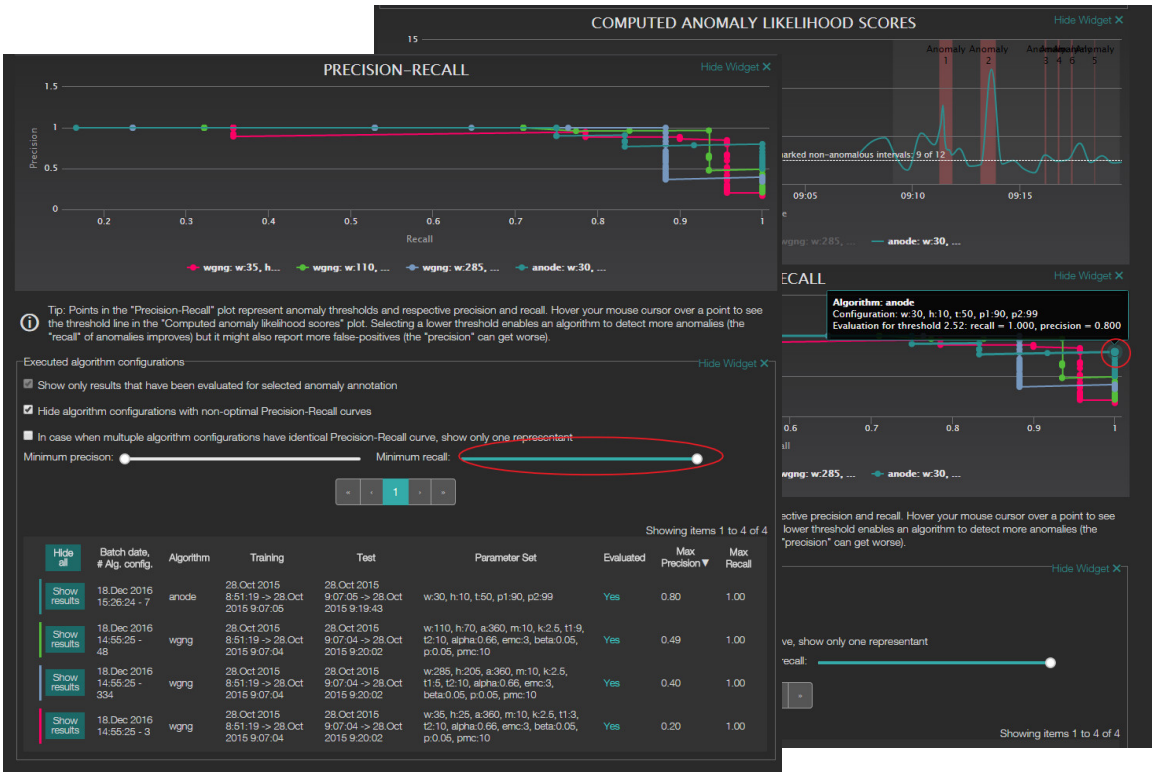

Figure 6.11: Completion of Task 21

6.6.2 Conditions During Testing

The test sessions with participants took place on December 19, 2016 from 13:30 to 18:30. The conditions for testing were good and the environment was quiet. The last two participants seemed tired. This might be due to the late time of testing and also because they had worked during the day before participating in the test.

6.7 Sessions with Participants

This section describes sessions with participants and the insights that I have learned from individual sessions. During the sessions I was present as moderator only. To assess the sessions properly I have watched the captures of the sessions where I assume a role of an observer. The complete transcripts of voice recorded during the sessions as well as logs of observations that I noted while analyzing the captures are provided in Appendix E.

6.7.1 Participant 1

The first participant took 43 minutes to complete all tasks (time for filling in questionnaires is not included).

6.7.1.1 Evaluation of the Pre-Test Questionnaire

Based on the answers from the pre-screen questionnaire, the first participant has a degree or pursues a specialization in computer science, she has extensive knowledge in using software tools for machine learning tasks, however she evaluates her experience with anomaly detection in time series as basic only. She lists image processing, machine learning, big data analysis, mathematics and precision-recall as her computer skills.

6.7.1.2 Session Assessment

The participant was able to complete all tasks. During the completion of the tasks she took occasional pauses to understand the user interface. Section E.1.1 contains a log of all observations. Few situations deserve to be description here in more detail.

A problem occurred when the participant had to label the anomaly in Task 9. The software offers a slider user interface element to select the range. The participant had trouble selecting an anomaly that has a very short duration. At the level of zoom of the plot the slider did not offer sufficient precision. It is possible to increase the precision of the sliders by zooming the plot of captured data but this was not apparent to the participant. The moderator had to advise the participant to use the zoom feature of the plot. Figure 6.12 illustrates the problem. The left side of the figure shows the initial situation when the plot of captured values is not zoomed. In this situation the slider start and end values match the earliest and latest time of the plot. Since the plot shows over 2 hours of data, it is difficult to select few seconds with the slider below the plot. Figure 6.12 also shows the solution to the problem. Users are supposed to select a part of the captured values plot using the mouse (in the figure depicted by red line with arrows). Then the plot zooms as is shown in the right area of the figure. Slider start and end values are updated to match the plot zoom and users can select shorter intervals easily. However, this solution was not apparent to the user and the problem occurred in slightly different forms with all the participants of this user testing. This problem should be solved in the future either by making the zoom possibility of the plot more apparent or by providing a different type of interaction elements to select anomaly intervals. One option would be to select intervals directly in the plot and include zoom and markup buttons in the corner of the captured values plot.

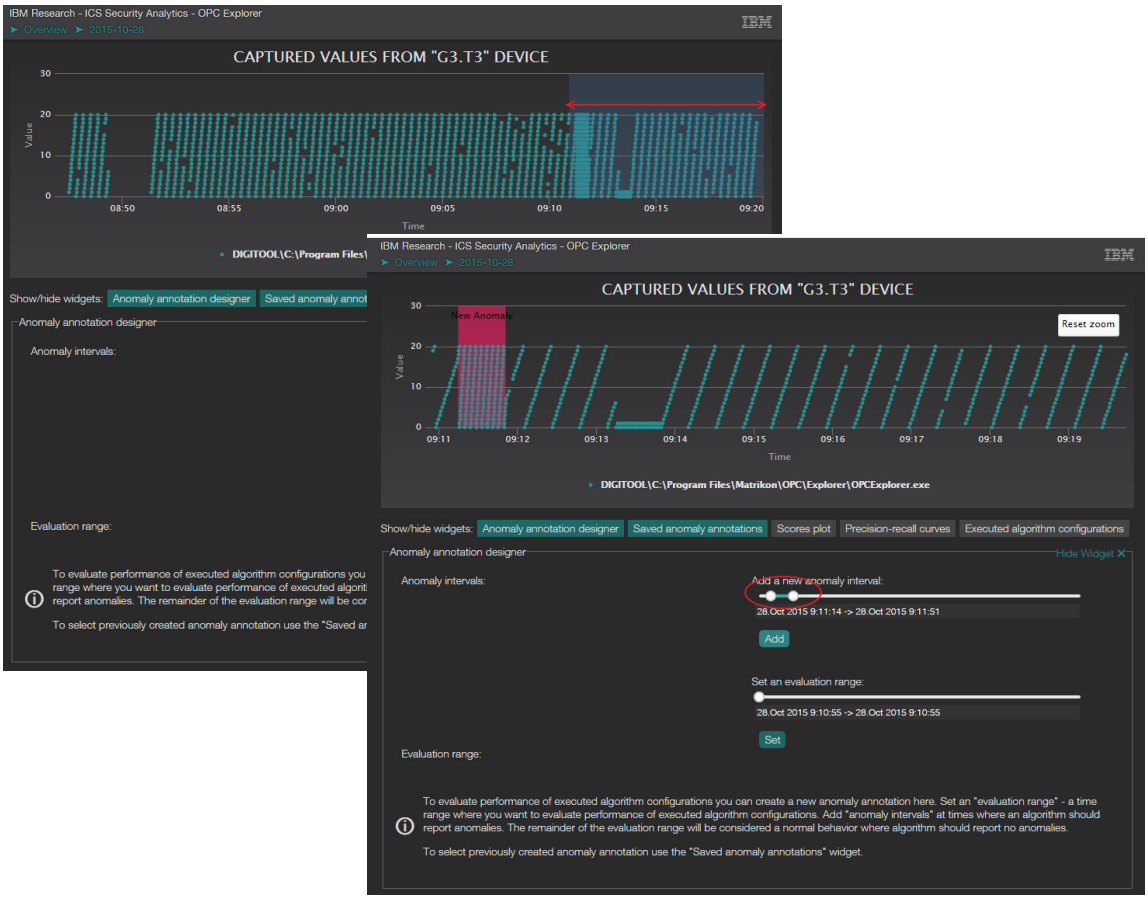

Figure 6.12: UI problem - slider precision

A task that deserves attention is the last task, where the user is supposed to compare results of A-node and Wgng algorithms with a set minimum recall. At this point the user has some experience working with the platform. The participant was able to make use of the scores and precision-recall interactive plots to compare the algorithms. Even though the Wgng algorithm provides better precision, participant did not like the high fluctuations of the Wgng algorithm scores and preferred A-node algorithm. This shows number of things. The participant was able to compare precision-recall curves and scores, but considered smoothness of a scores more important than the better precision. The A-node scores are, in fact, smoother because the two compared configurations of algorithms have different window size parameter. A-node scores have less frequent values than the Wgng scores. The situation is presented in Figure 6.13.

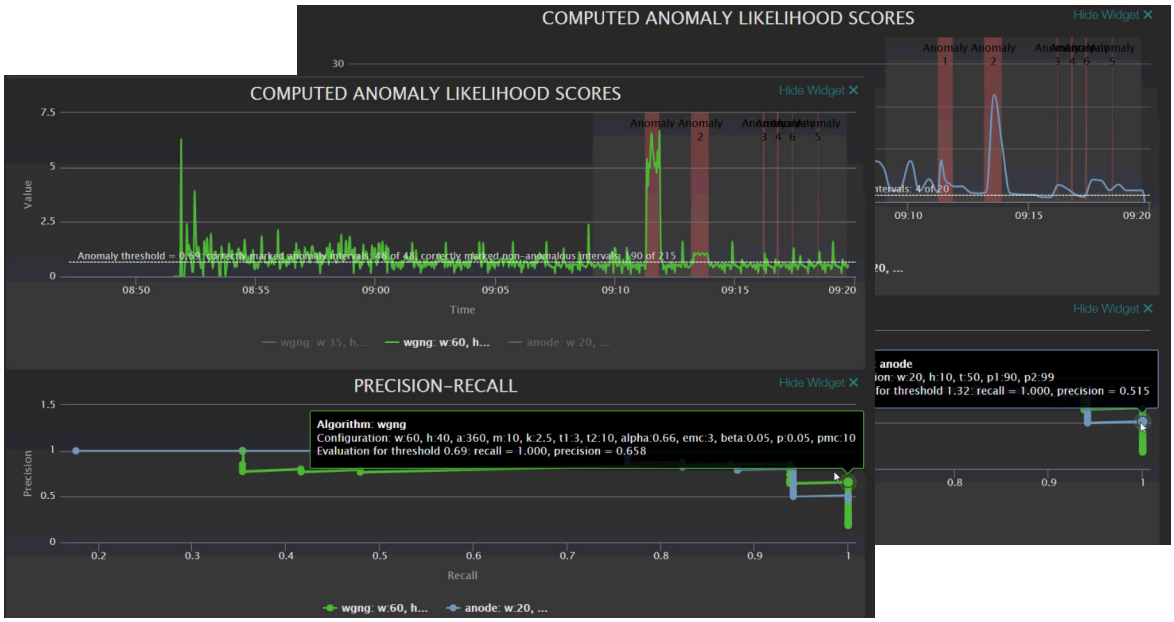

Figure 6.13: Comparing Wgng and A-node

6.7.1.3 Evaluation of the Post-Test Questionnaire

Based on the marked options in the post-test questionnaire, the participant believes that she would not be able to complete some tasks without help (problem with setting up anomaly intervals precisely). The participant considers the tasks and working with the software slightly complicated. She has no suggestions to improve the software and agrees that the information guide provided before the tasks was helpful.

Answers to the questions from the set B (Table 6.3) of the post-test questionnaire are discussed in Section 546.8.3.

6.7.2 Participant 2

The second participant took 32 minutes to complete all tasks (time for filling in questionnaires is not included).

6.7.2.1 Evaluation of The Pre-Test questionnaire

The second participant is a specialist in computer science, he has only basic knowledge in using software tools for machine learning tasks and he evaluates his experience with anomaly detection in time series as intermediate. He listed programming, networking and security as his computer skills.

6.7.2.2 Session Assessment

The participant completed all the tasks exceptionally well and fast. The only greater problem that occurred is the problem at Task 9 – same as with the first participant. I had to explain how zooming the plot allows to markup anomaly intervals with necessary precision. Additionally, the user would appreciate a clear feedback from the platform when pressing the "Evaluate all algorithm outputs with the anomaly markup" button in Task 12 and had stumbled shortly before finding where he can see the results of the algorithm evaluation. This issue could be solved by a better guidance from the platform, e.g. a notification or highlighting the table with computed evaluations,

6.7.2.3 Evaluation of the Post-Test Questionnaire

In the post-test questionnaire the participant marked that he appreciated the help of the moderator but did not mark an option saying that it was necessary. The participant considers the tasks slightly complicated. He claims that he found his way around the software easily. In the open ended questions, he suggests to make the zooming feature of the captures values plot more apparent. It was not clear to him when the experiments (i.e. computation of anomaly likelihood scores) were finished. He suggests adding progress bar to indicate completion of the experiments. He would appreciate the options to markup anomalies directly in the plot. He did not like that the right knob of the anomaly interval slider is blocking the left knob from moving to the right. I.e. when the participant wants to mark an anomaly that starts after the current position of the end slider knob, the user has to first move the end slider knob to the right. The referred slider can be seen in Figure 6.12.

Answers to the questions from the set B (Table 6.3) of the post-test questionnaire are discussed in Section 546.8.3.

6.7.3 Participant 3

The third participant completed all tasks in 37 minutes (time for filling in questionnaires is not included).

6.7.3.1 Evaluation of The Pre-Test questionnaire

The answers from the pre-screen questionnaire inform us that the third participant has or pursues a specialization in computer science, he has extensive knowledge in using software tools for machine learning tasks, as well as with anomaly detection in time series. He lists

communication and computer networks, security, embedded systems and digital signal processing as his computer skills. The participant is a developer of the IBM's analysis and forensics platform but has not worked with the final version of the assistant platform.

6.7.3.2 Session Assessment

The third participant had a good understanding of the platform features and the domain of ICS security. He was able to complete all the tasks and provided commentary.

In Task 14 where he is supposed to use the "Hide algorithm configurations with non-optimal Precision-Recall curves" filtering option (illustrated in Figure 6.10) he comments that he does not see a description of what "optimal" means in this context. He further comments that one could optimize precision-recall curves by their area. An explanation tooltip that would explain this filtering option in detail could solve this problem.

6.7.3.3 Evaluation of the Post-Test Questionnaire

The participant answered the post-test questionnaire as follows. He appreciated the help of the moderator but does not say it is necessary. The participant considers the tasks slightly complicated. He claims that working with the software was slightly complicated. He suggests to remove complexity of the platform by introducing sensible defaults. He suggests adding filtering functionality to filter the displayed results by algorithm. Finally, he appreciated the information guide provided before the test.

Answers to the questions from the set B (Table 6.3) of the post-test questionnaire are discussed in Section 546.8.3.

6.7.4 Participant 4

The fourth participant took 27 minutes to complete all tasks (time for filling in questionnaires is not included).

6.7.4.1 Evaluation of The Pre-Test questionnaire

The fourth participant is specialized in applied sciences other than computer science. He has intermediate knowledge in using software tools for machine learning tasks and he evaluates his experience with anomaly detection in time series as basic. He lists "C++", "Python", "Ruby" and "OpenCU" as his computer skills.

6.7.4.2 Session Assessment

The fourth completed all Tasks except Task 12. He skipped this step by accident. Since the completion of task is not required to complete further tasks, he was able to finish the rest of the tasks.

The participant, similar to the first and second participant, also had problem selecting short anomaly intervals in the Task 9 and was not able to use the plot zoom option to his benefit. I described this problem in detail in Section 6.7.1.2.

The participant was also confused when applying the minimum precision and minimum recall filtering options in Task 14. He did not realize that by setting minimum precision value to 0.95 the algorithm configuration results are sorted by maximum achievable recall. Since, he did not correctly understand what "Max precision" and "Max recall" columns represent, which can be seen in Figure 6.10. To solve this problem, an information icon with a pop-up tooltip could be added to the column headers, however to understand this feature well, one should probably understand the underlying sorting principle.

6.7.4.3 Evaluation of the Post-Test Questionnaire

The fourth participant stated in the post-test questionnaire that he felt more tense in an environment with a moderator. The participant considered the tasks slightly complicated. He claims that working with the software was easy. He suggests adding keyboard shortcuts functionality and would like to type hours, minutes and seconds to setup the anomaly intervals. He appreciated the information guide provided.

Answers to the questions from the set B (Table 6.3) of the post-test questionnaire are discussed in Section 546.8.3.

6.7.5 Participant 5

The third participant completed all tasks in 35 minutes (time for filling in questionnaires is not included).

6.7.5.1 Evaluation of The Pre-Test questionnaire

The last participant is specialized in computer science; he has extensive knowledge in using software tools for machine learning tasks but low or no experience with anomaly detection in time series. He lists "C++", "Python" and "Matlab" as his computer skills.

6.7.5.2 Session Assessment

The participant was able to complete all tasks. When setting training and test intervals in Task 3, he tried to select intervals with the mouse cursor in the captured data plot instead of using the training and test interval slider knobs. The screen on which this situation happened can be seen in Figure 6.3. He found out that this way does not work and managed to complete the task on his own.

The fifth participant also struggled to set the short anomaly intervals in Task 9 and was instructed by moderator to zoom the plot. The problem was described in Section 6.7.1.2.

The participant got into tricky situation in the Task 21. Normally, if all previous tasks are completed precisely by instructions there should be both A-node and Wgng configurations shortlisted as is the case in Figure 6.13. However, in this case, he probably set some time intervals differently and no A-node configuration was present in the list of optimal precision-recall curve algorithms. Nevertheless, even in this situation he was able to understand the results correctly. This means that he understood well how the system works and interpreted even such an unexpected result correctly.

6.7.5.3 Evaluation of the Post-Test Questionnaire

The participant answered the post-test questionnaire as follows. He claims that he tried less than he normally would. He appreciated the help and found the tasks easy. The participant considers working with the software slightly complicated. He suggests a better feedback mechanism to make the system more user friendly. Finally, he claims that he did not need the information guide because he had understood the concepts already before reading it.

Answers to the questions from the set B (Table 6.3) of the post-test questionnaire are discussed in Section 6.8.3.

6.8 Results

Results of testing with helps to validate the quality and usability of the implemented assistant platform. The participants went through the variety of tasks which helped pinpoint issues and also provided insight about how well users understand the concepts of the platform. This section summarizes the user interface issues, recapitulates feedback from users and presents the results of the *Perceived Usefulness and Ease of Use* [46] questions – the second part of the post-test questionnaire (listed in Table 6.3).

6.8.1 User Interface Issues

6.8.1.1 Using slider to select anomaly interval

Using sliders to select short anomalies in the anomaly annotation widget is cumbersome. The slider does not provide sufficient precision. For better resolution, users can zoom to the area around anomaly using the zoom function of the captured values plot. However, this is not apparent.

6.8.1.2 Feedback after clicking on the "Evaluate all algorithm outputs with the anomaly markup" button

When users click on the "Evaluate all algorithm outputs with the anomaly markup" the "Executed algorithm configurations" table shows up. The table contains the evaluations of algorithm outputs. However, this is not apparent enough. A message should be added that explains where to look for the computed evaluations.

6.8.1.3 Unclear progress notification about computation of algorithm

When users of the assistant platform execute algorithm configurations, the new results appear in the table of the "Results tab". Since the new results are added at the end of the table it is not apparent whether all computation have been finished.

6.8.1.4 "Hide algorithm configurations with non-optimal Precision-Recall curves" filtering option

When selecting the "Hide algorithm configurations with non-optimal Precision-Recall curves" filtering option it is not apparent what "optimal" means and what this function does.

6.8.1.5 Confusion about minimum precision and minimum recall filter options

After setting the minimum recall or minimum precision filter, the rows in the "Executed algorithm configurations" table are sorted by the metric that has none or lower filtering value set. For instance, when users set minimum precision to 0.9, I assume that they want to find configuration with precision at least 0.9 and recall as high as possible. Thus the table is automatically sorted by recall, while meeting the precision minimum as well. This might not be apparent.

6.8.2 Suggestions from participants

Participants suggested:

- adding progress bar to indicate completion of experiments
- marking-up anomalies directly in the captured values plot
- introducing sensible defaults for the parameters
- adding filtering of the results based on the algorithm and parameters
- keyboard shortcuts
- ability to type-in start and end time of anomalies
- better feedback

6.8.3 Post-Test questionnaire Results Summary

Table 6.5 contains answers to the *Perceived Usefulness and Ease of Use* [46] questions. For each question, participants could choose a number from one to seven, one meaning *not likely* and seven meaning *likely*. Some participants used the whole scale, but others only numbers five to seven. To address this, Table 6.6 presents the normalized scores where I stretch values given by user. The lowest score is mapped to 1 and their highest score is mapped to 7.

Sum and mean statistics in both tables show following statements as likable:

- It would be easy for me to become skillful at using the system (mean 6.2).
- Using the system in my job would enable me to accomplish tasks more quickly (mean 6.0).
- Using the system would make it easier to do my job (mean 5.8).

The following statement received the lowest scores in both tables:

- I would find the system to be flexible to interact with (mean 4.6).

6.8.4 Summary

All users were able to complete the presented tasks and successfully compare and identify fitting algorithm configurations for the given anomaly annotation of the data. They were able to reduce number of algorithm configurations from four hundred to four using the functions of the platform. The problems that occurred were mostly related to user-interface glitches but did not prevent assistant platform from helping users with the specified goals. The collected data can be used to fix the issue and improve the usability of the platform. Users stated that it would be easy for them to become skillful at using the system and the platform would help them to accomplish tasks in their jobs quicker and easier. Improving flexibility of the system should be considered, since this point received the lowest score.

#	Question	Participant					Sum	Mean
		1	2	3	4	5		
B1	Using the system in my job would enable me to accomplish tasks more quickly	7	6	6	4	7	30	6.0
B2	Using the system would improve my job performance	6	6	3	3	7	25	5.0
B3	Using the system in my job would increase my productivity	6	6	5	3	7	27	5.4
B4	Using the system would enhance my effectiveness on the job	7	6	2	3	7	25	5.0
B5	Using the system would make it easier to do my job	7	7	6	2	7	29	5.8
B6	I would find the system useful in my job	6	7	6	2	7	28	5.6
B7	Learning to operate the system would be easy for me	5	7	7	5	4	28	5.6
B8	I would find it easy to get the system to do what I want it to do	5	6	6	4	5	26	5.2
B9	My interaction with the system would be clear and understandable	6	7	5	3	5	26	5.2
B10	I would find the system to be flexible to interact with	6	5	2	4	6	23	4.6
B11	It would be easy for me to become skillful at using the system	7	6	7	5	6	31	6.2
B12	I would find the system easy to use	6	6	5	6	5	28	5.6

Table 6.5: Results and statistics for Post-Test set B

#	Question	Participant					Sum	Mean
		1	2	3	4	5		
B1	Using the system in my job would enable me to accomplish tasks more quickly	7.0	4.0	5.8	4.0	7.0	27.8	5.6
B2	Using the system would improve my job performance	4.0	4.0	2.2	2.5	7.0	19.7	3.9
B3	Using the system in my job would increase my productivity	4.0	4.0	4.6	2.5	7.0	22.1	4.4
B4	Using the system would enhance my effectiveness on the job	7.0	4.0	1.0	2.5	7.0	21.5	4.3
B5	Using the system would make it easier to do my job	7.0	7.0	5.8	1.0	7.0	27.8	5.6
B6	I would find the system useful in my job	4.0	7.0	5.8	1.0	7.0	24.8	5.0
B7	Learning to operate the system would be easy for me	1.0	7.0	7.0	5.5	1.0	21.5	4.3
B8	I would find it easy to get the system to do what I want it to do	1.0	4.0	5.8	4.0	3.0	17.8	3.6
B9	My interaction with the system would be clear and understandable	4.0	7.0	4.6	2.5	3.0	21.1	4.2
B10	I would find the system to be flexible to interact with	4.0	1.0	1.0	4.0	5.0	15.0	3.0
B11	It would be easy for me to become skillful at using the system	7.0	4.0	7.0	5.5	5.0	28.5	5.7
B12	I would find the system easy to use	4.0	4.0	4.6	7.0	3.0	22.6	4.5

Table 6.6: Results and statistics for Post-Test set B - Normalized

Chapter 7

Conclusion

In this thesis I have proposed and evaluated features and user interface design of an interactive assistant platform for tuning behavior based algorithms in the domain of Industrial Control Systems (ICS). The platform assists users with generating combinations of parameters for anomaly detection modules. The platform offers an interactive user interface. It enables users to discover the best anomaly detection module and configuration set for a provided anomaly annotation. The evaluation with test users shows that the platform is capable of helping users to shortlist the best performing configuration sets and anomaly detection modules. The platform enables users to find an anomaly detection module fitting their expectations of anomalous and normal behavior classification. It provides interactive elements for fast exploration of precision recall curves of algorithms setups and allows user to shortlist the possible configurations using filtering and sorting options.

Future development of the platform is possible in few directions: adding more anomaly detection modules to the platform, expanding the feature set of the platform, automating the exploration of algorithm parameter space or intelligent learning of the platform based on the inputs of expert users.

Bibliography

[1] T. Macaulay and B. L. Singer, Cybersecurity for industrial control systems: SCADA, DCS, PLC, HMI, and SIS, CRC Press, 2011.

[2] R. R. R. Barbosa, Anomaly detection in SCADA systems: a network based approach, Enschede: University of Twente, 2014.

[3] K. Stouffer, J. Falco and K. Scarfone, "Guide to industrial control systems (ICS) security," *NIST special publication,* vol. 800, no. 82, pp. 16-16, 2011.

[4] R. L. Krutz, "Securing SCADA systems," *John Wiley & Sons,* 2005.

[5] ICS CERT, "Monthly Monitor October December," 2012.

[6] V. M. Igure, S. A. Laughter and R. D. Williams, "Security issues in SCADA networks," *Computers & Security,* vol. 25, no. 7, pp. 498-506, 2006.

[7] Ponemon Institute, "State of Security, Study of Utilities and Energy Companies," 2011.

[8] W. Beckner, "NRC Information Notice 2003-14: Potential Vulnerability of Plant Computer Network to Worm Infection," *United States Nuclear Regulatory Commission,* 2003.

[9] N. Falliere, L. O. Murchu and E. Chien, "W32. stuxnet dossier," *White paper, Symantec Corp., Security Response,* vol. 5, 2011.

[10] B. Genge, D. A. Rusu and P. Haller, "A connection pattern-based approach to detect network traffic anomalies in critical infrastructures," in *Proceedings of the Seventh European Workshop on System Security*, 2014.

[11] F. Schuster, A. Paul and H. König, "Towards learning normality for anomaly detection in industrial control networks," in *IFIP International Conference on Autonomous Infrastructure, Management and Security*, 2013.

[12] J. Bigham, D. Gamez and N. Lu, "Safeguarding SCADA systems with anomaly detection," in *International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security*, 2003.

[13] S. Cheung, B. Dutertre, M. Fong, U. Lindqvist, K. Skinner and A. Valdes, "Using model-based intrusion detection for SCADA networks," in *Proceedings of the SCADA security scientific symposium*, 2007.

[14] N. Goldenberg and A. Wool, "Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems," *International Journal of Critical Infrastructure Protection,* vol. 6, pp. 63-75, 2013.

[15] M. Caselli, E. Zambon and F. Kargl, "Sequence-aware intrusion detection in industrial control systems," in *Proceedings of the 1st ACM Workshop on Cyber-Physical System Security*, 2015.

[16] S. Ponomarev and T. Atkison, "Industrial Control System Network Intrusion Detection by Telemetry Analysis," *IEEE Transactions on Dependable and Secure Computing,* vol. 13, pp. 252-260, 2016.

[17] R. R. R. Barbosa, R. Sadre and A. Pras, "Flow whitelisting in SCADA networks," *International journal of critical infrastructure protection,* vol. 6, pp. 150-158, 2013.

[18] A. Amrein, V. Angeletti, A. Beitler, M. Német, M. Reiser, S. Riccetti, M. P. Stoecklin and A. Wespi, "Security intelligence for industrial control systems," *IBM Journal of Research and Development,* vol. 60, pp. 13-1, 2016.

[19] M. Demarne, "Industrial Control System Security," 2016.

[20] "R project," [Online]. Available: https://www.r-project.org/.

[21] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Network Modeling Analysis in Health Informatics and Bioinformatics,* vol. 5, pp. 1-16, 2016.

[22] ABB, "REC 501 RP 570 Protocol Description," 1997.

[23] "Profibus," [Online]. Available: http://www.profibus.com/.

[24] "Modus," [Online]. Available: http://www.modbus.org/.

[25] "Distributed Network Protocol," [Online]. Available: https://www.dnp.org/.

[26] "OPC Foundation," [Online]. Available: https://opcfoundation.org/.

[27] "TCPdump," [Online]. Available: https://www.tcpdump.org/.

[28] B. Fritzke and others, "A growing neural gas network learns topologies," *Advances in neural information processing systems,* vol. 7, pp. 625-632, 1995.

[29] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006.

[30] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research,* vol. 13, pp. 281-305, 2012.

[31] J. S. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, 2011.

[32] "ReactJS," [Online]. Available: https://facebook.github.io/react/.

[33] "ReduxJS," [Online]. Available: http://redux.js.org/.

[34] M. S. Mikowski and J. C. Powell, "Single Page Web Applications," *B and W,* 2013.

[35] "Webpack," [Online]. Available: https://webpack.github.io/.

[36] "Babel," [Online]. Available: https://babeljs.io/.

[37] "JSX," [Online]. Available: https://jsx.github.io/.

[38] "ECMA Internaional," [Online]. Available: http://www.ecma-international.org/.

[39] "SocketIO," [Online]. Available: http://socket.io/.

[40] "Mongodb," [Online]. Available: https://www.mongodb.com/.

[41] "RabbitMQ," [Online]. Available: https://www.rabbitmq.com/.

[42] "Python," [Online]. Available: https://www.python.org/.

[43] "Pandas," [Online]. Available: http://pandas.pydata.org/.

[44] "NumPy," [Online]. Available: http://www.numpy.org/.

[45] "Docker," [Online]. Available: https://www.docker.com/.

[46] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly,* pp. 319-340, 1989.

[47] "Microsoft," [Online]. Available: https://www.microsoft.com.

[48] "Camstudio," [Online]. Available: http://camstudio.org/.

[49] "Google Chrome," [Online]. Available: https://www.google.com/chrome.

[50] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE,* vol. 78, pp. 1464-1480, 1990.

[51] "NodeJS," [Online]. Available: https://nodejs.org/en/.

Appendix A

List of abbreviations

ICS	Industrial Control Systems
SCADA	Supervisory Control and Data Acquisition
DCS	Distributed Control Systems
PCS	Process Control Systems
IBM	International Business Machines
RP-570	RTU Protocol based on IEC 57 part 5-1 (present IEC 870) version 0 or 1
OPC	OLE for Process Control
WGNG	Windowed Growing Neural Gas
RTU	Remote Terminal Unit
PLC	Programmable Logic Controller
MTU	Master Terminal Unit
HMI	Human-Machine Interface
API	Application Programming Interface
HTTP	Hypertext Transfer Protocol

Appendix B

Contents of CD

<dir> doc – contains thesis document source

<dir> pdf – contains PDF file

Appendix C

User interface screenshots

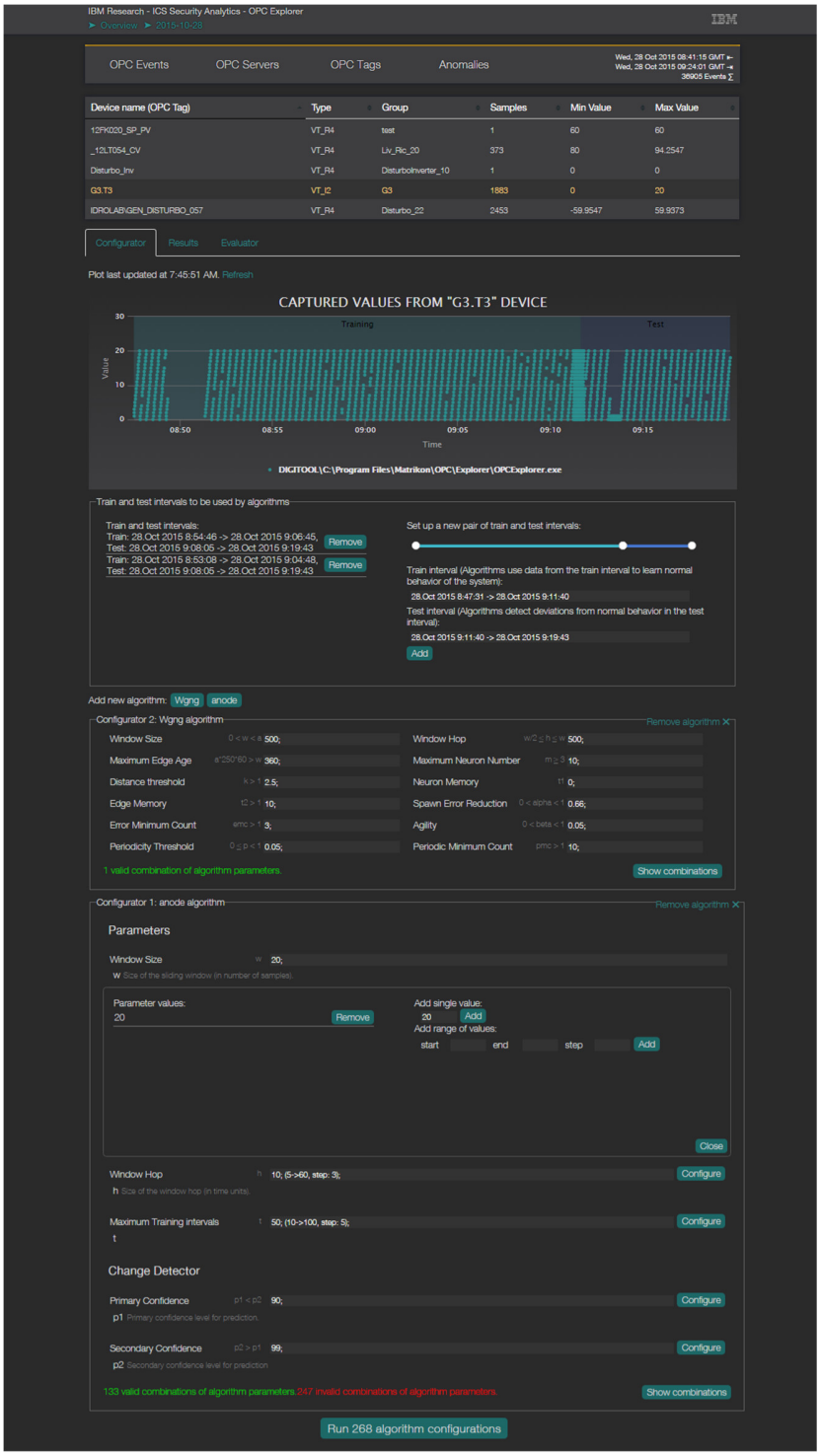

Figure C.1: UI - Configurator tab

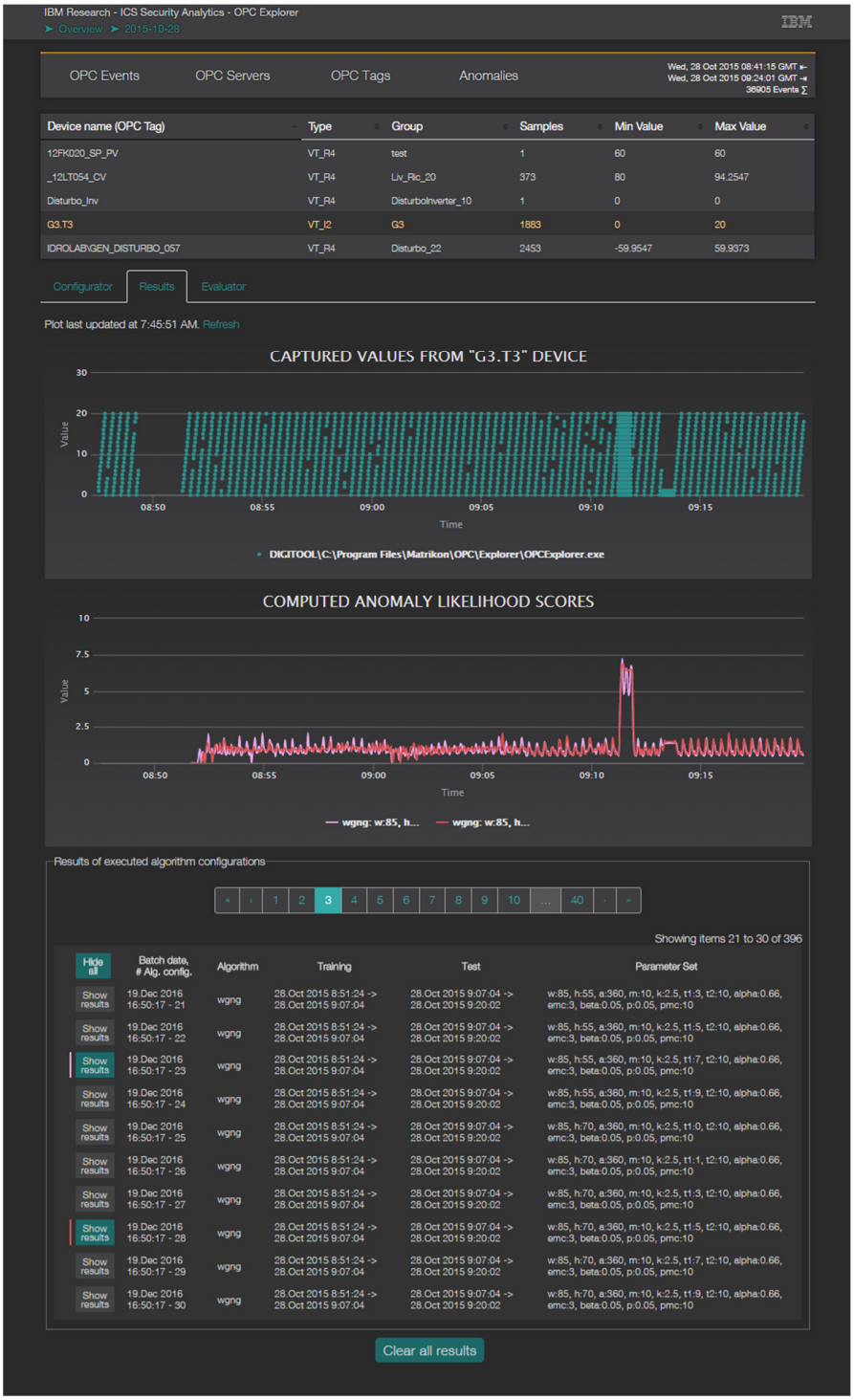

Figure C.2: Results Tab

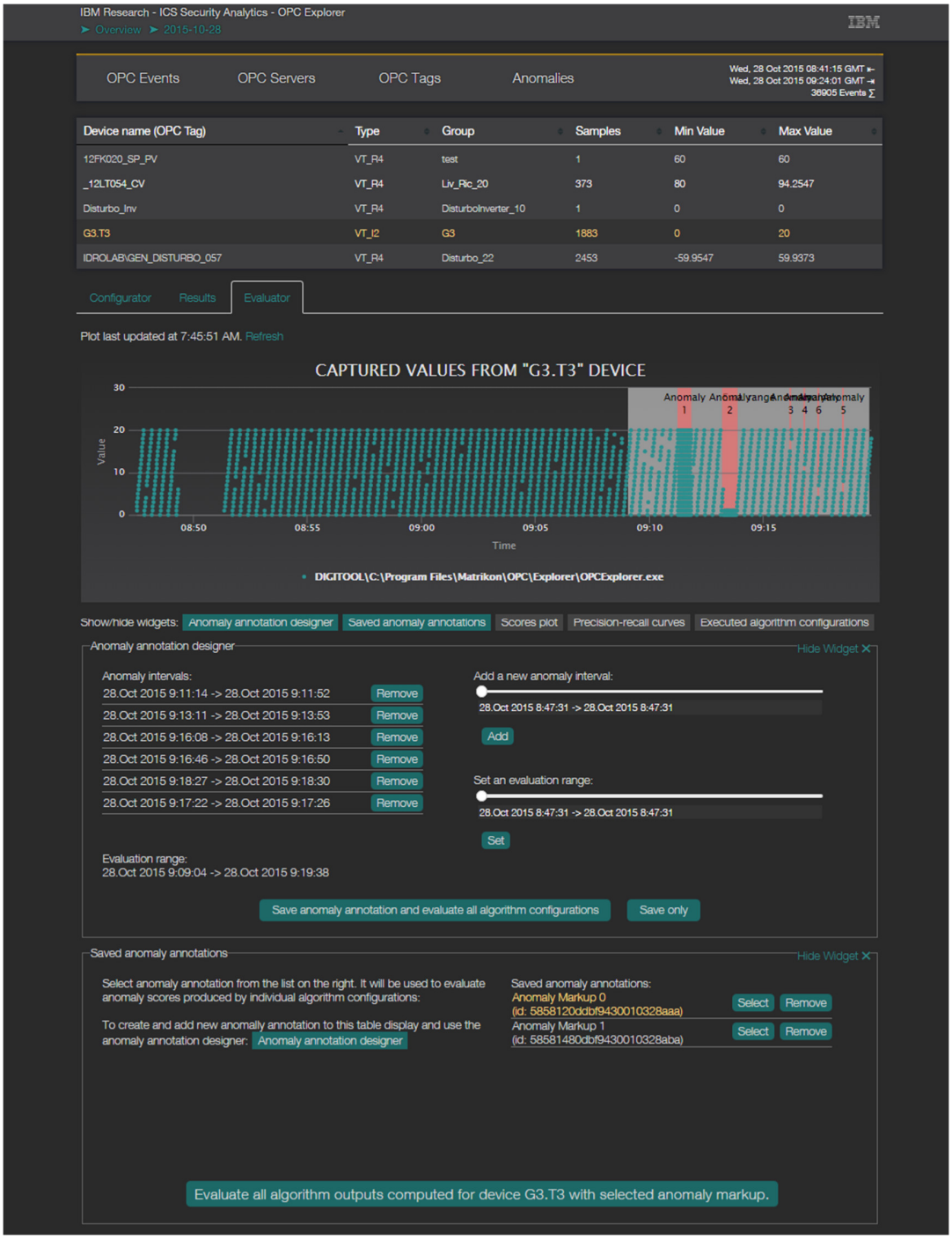

Figure C.3: Evaluator view - anomaly annotation setup

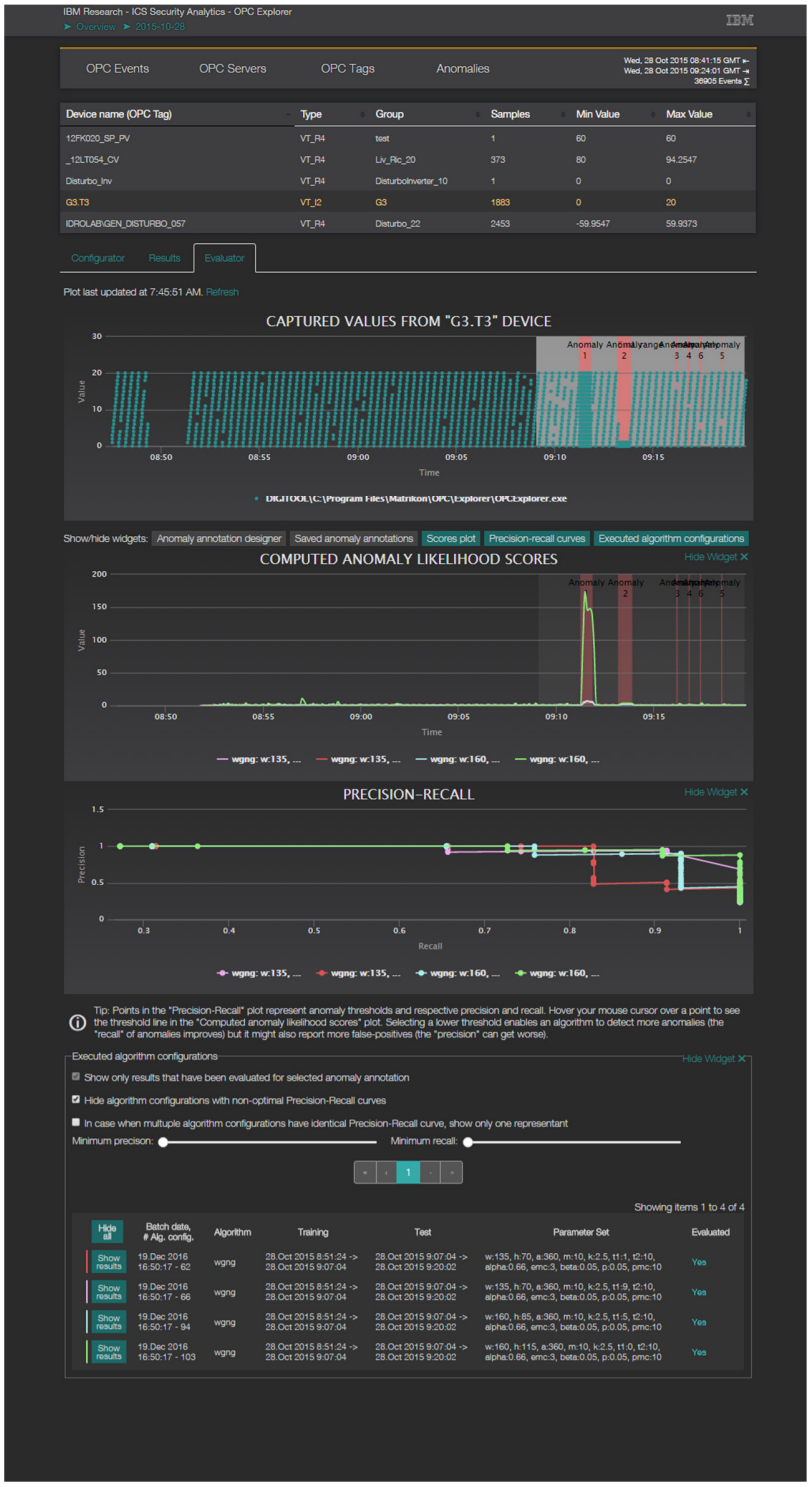

Figure C.4: Evaluator view – exploring evaluated algorithm configurations

Appendix D

User Testing Questionnaires

This appendix presents following documents:

- Document D.1: Screener questionnaire
- Document D.2: Pre-test questionnaire
- Document D.3: Post-test questionnaire - first page
- Document D.4: Post-test questionnaire - second page
- Document D.5: Information Guide for UI-Test Participants - first page
- Document D.6: Information Guide for UI-Test Participants - second page
- Document D.7: Information Guide for UI-Test Participants - third page

Participant ID:

#	Question	Answer options
1	Do you currently pursue or have you previously completed a higher education degree (university/university of applied sciences/other post-secondary education)?	Yes No I cannot answer
2	What is your experience reading mathematical plots (graphs)?	High (From university Math courses or similar) Intermediate (From secondary school Math courses) Basic (no formal education) Lower than previous options or none
3	How comfortable do you feel using modern interactive websites (for example any of following: gmail.com, maps.google.com, google drive, drop box/box/iCloud or purchasing airplane tickets online)?	High (I feel comfortable and I use them more than once a week) Intermediate (I can use them when I need to, with an occasional hesitation) Basic (I use them less than once a week, I find them complicated) Lower than previous options or none
4	Do you have a work experience or have you completed a university course in statistics, machine learning, statistical learning or anomaly detection?	Yes No I cannot answer
5	What is your command of English?	Very high (I can understand technical English well) High (I can understand a movie in English) Intermediate (I can speak in future and past tense) Basic (I can say my name, where I come from) Lower than previous options or none)

Document D.1: Screener questionnaire

Participant ID:

#	Question	Answer options
1	What is your degree specialization / main professional expertise?	Applied sciences - with focus on computer science or similar Applied sciences – other (mechanical/civil engineering/…) Natural sciences – physical sciences (physics/chemistry/...) Natural sciences – life sciences (physics/chemistry) Social sciences Formal sciences (mathematics/logic/statistics) Other
2	What is your expertise in using software (MATLAB/R/Python/similar) for machine learning or similar tasks?	Extensive (In projects and over 10 hours) Intermediate (One project/ coursework, over 3 hours) Basic (Less than 3 hours of using such software) Lower than previous options or none
3	What is your expertise in using software for detecting anomalies in time series?	Extensive (Projects and over 10 hours) Intermediate (Coursework, over 3 hours) Basic (Less than 3 hours of using such software) Lower than previous options or none
4	List some examples of your computer skills	Write briefly:

Document D.2: Pre-test questionnaire

Participant ID:

Please fill in the survey carefully. In case you do not understand a question ask the assistant. In questions A1, A2 and A3 multiple choices can be selected. You can write additional comments to any question.		
#	**Question**	**Answer options**
A1	How do you think a presence of an assistant affected you?	No effect I tried harder than I would normally try I tried less than I would normally try I was more tense than usually It made completing the tasks more complicated I appreciated the help I would not be able to complete some tasks without help Other – please write down:
A2	I considered the tasks	Easy Slightly complicated Complicated I did not understand the instructions
A3	How would you describe your experience working with the software?	I found my way around the software easily Working with software was easy Working with software was slightly complicated Working with software was complicated I could not find my way around the software, I was confused Other – please write down:
A4	Do you have any suggestions for improving the software?	Yes – please write down: No
A5	Do you have any suggestions for new functionality of the software?	Yes – please write down: No
A6	Please evaluate following statement: Information guide provided before the testing helped me in completing the tasks.	Strongly Agree Agree Undecided / Neutral Disagree Strongly Disagree

Document D.3: Post-test questionnaire - first page

Participant ID:

For the following questions: If your occupation is not related to the functionality of the software, please assume that your job is to select algorithm and parameters to use for detecting anomalies in time series.											
#	**Question**	**unlikely**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Likely**	**Not applicable**
B1	Using the system in my job would enable me to accomplish tasks more quickly										
B2	Using the system would improve my job performance										
B3	Using the system in my job would increase my productivity										
B4	Using the system would enhance my effectiveness on the job										
B5	Using the system would make it easier to do my job										
B6	I would find the system useful in my job										
B7	Learning to operate the system would be easy for me										
B8	I would find it easy to get the system to do what I want it to do										
B9	My interaction with the system would be clear and understandable										
B10	I would find the system to be flexible to interact with										
B11	It would be easy for me to become skillful at using the system										
B12	I would find the system easy to use										

Document D.4: Post-test questionnaire - second page

Information for the participant of user testing

The software that you will work with allows you to use anomaly detection algorithms. It permits you to configure such algorithms and evaluate their performance. In this guide you can find information about topics related to anomaly detection in time series. It can help you understand the terminology and functions of the software.

Time series data

The software allows you to work with captured data from various industrial devices. The data is in time series format, i.e. it contains values that were recorded at certain times. Such data may contain anomalies. For an industrial client it is useful to be able to detect such anomalies.

Figure 1 shows how a captured time series data might look. First 17 seconds (the left half of the plot) represent a normal behavior of a device. The rest of the presented data contains anomalous values.

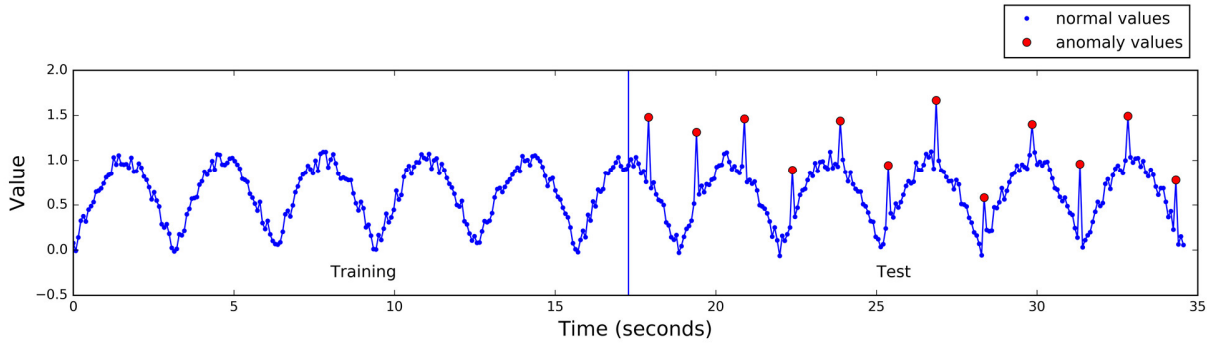

Figure 1

Training and test data

Anomaly detection algorithms that are included in the software need an example of normal behavior to learn characteristics of a device when it operates normally. This data is called a training data or a training interval. After a training process, algorithms can analyze new data (test data) and report whether it contains deviations from normal behavior (anomalies).

Evaluating performance of anomaly detection algorithms - Recall and Precision

The software uses recall and precision statistics to evaluate performance of algorithms. An algorithm with high recall capability is able to identify majority of anomalies. High precision means that an algorithm is returning accurate results – majority of the values that algorithm marks as anomaly are true anomalies. Examples presented on the next page should allow you to gain intuition about precision and recall. You can read following mathematical definitions.

Recall (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n).

$$R = \frac{T_p}{T_p + F_n}$$

Precision (P) is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p).

$$P = \frac{T_p}{T_p + F_p}$$

Document D.5: Information Guide for UI-Test Participants - first page

Consider an example of a simple algorithm that marks every value greater than a certain threshold as an anomaly. Figure 2 illustrates such a situation with a threshold value of 1.313. Such an algorithm is able to detect 58 percent of anomalies (all values above the threshold). A recall of such an algorithm is 58% (R = 0.58). All values that are marked as anomaly are truly anomalies. Thus, a precision of such an algorithm is 100% (P = 1).

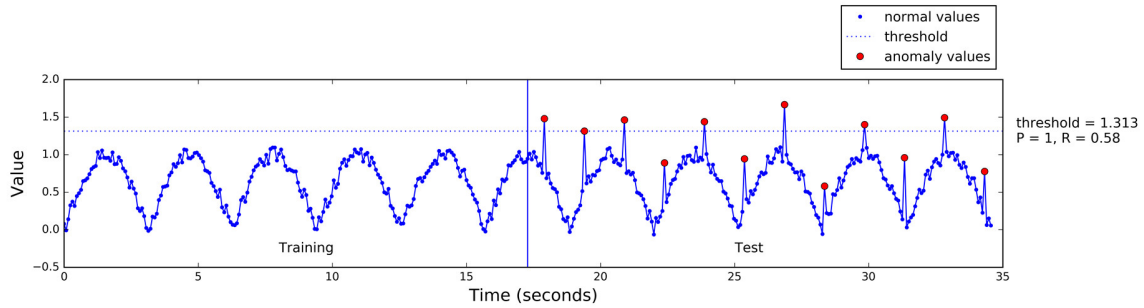

Figure 2

Consider a version of such an algorithm where a threshold would be 0.581. Again, all the values above the threshold are marked as anomaly by this algorithm. The situation is depicted in Figure 3. This version of the algorithm would be able to find all anomalies (recall is 100%, R = 1). However, it would also mark many normal values as an anomaly. Only 8 % of values that algorithm marks as anomaly are true anomalies. Thus precision is 8% (P = 0.08).

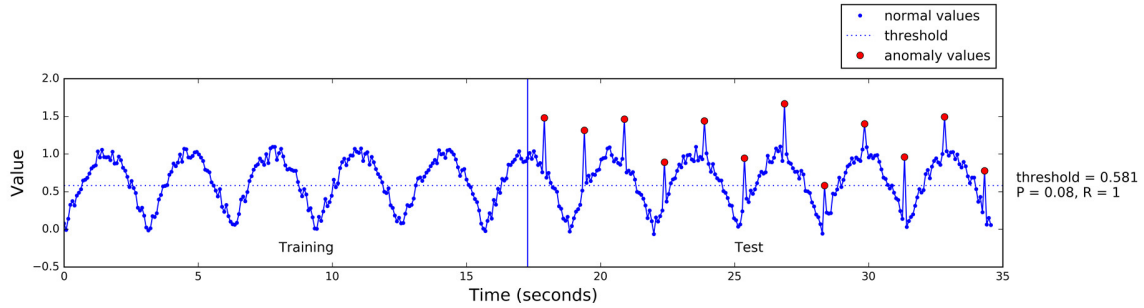

Figure 3

Precision-recall curves

It is possible to try various values of the threshold. However, this algorithm is not able to reach 100% precision and 100% recall at the same time with any threshold. To reach higher recall, we need to sacrifice precision and vice-versa. All possible pairs of precision and recall that an algorithm can reach can be plotted as a curve called *precision-recall curve*. Figure 4 shows a precision-recall curve of the presented simple algorithm for the presented time series.

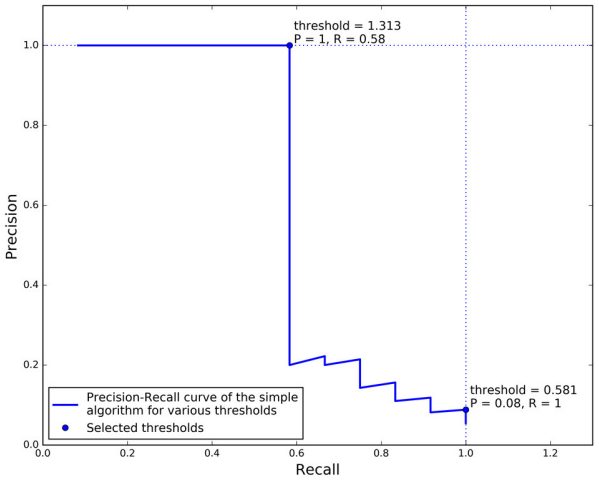

Figure 4

Document D.6: Information Guide for UI-Test Participants - second page

A more advanced algorithm that models a normal behavior as an area between bands is illustrated on Figure 5. All the values outside the grey area would be marked as an anomaly by the algorithm. Such an algorithm would be able to recognize all the anomalies (recall is 100%, R = 1). At the same time, 100% of the values that were marked are truly anomalies. Thus, precision is 100%, P = 1.

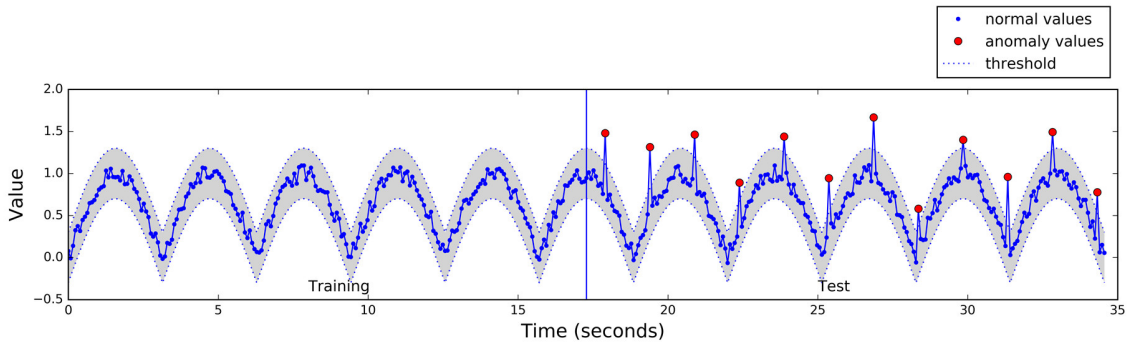

Figure 5

Figure 6 shows a comparison of precision-recall curves of simple and advanced algorithm. For the presented time series data, the advanced algorithm is able to reach better precision and recall values.

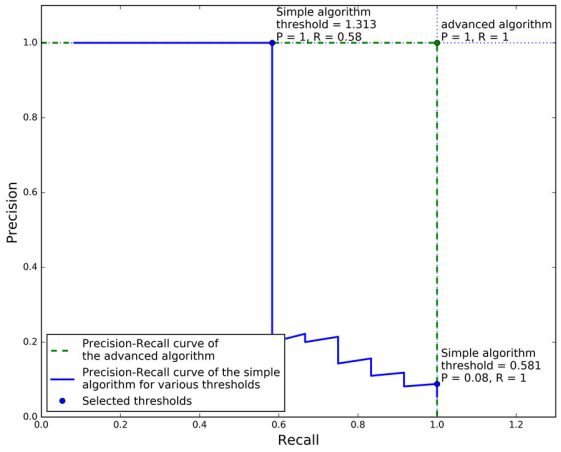

Figure 6

Annotating time series

In the time series that was presented in Figure 1, anomalies are already marked up by red dots. The software that you will work with requires that anomalies are marked up by an operator of the software (you). The instructions on which parts of the time series you should annotate as anomalies will be included in the list of tasks.

Summary of terms

Time series – values together with times when they were recorded
Anomalies – deviations from normal behavior
Training data/interval – data/time interval used by anomaly detection algorithm to learn about normal behavior
Test data/interval – data/time interval where algorithm marks anomalies
Recall – capability of algorithm to mark majority of anomalies
Precision – accuracy of marked anomalies
Precision-recall curve – plot of pairs of precision and recall for possible setups of an algorithm
Anomaly annotation – a markup of anomalies created by user

Thank you and questions

Thank you for reading through this guide. Now is a good time to ask any questions if anything is not clear.

Document D.7: Information Guide for UI-Test Participants - third page

Appendix E

Data from user testing

E.1 Participant 1

E.1.1 Log for Participant 1

Time	Task		Detail	Note
0:01:58	Task 1		Success	Successfully finished the task.
0:02:27	Task 2		Observer	Comment of participant: participant finds the interface professional looking
0:04:32	Task 2		Success	Successfully finished the task.
0:06:12	Task 2		Observer	Participant has trouble seeing the user interface element for setting up the training and test interval because the browser zoom is set to 150%. Moderator instructs the participant to scroll down to see more elements of the user interface.
0:08:25	Task 3		Success	Successfully finished the task.
0:09:40	Task 4		Observer	Adds window size range. Doesn't delete the default value.
0:10:53	Task 4		Observer	Adds window hop. Deletes default values in both window size and hop.
0:11:25	Task 4		Observer	Tries to add values for Neuron memory as range.
0:11:51	Task 4		Observer	Adds values for Neuron memory correctly as single values.
0:12:10	Task 4		Success	Successfully finished the task.
0:13:14	Task 5		Success	Successfully finished the task.
0:13:26	Task 6		Success	Successfully finished the task.
0:15:11	Task 7		Success	Successfully finished the task.
0:15:34	Task 8		Success	Successfully finished the task.
0:18:12	Task 9		Observer	Has trouble setting up time intervals for anomaly annotation precisely. Moderator needs to explain how to do it.
0:29:54	Task 9		Success	Successfully finished the task.
0:31:34	Task 10		Success	Successfully finished the task.
0:31:57	Task 11		Success	Successfully finished the task.
0:32:39	Task 12		Success	Successfully finished the task.

0:33:24	Task 13		Success	Successfully finished the task.
0:34:50	Task 14		Success	Successfully finished the task.
0:34:58	Task 15		Success	Successfully finished the task.
0:35:47	Task 16		Success	Successfully finished the task.
0:36:45	Task 17		Observer	Successfully finishes the task but then for some reason adds the values from the range as single values as well.
0:37:15	Task 18		Success	Successfully finished the task.
0:37:28	Task 19		Success	Successfully finished the task.
0:39:46	Task 20		Observer	Operator has to explain where to find the user interface element that evaluates results of A-node algorithm with the previously created anomaly annotation.
0:40:03	Task 20		Success	Successfully finished the task.
0:45:25	Task 21		Success	Successfully finished the task.

E.1.2 Transcript for Participant 1

Time	Task	Speaker	Notes
0:00:37		Moderator	These are the instructions. If you have any questions you should ask now because from now you should work on your own.
0:00:57		Participant	Can you give me brief introduction?
0:00:58		Moderator	Yes, you have a list of tasks and you should try to complete all the tasks. They will tell you what to do with the system.
0:01:10		Moderator	Is that ok?
0:01:11		Participant	Yes.
0:01:15		Participant	Is this a test?
0:01:16		Moderator	Well actually I am the one being tested.
0:01:33		Participant	I think that I am not allowed to use this tab.
0:01:35		Moderator	Exactly, that leads to other parts of the system which I was not working on so you will be testing someone else's work.
0:01:47		Moderator	You can use the keyboard and a mouse.
0:02:06	Task 1	Moderator	Another important thing. Try all the time to think out loud and comment what you are doing. If you find or consider something etc.
0:02:27		Participant	I like the interface. I find it professional, it is simple and I can easily see what are the training and test data.
0:03:00	Task 2	Participant	These are the time series, right?
0:03:06		Moderator	I should not answer this. You should try to figure out and if it is not possible then I can help you.
0:03:16		Participant	That's OK.
0:03:19		Moderator	One more thing. Every time you start doing the task, read the number of the task.
0:03:31		Participant	I have completed task 1 and I have to understand the captured values (task 2).

0:03:46		**Moderator**	And again you can think out loud, say what you think etc.
0:04:16		**Participant**	I am trying to understand why there are some missing information and what is the pattern here. I don't know if this is what I should do.
0:04:32		**Moderator**	Yes, I think that is enough for task 2. You can go to the next task, can you read it?
0:04:44	Task 3	**Participant**	Reads the task.
0:05:45		**Participant**	I am trying to figure out how to understand how to setup the timing interval for the test data.
0:06:12		**Moderator**	You have scroll down.
0:07:00		**Participant**	I want to set the time interval.
0:07:19		**Moderator**	You should try to set it up very close the values that are on the paper.
0:08:30		**Moderator**	You can go to the next task.
0:08:34	Task 4	**Participant**	Reads task.
0:10:08		**Participant**	I am doing task 4, I am now adding the value for the four parameters.
0:10:55		**Participant**	I have to remove the default value.
0:11:39		**Moderator**	Actually if you look closely at the numbers its 0, 1,..
0:12:10		**Participant**	I think I finished task 4.
0:12:13		**Moderator**	Yes, can you read the next one?
0:12:16	Task 5	**Participant**	Reads task.
0:12:50		**Participant**	I want to check if I can find here only valid combinations or the invalid? Yes, here are all of them.
0:13:11		**Moderator**	Good, you can move to the next task.
0:13:20	Task 6	**Participant**	Reads task... RUN
0:13:38	Task 7	**Participant**	There was a results tab here.
0:14:06		**Participant**	Now I am trying to find show results button.
0:14:14		**Participant**	Oh here it is.
0:14:38		**Participant**	So I did the task 7. I have examined the results. In my opinion, this is Gaussian distribution but I am not sure, maybe it is normal distribution and these are some anomalies.
0:15:16	Task 8	**Participant**	Reads task.
0:15:36	Task 9	**Participant**	Reads task.
0:17:40		**Participant**	The only think I can remark is that I have trouble setting up the seconds correctly... maybe it can be more sensible?
0:19:43		**Moderator**	Try to set it up as precisely as possible.
0:19:55		**Participant**	Yes, maybe it is my hand or I need to make the window larger in order to set this.
0:20:12		**Moderator**	Can you maybe try to fix the previous one as well?
0:20:14		**Participant**	Yes.
0:20:16		**Moderator**	Just the last one maybe.
0:21:39		**Moderator**	Zoom out please.
0:21:41		**Participant**	Yes, I don't know.
0:21:43		**Moderator**	Go back to 150%.

0:22:25		Moderator	You can use the mouse to zoom the plot. Not the scroll though. Like the select part. Like drag and select the interval.
0:22:51		Participant	I understood.
0:24:14		Participant	It's interface problem. The testers often complain about it. Changing the UI.
0:24:24		Moderator	What do you mean?
0:24:26		Participant	They are working with a UI and at some point several changes are made. They have some problems in the first days to get used with it. This is my problem too because I don't know how the system is working and I think I prefer by setting this (Add a new anomaly interval) than the plot. Because I don't know exactly how it functions. So I didn't complete the task 9.
0:25:10		Moderator	You have to complete it to continue. So you can use the zoom in the plot and then the sliders are zoomed according to the plot so you can then in better detail select the times using the sliders.
0:25:40		Participant	Ok.
0:26:43		Moderator	You can zoom in the plot. For example, try to zoom to the second third of the plot.
0:27:13		Participant	So the second third ...
0:27:18		Moderator	And now you can try to set the anomaly.
0:27:35		Participant	Now I understand.
0:29:55	Task 10	Participant	Reads task.
0:30:51		Participant	I have to reset the range.
0:31:08		Moderator	Yes, that's good.
0:31:35	Task 11	Participant	Reads task.
0:32:06		Moderator	What is the next task?
0:32:07	Task 12	Participant	Reads task.
0:32:40	Task 13	Participant	Reads task.
0:33:05		Participant	We can't have high precision and high recall because this isn't ideal case. I think this is the best that I can have.
0:33:28	Task 14	Participant	Reads task.
0:33:43		Participant	But I think this should...
0:33:49		Moderator	Yes, you can reset the ...
0:33:51		Participant	It selects only the best configurations that have good precision and recall.
0:34:16		Moderator	Can you try to view them?
0:34:18		Participant	We have to press "show results".
0:35:00	Task 15	Participant	Reads task.
0:35:08	Task 16	Participant	Reads task.
0:35:48	Task 17	Participant	Reads task.
0:37:13	Task 18	Participant	Reads task.
0:37:20	Task 19	Participant	Reads task.
0:37:37	Task 20	Participant	Reads task.
0:37:50	Task 19	Moderator	Can you read 19 again?

0:37:52		Participant	Reads task 19 again... I realized that I didn't wait one minute.
0:38:18		Participant	My question is when do I know that the algorithm is finished?
0:38:39		Moderator	Yes, you can go to the next step now.
0:38:45	Task 20	Participant	Reads task 20 again.
0:39:46		Moderator	You have to open saved annotations.
0:40:01		Participant	This one, right?
0:40:11		Moderator	Good, you can go to the next task.
0:40:18	Task 21	Participant	Reads task.
0:41:18		Participant	So maybe I should "show results" for both.
0:42:46		Participant	I think for the A-node the performance is better given that that we have plateau for the precision.
0:43:22		Participant	Now I understand, so it was the minimum recall so here is when the recall is 100%.
0:44:25		Moderator	So what are you thinking?
0:44:28		Participant	I think that for Wgng the results are fluctuating, they don't have steady representation. And for the A-node there is a steady representation although it reaches some peaks. I can see that at some point we have a plateau like here. Data don't fluctuate a lot in A-node case compared to Wgng.
0:45:25		Moderator	Ok, good.

E.2 Participant 2

E.2.1 Log for Participant 2

Time	Task	Detail	Note
0:06:30	Task 1	Success	Successfully finished the task.
0:07:12	Task 2	Success	Successfully finished the task.
0:09:15	Task 3	Success	Successfully finished the task.
0:12:22	Task 4	Success	Successfully finished the task.
0:13:13	Task 5	Success	Successfully finished the task.
0:13:19	Task 6	Success	Successfully finished the task.
0:15:34	Task 7	Success	Successfully finished the task.
0:16:48	Task 8	Success	Successfully finished the task.
0:18:51	Task 9	Observer	Moderator needs to explain how it is possible to achieve better precision when setting up anomaly intervals (plot of captured data needs to be zoomed).
0:20:43	Task 9	Success	Successfully finished the task.
0:21:27	Task 10	Success	Successfully finished the task.
0:21:39	Task 11	Success	Successfully finished the task.
0:25:05	Task 12	Success	Successfully finished the task.
0:26:30	Task 13	Success	Successfully finished the task.
0:28:40	Task 14	Success	Successfully finished the task.

0:29:02	Task 15	**Success**	Successfully finished the task.
0:29:17	Task 16	**Success**	Successfully finished the task.
0:30:24	Task 17	**Success**	Successfully finished the task.
0:30:33	Task 18	**Success**	Successfully finished the task.
0:32:10	Task 19	**Success**	Successfully finished the task.
0:34:12	Task 20	**Success**	Successfully finished the task.
0:37:33	Task 21	**Success**	Successfully finished the task.

E.2.2 Transcript for Participant 2

Time	Task	Speaker	Notes
0:05:17		**Moderator**	So the next part is you doing the tasks. I shouldn't be assisting you with the tasks. I will constantly remind you to think out loud and to comment on the system. And yea so basically do the tasks and every time in the beginning of the task just read the number of the task and read the task out loud so that I know which task are you working on.
0:05:43		**Participant**	Ok.
0:05:47		**Moderator**	You can use this machine.
0:06:11	Task 1	**Participant**	Reads the task.
0:06:22		**Participant**	Well I guess the device name – device list is here. So I guess that's selected.
0:06:38	Task 2	**Participant**	Reads task.
0:06:46		**Participant**	Based on the background this is a time series. There is some time information on the X axis and each one of these points represents a value at that time.
0:07:16	Task 3	**Participant**	Reads task 3.
0:09:05		**Participant**	I guess that's task 3.
0:09:24	Task 4	**Participant**	Reads task.
0:10:06		**Participant**	I don't understand. Where is a step? I see it now.
0:12:24		**Moderator**	Good, you can go to the next task.
0:12:35	Task 5	**Participant**	Reads tasks.
0:12:41		**Participant**	There are 366 valid combinations and 888 invalid combinations.
0:13:09		**Participant**	There are some restrictions of the algorithm in order to get accepted.
0:13:13	Task 6	**Participant**	Reads task.
0:13:33		**Participant**	I like it because it uses machine learning. I think it's working.
0:13:43	Task 7	**Participant**	Reads task.
0:13:49		**Participant**	So this is the results: "show results"
0:14:06		**Participant**	This might be the anomalies region...
0:14:20		**Participant**	I will try to see a different one. I can add different ones at the same time. How many can I see at once?
0:14:50		**Participant**	Every one of these is outputting a different anomaly value, likelihood value. Some of them look to be more precise than others because here it's attracting an anomaly where doesn't look like

			there is one from this dataset. This one seems to be found by more of these combinations.
0:15:12		**Participant**	Let me see another one. There are results with different estimations of the anomalies for these time series with different combinations of the parameters
0:15:35	Task 8	**Participant**	Reads task.
0:16:39		**Participant**	The evaluation range should be the whole range.
0:16:49	Task 9	**Participant**	Reads task.
0:18:49		**Participant**	These are getting progressively more difficult to select.
0:18:51		**Moderator**	You can zoom the plot and then the sliders will zoom as well.
0:18:57		**Participant**	How do I zoom the plot? Is there a zoom function here? Do I select it somehow?
0:19:33		**Participant**	This is much better.
0:20:45	Task 10	**Participant**	Reads task.
0:20:51		**Participant**	I need to reset the zoom.
0:21:28	Task 11	**Participant**	Reads task.
0:21:46	Task 12	**Participant**	Reads task.
0:22:00		**Moderator**	Can you do a comment? Think out loud about what you're thinking.
0:22:03		**Participant**	I ran the "save the annotation and evaluate", so I did this and now I have to press "explore" to explore the evaluations for some of the configurations. I explore some evaluations but I don't know where that is.
0:22:28		**Participant**	I am looking for "computed evaluation" button. Is it this?
0:22:47		**Participant**	Is it "Evaluate all algorithm outputs with the anomaly markup"? I guess so. (clicks)
0:23:07		**Participant**	Did something happen? It's not working. Maybe I should select this thing? Maybe it's this tab but that was already visible.
0:23:31		**Participant**	I am not really sure what I am supposed to be looking for. I am not sure if the task is just not clear to me. I don't know what I am supposed to look at.
0:23:50		**Moderator**	Before you create some configurations and you run them, the algorithms will give you some results. Now you evaluated them and you should see the evaluations soon.
0:24:09		**Participant**	I guess these ones, right? Ok.
0:24:10		**Participant**	If I start clicking on these then it's showing me those curves – precision and recall for each one. I want all of them, but there's too many.
0:24:51		**Participant**	It's also highlighting here the score of one of these steps which is interesting because these help me to figure out which one of these parameters perform the best. So I understand now.
0:25:07	Task 13	**Participant**	Reads task.
0:25:32		**Participant**	You are asking for 95 %. I need to bring this up to 95. You can sort them by recall. This has already 0.94 so this is the on... but other ones are still selected.

0:26:05		**Participant**	For sure this is good. I think this is the best parameter, because the second one is 0.92
0:26:28		**Participant**	These are the parameters.
0:26:32		**Moderator**	Can you go to the next task?
0:26:37	Task 14	**Participant**	Reads task.
0:26:59		**Participant**	"Hide algorithm..." what's it getting rid of?
0:27:13		**Participant**	It only leaves me five of them.
		Participant	Obviously it's removing some of them from the list, but I am not sure which ones it's getting rid of.
0:27:51		**Participant**	I am not sure what optimal means here.
0:28:01		**Participant**	For some reason these two that I selected that have precision 1 and recall 0.84 and 0.82 are being hidden when I hide non-optimal curves. But I am not sure what an optimal curve means.
0:28:21		**Participant**	Can I zoom in this?
0:28:39		**Participant**	I am not sure if I understand what that does.
0:28:42		**Moderator**	Ok. You can go to the next task.
0:28:51	Task 15	**Participant**	Reads task.
0:29:09	Task 16	**Participant**	Reads task.
0:29:24	Task 17	**Participant**	Reads task.
0:29:34		**Participant**	I assume we are still using these training and test intervals so I just go A-node.
0:29:41		**Participant**	And now we are removing the default value for windows size and we are going to add between 20 and 40 with step 5.
0:30:10		**Participant**	For window hop I remove the default value and add between 10 and 20 with step 5 and the other default value.
0:30:28	Task 18	**Participant**	Reads task.
0:30:39	Task 19	**Participant**	Reads task.
0:31:41		**Participant**	How do we know when it's finished? There is no UI for it?
0:31:54		**Participant**	I guess we are almost done here. Once the fan stops spinning.
0:32:10		**Moderator**	It's done and you can go to the next.
0:32:13		**Participant**	I guess if you know how many combinations there's going to be, then you have a progress bar that tells you how many results are computed.
0:32:19		**Moderator**	Yes.
0:32:21	Task 20	**Participant**	Reads task.
0:32:56		**Participant**	Should I just run this? Because the annotation is already there. So then I can remove the widget for Wgng.
0:33:29		**Participant**	This is probably going to run it.
0:33:41		**Participant**	Well here's an A-node one.
0:33:53		**Participant**	I guess if I show that this is the one that I have picked.
0:34:12		**Participant**	I think that's comparing them.
0:34:15		**Moderator**	Can you read the whole task?
0:34:25	Task 21	**Participant**	Reads task.

0:35:41	Participant	I am only seeing one of the A-node ones, which are not hidden based on this. But its recall is lower than any of these two Wgng.
0:36:04	Participant	Minimum recall to 1. If I do that then everything goes away.
0:36:13	Moderator	You also have to set the precision to 0.
0:36:18	Participant	Ok, I set that to 1.
0:36:30	Participant	Now we have Wgng. The recall is 1 but precision is lower.
0:36:44	Participant	This is the tradeoff between precision and recall. You either have 100 % recall or 100 % precision with one of the algorithms.
0:36:59	Moderator	If I wanted to reach recall 1 how would the setups compare?
0:37:13	Participant	You would have either 88% precision or 64% precision depending on which algorithm you pick. If they're sorted by precision, there.
0:37:33	Moderator	Ok that was the last task
0:37:34	Participant	That was last task? I really like the project.

E.3 Participant 3

E.3.1 Log for Participant 3

Time	Task	Detail	Note
0:07:34	Task 1	Success	Successfully finished the task.
0:08:18	Task 2	Success	Successfully finished the task.
0:09:48	Task 3	Success	Successfully finished the task.
0:12:35	Task 4	Success	Successfully finished the task.
0:13:45	Task 5	Success	Successfully finished the task.
0:13:53	Task 6	Success	Successfully finished the task.
0:15:28	Task 7	Success	Successfully finished the task.
0:16:11	Task 8	Success	Successfully finished the task.
0:20:27	Task 9	Success	Successfully finished the task.
0:21:15	Task 10	Success	Successfully finished the task.
0:21:28	Task 11	Success	Successfully finished the task.
0:24:12	Task 12	Success	Successfully finished the task.
0:25:48	Task 13	Success	Successfully finished the task.
0:30:12	Task 14	Success	Successfully finished the task.
0:30:32	Task 15	Success	Successfully finished the task.
0:30:50	Task 16	Success	Successfully finished the task.
0:31:48	Task 17	Success	Successfully finished the task.
0:31:56	Task 18	Success	Successfully finished the task.
0:33:40	Task 19	Success	Successfully finished the task.

0:35:29	Task 20	**Observer**	Clicks on “Save anomaly annotation...” instead of selecting the original anomaly annotation and clicking “Evaluate all algorithm outputs...”
0:37:43	Task 20	**Success**	Successfully finished the task.
0:43:26	Task 21	**Success**	Successfully finished the task.

E.3.2 Transcript for Participant 3

Time	Task	Speaker	Notes
0:06:18		**Moderator**	So these are the instructions. Basically I am here in a role of an assistant so I am not here to evaluate you. I am just here to help you. You should always read number of the task and the name of the task and then proceed on doing it. And it would be great if you could think out loud and comment on the system. I will remind you to do it if you forget.
0:06:54		**Participant**	Ok. Is this being recorded?
0:06:58		**Moderator**	Yes.
0:06:58		**Participant**	Ok.
0:07:08		**Participant**	Should I also read which step I am taking?
0:07:13		**Moderator**	Yea exactly.
0:07:26	Task 1	**Participant**	Reads task.
0:07:39		**Participant**	A new set of views is showing up.
0:07:43	Task 2	**Participant**	Reads task.
0:07:57		**Participant**	I can zoom in and zoom out.
0:08:03		**Participant**	It’s a ??? pattern with type time series, missing dots here and there.
0:08:21	Task 3	**Participant**	Reads task.
0:09:37		**Participant**	Ok. That’s good. And I think I have to hit the “add” button to add this interval because I have the option to add more than one.
0:09:59	Task 4	**Participant**	Reads task.
0:10:15		**Participant**	Ok. Hit the Wgng button.
0:10:20		**Participant**	And the window size ... Ok remove the default value.....
0:11:04		**Participant**	Now window hop....
0:11:35		**Participant**	Neuron memory (t1)...
0:11:55		**Participant**	Is that allowed to use the range instead of specifying single values
0:12:03		**Moderator**	It is. Go on.
0:12:09		**Participant**	Oh I cannot do that because it is not evenly spaced.
0:12:14		**Participant**	So I have to add 0,1 and then 3, because it is not evenly spaced I cannot use the range specifier
0:12:40		**Participant**	And I leave the default values for the rest
0:12:44	Task 5	**Participant**	Reads task.

0:12:51		**Participant**	Ok, I have scrolled down to the bottom and I see 366 valid combinations because some combinations are not valid , 888 invalid are not valid
0:13:10		**Participant**	Ok, I guess I don't care about those cases so I just ignore this warning. Because I only care about the subset of combinations that are valid and I ignore the ones that are invalid.
0:13:28		**Participant**	I can hit the "show combinations" button to see which combinations are actually valid and which are invalid and for what reason they are invalid.
0:13:46	Task 6	**Participant**	Reads task.
0:14:06	Task 7	**Participant**	Reads task.
0:14:33		**Participant**	Ok so I look at the list of results, they are ordered by I guess some artificial index.
0:14:47		**Participant**	And I show the first one – show results – and it shows me the computed anomaly likelihood score aligned with the actual data
0:15:01		**Participant**	I can zoom in and see that where I see an anomaly in the original time series I also see spike in the anomaly score
0:15:16		**Participant**	Ok I activate second result, third one... ok... and hide all... I think I understand what the results are
0:15:30	Task 8	**Participant**	Reads task.
0:16:12	Task 9	**Participant**	Reads task.
0:16:41		**Participant**	I think you probably mean I add an anomaly interval.
0:16:51		**Participant**	I add anomaly interval number 1 which ranges from 9:11 ...
0:17:18		**Participant**	I think that's good enough
0:18:24		**Participant**	That's quite tedious, I only have 3 second time range.
0:18:47		**Participant**	Can I zoom in? ... ok
0:18:57		**Participant**	That's something I should have discovered before (laughs)
0:19:22		**Participant**	Can I scroll? ... No I can't scroll.
0:19:33		**Participant**	So I zoom into time range to have higher resolution that makes it much easier to select the anomaly – select the time interval where I annotate the anomaly
0:20:32	Task 10	**Participant**	Reads task.
0:21:19	Task 11	**Participant**	Reads task.
0:21:35	Task 12	**Participant**	Reads task.
0:22:08		**Participant**	I probably select this because it's the one I created.
0:23:21		**Participant**	So these are evaluated... yes... so I can select it and the I see the precision-recall curve
0:24:12		**Participant**	I think that completes task 12.
0:24:18	Task 13	**Participant**	Reads task.
0:24:27		**Moderator**	Can you read the whole task? Not out loud but it's good if you read it completely... Sometimes the task name is too short.
0:24:56		**Participant**	Yea... I think I get it.
0:25:05		**Participant**	So I need a minimum precision of 95 so I change the minimum precision slider to 0.95

0:25:28		**Participant**	And at the same the recall should be as high as possible... So should I should sort the list by maximum recall... which is already done
0:25:38		**Participant**	So this result... this parameter set is the one that we're looking for.
0:26:00		**Participant**	Yea I am looking at the anomaly score
0:26:12		**Participant**	So when I hover points in precision-recall curve I see the threshold that it corresponds to, that's very nice... that's actually very nice
0:26:38	Task 14	**Participant**	Reads task.
0:27:16		**Participant**	Ok this reduces the results set very significantly.
0:28:15		**Participant**	I think it's three results that are considered optimal... to non-optimal results judging by the precision-recall curve
0:28:35		**Participant**	It also looks like the three which are...
0:28:47		**Participant**	I think it's not obvious... I know what these are... And to what extend these are considered optimal... but it's not apparent... at least I don't see a description of what "optimal" means in this context.
0:29:06		**Participant**	Because one could optimize precision-recall curves by their area maybe...
0:29:44		**Participant**	Yea but these three have points... that the boundary of all those precision-recall curves from all results at the boundary what's the top-right – which is also known as a Pareto-front.
0:30:08		**Participant**	This is not obvious but that's probably what's been done here.
0:30:23	Task 15	**Participant**	Reads task.
0:30:36	Task 16	**Participant**	Reads task.
0:30:52	Task 17	**Participant**	Reads task.
0:31:00		**Participant**	I suppose for the same training and test intervals.
0:31:02		**Moderator**	Yea.
0:31:16		**Participant**	Ok so the windows size...
0:31:35		**Participant**	Window hop...
0:31:48		**Participant**	And the others I leave default values.
0:31:52	Task 18	**Participant**	Reads task.
0:32:01	Task 19	**Participant**	Reads task.
0:32:43		**Participant**	I don't remember how many combinations I had so I don't know for how many I have to wait so I switch back to the configurator tab and I see I have 15 combinations.
0:33:12		**Participant**	Aha 15... so the index 15... so the 15^{th} one executed so all of them must have been executed that's my assumption that they are executed in sequential order but browsing through this quick I see that all indices are there so I suppose all A-node executions have completed and the results are available.
0:33:43	Task 20	**Participant**	Reads task.
0:34:09		**Participant**	Ok I use the same markup annotation so that's fine.

0:34:49	**Participant**	Oh I have to run the evaluation, because here I see that they are not evaluated yet so I have to rerun the evaluation that's why I don't see precision-recall curve here, which makes sense.
0:35:08	**Participant**	I have to rerun the evaluation.
0:35:15	**Participant**	Question is if I click here it will probably do the evaluation only for the ones that have not been done yet is that correct?
0:35:28	**Participant**	I will try that.
0:35:35	**Participant**	Otherwise they would have been evaluated twice – the one that have been evaluated before. Which is ok for me too, I don't mind.
0:35:48	**Participant**	Now I select anomaly markup 1 and I see that they are all evaluated. I switch back to 0 and I see that these are not evaluated. Ok.
0:36:06	**Participant**	But for me number one is more interesting
0:36:15	**Participant**	And then I want their results. Would be nice to have a filtering method for these columns – for example only certain types of algorithms to be shown.
0:36:41	**Participant**	Let's look at some of them. Ok.
0:36:46	**Participant**	Let's see if the new algorithm has some results that are optimal. No – so I general this second algorithm performs worse than the first one because it doesn't have any optimal results.
0:37:06	**Participant**	But I think that's my evaluation of the new algorithm results task 20. That given the parameter sets that I have configured the new algorithm performed generally worse. So A-node performed generally worse than Wgng. Since none of the results of A-node evaluations appear in the optimal set in the prr curve. If that makes sense.
0:37:44 Task 21	**Participant**	So task 21. Aha that's the comparison... Reads task.
0:38:15	**Participant**	Ok I think that's what I have stated before so in the optimal set there are only the Wgng results which makes A-node a none optimal choice for this data and for this parameter sets and for these annotations that I've used.
0:39:14	**Participant**	Ok there are some instances of A-node that have higher precision at recall 1.
0:39:45	**Participant**	Minimum recall set to 1 so I am looking at the results where the recall is 1.
0:40:09	**Participant**	So still even for the where I require the recall 1 which means I don't have any false right? With recall 1 I don't have any false positives if I remember correctly from the definition of the recall.
0:40:46	**Participant**	Nee, I don't have any false negatives.
0:41:25	**Participant**	So it didn't detect all anomalies... No... Recall 1 means I detected all anomalies even at the cost of many false positives right?
0:41:45	**Participant**	So the precision gives me an indication of false positives the higher the number the fewer false positives I have right?

0:41:58	Participant	Minimum recall 1 says I am looking at the anomaly detection executions where I have detected 100 % of the anomalies that I have annotated but at the cost of false positives.
0:42:28	Participant	So the number of false positives using the Wgng algorithm with this parameter set is still lower than A-node but A-node performs comparably well also at the range of 0.84 at this scheme.
0:42:53	Participant	So given these prerequisites that I need to detect all anomalies A-node does compare to Wgng in terms of false positives but none of the results from the A-node anomaly detection execution are optimal. That's my comparison of A-node and Wgng.
0:43:22	Moderator	Thanks.
0:43:23	Participant	And this concludes task 21 which is the last task.
0:43:26	Moderator	Yes, great.

E.4 Participant 4

E.4.1 Log for Participant 4

Time	Task	Detail	Note
0:08:24	Task 1	Success	Successfully finished task.
0:09:13	Task 2	Success	Successfully finished task.
0:10:52	Task 3	Observer	Finished task but didn't click "add". (corrected later)
0:12:52	Task 4	Observer	Incorrect step value in window hop (corrected later).
0:14:05	Task 4	Success	Successfully finished task.
0:14:51	Task 5	Success	Successfully finished task.
0:17:27	Task 6	Success	Successfully finished task.
0:19:02	Task 7	Success	Successfully finished task.
0:19:15	Task 8	Success	Successfully finished task.
0:22:08	Task 9	Observer	Moderator explains how it is possible to set up time intervals with greater precision by zooming the time series plot.
0:24:20	Task 9	Success	Successfully finished task.
0:25:30	Task 10	Success	Successfully finished task.
0:25:38	Task 11	Success	Successfully finished task.
0:26:25	Task 12	Skipped	By accident.
0:27:16	Task 13	Success	Successfully finished task.
0:30:24	Task 14	Observer	Participant is confused about how precision and recall works.
0:30:48	Task 14	Success	Successfully finished task.
0:30:56	Task 15	Success	Successfully finished task.
0:31:10	Task 16	Success	Successfully finished task.
0:32:08	Task 17	Success	Successfully finished task.
0:32:11	Task 18	Success	Successfully finished task.
0:32:22	Task 19	Success	Successfully finished task.
0:33:17	Task 20	Success	Successfully finished task.
0:34:29	Task 21	Success	Successfully finished task.

E.4.2 Transcript for Participant 4

Time	Task	Speaker	Notes
0:07:06		Moderator	From now on you should work on your own. These are the instructions, the tasks that you should follow. You should always read the number of the task and the name of the task. You can also read the whole text of the task but you don't have to. But you should first read the whole task at least for yourself quietly before you start doing it. And comment on what you're doing like not just the task but how do you perceive it. What do you think that the program is doing right now etc. So I know what you're thinking.
0:07:47		Moderator	Are you a mac user or windows user?
0:07:49		Participant	Linux normally but now I got a Mac. Don't worry.
0:07:58		Participant	So basically just the first instruction is to not to use the top part of the system. Because it leads to functions we don't want to test.
0:08:14		Participant	Ok. Let's start then.
0:08:17	Task 1	Participant	Reads task.
0:08:22		Participant	So basically I guess I am clicking here.
0:08:25	Task 2	Participant	Reads task.
0:08:38		Participant	So it's a time series.
0:08:44		Participant	And I am trying to understand the points but doesn't make sense.
0:08:54		Participant	Ah ok. I didn't see the training and the test maybe it should be here.
0:09:05		Moderator	Ok.
0:09:08		Participant	So time series with training part and test part.
0:09:14	Task 3	Participant	Reads task.
0:09:42		Participant	Can we edit that? ... No, ok.
0:10:11		Participant	Does it have to be precise?
0:10:14		Moderator	Within +- 5 seconds is ok.
0:10:54	Task 4	Participant	Reads task.
0:11:52		Participant	So there I should put a window size. Actually I should put several numbers....
0:12:04		Participant	Ah ok looks more clear now.
0:13:19		Participant	That is not true actually.
0:13:40		Moderator	Can you comment?
0:13:41		Participant	Ye, ye sorry.
0:13:43		Moderator	No problem I am here to remind you that's my job.
0:13:49		Participant	I just added the number there. It's a bit annoying to click "add" every time.
0:14:13		Moderator	If you finish and go to next task always read...
0:14:22	Task 5	Participant	Reads task.
0:14:39		Participant	That to me looks a bit cryptic.

0:14:52	Task 6	**Participant**	Reads task.
0:14:56		**Participant**	The thing is I don't understand why those parameters are right or wrong and how to execute it.
0:15:57		**Participant**	Did I do a mistake with the parameters?
0:16:07		**Participant**	I go back to check the parameters I missed.
0:16:43		**Participant**	Ok I am going back to task 5.
0:17:06		**Moderator**	Ok so you have to actually check the task 3. That's the one where the problem is.
0:17:21		**Participant**	Hmm... Ok thanks.
0:17:31		**Moderator**	No problem.
0:17:36		**Participant**	Ok let's go to examine the results I guess.
0:17:41	Task 7	**Participant**	Reads task.
0:18:28		**Participant**	It's not very clear what does it mean there but I guess...
0:18:48		**Participant**	There are some fittings there and there evaluation of the parameters.
0:19:05	Task 8	**Participant**	Reads task.
0:19:21	Task 9	**Participant**	Reads task.
0:20:39		**Participant**	It's ??? within 5 second range.
0:20:43		**Moderator**	Well if you look at next one it's gonna be kind of problematic. Because it's very short.
0:20:51		**Participant**	Ye ok. But for the other one it's fine?
0:20:54		**Moderator**	It's fine.
0:20:58		**Participant**	Ah ye problematic indeed.
0:21:33		**Moderator**	Can you comment on what you're doing?
0:21:35		**Participant**	I am trying to select ??? anomaly
0:21:42		**Participant**	Actually it's not that easy.
0:21:51		**Participant**	08 will do instead of 10?
0:21:54		**Moderator**	Yea but I'll give you a tip because this is a common problem. So you can zoom with your mouse in the plot and the sliders zoom as well.
0:22:08		**Moderator**	Just select interval in the plot that you want to zoom in and then...
0:22:13		**Participant**	Hmm... yea ok... makes it much easier indeed.
0:22:38		**Participant**	Then next one.
0:23:40		**Participant**	Still I think it would be easier if you could just write them down.
0:24:35	Task 10	**Participant**	Reads task.
0:24:48		**Participant**	Ok basically that's the second one.
0:25:33	Task 11	**Participant**	Reads task.
0:25:36		**Participant**	Ok so I guess it's there.
0:25:56		**Moderator**	Can you comment on what you're doing or?
0:25:58		**Participant**	I am reading the text because apparently nothing happened.
0:26:23		**Participant**	Ah ok... I guess I am already at task 13.
0:26:28	Task 13	**Participant**	Reads task.

0:26:39		Participant	So easy you can take the precision slider and set it to 0.95.
0:26:53		Participant	Do we also want a recall that is high or not?
0:27:04		Participant	I guess we want the highest possible recall as well so I just set it high enough to get the first one and there it is.
0:27:19	Task 14	Participant	Reads task.
0:28:17		Participant	That I just don't understand.
0:28:20		Moderator	You're supposed to select the option that is written here.
0:28:27		Participant	Oh.
0:28:36		Participant	So basically this thing is the first one for you instead of you playing with that to get it.
0:28:48		Moderator	Try to remove the precision and recall. Like reset it to zero.
0:29:00		Participant	It's not always giving the same actually.
0:29:19		Participant	I guess there are other parameters account for it.
0:29:42		Participant	There is something weird there. If I set the precision to 0.95 and the recall high enough I have only that actually (one result).
0:30:04		Participant	But if I set it to zero... Not I zero but some value... I can have a max precision of 1 and max recall of 1as well... that's weird no? Ok, well...
0:30:24		Moderator	If you say so then it's weird of course. I mean that's the reason we are doing this.
0:30:30		Participant	For me I would have expected to have the... if I set it to 95 and something to recall those to come first... unless there is something problematic that... I guess there is
0:30:50	Task 15	Participant	Reads task.
0:30:58	Task 16	Participant	Reads task.
0:31:14	Task 17	Participant	Reads task.
0:31:48		Participant	How many of those have you done today?
0:31:51		Moderator	This is the fourth one.
0:32:14	Task 18	Participant	Reads task.
0:32:16		Participant	I already clicked on it.
0:32:20	Task 19	Participant	Reads task.
0:32:27	Task 20	Participant	Reads task.
0:33:30		Participant	Ok now I have both results so I can compare.
0:33:40	Task 21	Participant	Reads task.
0:34:08		Participant	Ok if I compare both I would say that A-node tends to be better because the precision under recall is closer to 1. That's from the info and that that I understand.
0:34:30		Moderator	Ok, that's it.

E.5 Participant 5

E.5.1 Log for Participant 5

Time	Task	Detail	Note

0:05:08	Task 1	**Success**	Successfully finished task.
0:06:15	Task 2	**Success**	Successfully finished task.
0:09:53	Task 3	**Success**	Successfully finished task.
0:12:28	Task 4	**Success**	Successfully finished task.
0:13:09	Task 5	**Success**	Successfully finished task.
0:13:12	Task 6	**Success**	Successfully finished task.
0:14:50	Task 7	**Success**	Successfully finished task.
0:14:58	Task 8	**Success**	Successfully finished task.
0:20:09	Task 9	**Observer**	Moderator explains to participant how to zoom the plot to be able to set up anomaly interval with required precision.
0:21:48	Task 9	**Success**	Successfully finished task.
0:22:42	Task 10	**Success**	Successfully finished task.
0:22:53	Task 11	**Success**	Successfully finished task.
0:24:05	Task 12	**Success**	Successfully finished task.
0:27:58	Task 13	**Success**	Successfully finished task.
0:28:33	Task 14	**Success**	Successfully finished task.
0:28:52	Task 15	**Success**	Successfully finished task.
0:29:56	Task 16	**Success**	Successfully finished task.
0:30:41	Task 17	**Success**	Successfully finished task.
0:30:58	Task 18	**Success**	Successfully finished task.
0:31:20	Task 19	**Success**	Successfully finished task, but didn't wait the recommended minute.
0:32:59	Task 20	**Success**	Successfully finished task – even though he didn't know that.
0:40:29	Task 21	**Success**	Successfully finished task.

E.5.2 Transcript for Participant 5

Time	Task	Speaker	Notes
0:05:05	Task 1	**Participant**	Reads task.
0:05:13		**Participant**	That's it, that's the first task right?
0:05:14		**Moderator**	Yea you should work on your own. If there is some problem or you get stuck I can help you but I shouldn't interfere very much. I will just remind you to think out loud etc.
0:05:27		**Participant**	Ok.
0:05:33		**Moderator**	You can read the task.
0:05:37	Task 2	**Participant**	Reads task.
0:05:47		**Participant**	Ok we have here the training data and the test data. 20 minutes of data we use for training and then 8 minutes of testing, something like that.
0:06:16	Task 3	**Participant**	Reads task.
0:07:04		**Participant**	Does it have to be sharp?
0:07:07		**Moderator**	It should be +- 5 seconds.

0:08:53		**Moderator**	Can you comment what you're doing and how you understand everything etc.?
0:08:57		**Participant**	First I thought we are setting up from this one (plot), but it allowed me to change the color but now I see there is a new setup here (sliders) and I think I should do it from here.
0:09:12		**Participant**	Which seem to make more sense.
0:10:08	Task 4	**Participant**	Reads task.
0:10:24		**Participant**	I guess I am to click this one to configure it.
0:10:32		**Participant**	I doesn't say add a new algorithm it says generate new combinations of parameters for ... Ok I guess I should start with the new algorithm
0:12:35	Task 5	**Participant**	Reads task.
0:12:46		**Participant**	Why are there so many invalid combinations?
0:12:59		**Participant**	Is it because I did something wrong or is it because....
0:13:04	Task 6	**Participant**	Ok... let's execute algorithm...
0:13:32	Task 7	**Participant**	Yes, I think the results displayed here. At least the correct ones.
0:13:45		**Participant**	I think it's this one.
0:14:04		**Participant**	Ok it shows results in different combinations.
0:14:16		**Participant**	Ok then here... what do they mean?
0:14:22		**Participant**	Training and testing... the training looks all the same... testing looks all the same... just the parameters are different
0:14:46		**Participant**	Ok they are consistent... more or less
0:14:52	Task 8	**Participant**	Reads task.
0:14:57		**Participant**	It's right here.
0:15:13	Task 9	**Participant**	Reads task.
0:15:33		**Participant**	I guess from here we can setup something
0:15:53		**Participant**	To create anomaly annotation...use anomaly designer...
0:16:00		**Participant**	It's this one why there's another here...
0:16:06		**Participant**	Ah it's already this one... this is a bit confusing but ye ok... this is anomaly annotator
0:17:58		**Participant**	Is it critical... should I try to be precise?
0:18:08		**Participant**	You're not supposed to talk? ... ok this is precise.
0:18:36		**Participant**	Can we use this stuff? No...
0:18:40		**Participant**	How can I use this more precisely?
0:19:00		**Participant**	Ok I go with this way...
0:19:03		**Moderator**	So you can use the mouse to select part in the plot and then the sliders will zoom together with the plot.
0:19:16		**Participant**	Ok, let's try this one now....
0:19:52		**Participant**	Now you say I can choose...
0:19:55		**Moderator**	You can zoom the plot with selecting the interval with the mouse.
0:20:09		**Moderator**	You just have to drag the mouse around the interval you want to zoom in... so for example you can zoom to everything from 9:10 to the end of the plot.
0:20:20		**Participant**	Ah I can actually choose like this...

0:20:24		**Moderator**	Ye.
0:20:28		**Participant**	But then from here... Ah ok got it.
0:20:48		**Participant**	This is now easy...
0:21:51	Task 10	**Participant**	Reads task.
0:22:03		**Participant**	Oh... Nope... I want to change this... Reset zoom... Ok.
0:22:44	Task 11	**Participant**	Reads task.
0:22:55	Task 12	**Participant**	Reads task.
0:23:09		**Participant**	I guess I already evaluated. I don't have to do that again... let me check.
0:23:42		**Participant**	There is not enough feedback so I don't know if it's running or not.
0:23:51		**Participant**	Ok so it probably is here.
0:24:15	Task 13	**Participant**	Reads task.
0:25:20		**Participant**	So this one is precision 96... Ok there is no precision... So it's this one... Now there's these two and for these two the recall is 35 %... so then I choose the precision 1
0:25:46		**Moderator**	Not really... Try to look at it again... maybe look around.
0:26:00		**Moderator**	Can you maybe read the whole description again? Of the task.
0:26:19		**Participant**	Precision is here...Recall is here...
0:26:30		**Moderator**	So right now you are using precision recall curve for one of the configurations. And you have to find the configuration that gives you the best
0:26:41		**Participant**	Ah ok.
0:26:47		**Participant**	Some of the configurations... so we check like this...
0:27:00		**Participant**	Which one of them is the... but there is like 300 configurations here
0:27:27		**Participant**	Minimum precision... 95%...
0:27:39		**Participant**	Basically we have... many... so let's try max recall
0:27:56		**Participant**	Ok this one now...
0:28:01		**Participant**	Hide all... and let's show just this one
0:28:19		**Moderator**	Yes this one, good.
0:28:21	Task 14	**Participant**	Reads task.
0:28:26		**Participant**	Ok I understood this part.
0:28:40	Task 15	**Participant**	Reads task.
0:28:53	Task 14	**Moderator**	So have you tried think about what this option means? Task 14
0:29:01		**Participant**	The 14? It's basically saying... if the curve is very bad... we don't count for it as an option... it's filtering basically... it's probably less than some threshold value.
0:29:18		**Moderator**	Ok.
0:29:31		**Participant**	It's basically just keeping the best ones.
0:29:52	Task 16	**Participant**	Reads task.
0:29:59	Task 17	**Participant**	Reads task.
0:30:05		**Participant**	Ok so same thing with this one.
0:31:03	Task 19	**Participant**	Reads task.

0:31:10	**Moderator**	So you're doing?
0:31:12	**Participant**	19 now.
0:31:20	**Participant**	Ok, I think now it's ok.
0:31:24 Task 20	**Participant**	Reads task.
0:32:18	**Participant**	Are these the new results or the old ones?
0:32:25	**Moderator**	The old ones should stay there but you should also evaluate new ones.
0:32:31	**Participant**	Ok.
0:32:46	**Participant**	I think I already completed everything. But it only shows Wgng.
0:32:53	**Participant**	Maybe I should evaluate all algorithm outputs...
0:33:33	**Participant**	Ok I don't understand.
0:33:37	**Moderator**	Try to comment maybe what you're trying to find.
0:33:40	**Participant**	So basically here I want to see the results of the new algorithm – A-node algorithm.
0:33:48	**Participant**	Because here in this one I only see the previous algorithm not the new one. So I would like to compare them.
0:33:58	**Participant**	But I'm not sure if the previous one, like from here. He also doesn't show Wgng.
0:34:16	**Participant**	Here we have this new algorithm and we ran it and it comes with the results here.
0:34:26	**Participant**	But then here it still shows the old one and the results of the old configurations.
0:34:46	**Participant**	Though I have to do something else.
0:35:03	**Participant**	So we wait until this gonna run.
0:35:08	**Moderator**	(sighs)
0:35:17	**Moderator**	So the results are there. They are in the last page because the table is sorted by the execution time of jobs.
0:35:31	**Participant**	Ok, but how am I supposed to know? ...ok.
0:35:45 Task 21	**Participant**	Reads task.
0:36:40	**Moderator**	Try to continue with the rest of the task... after hide anomalies.
0:36:47	**Participant**	Ok if not selected, select Hide ones....
0:36:50	**Moderator**	Ye.
0:36:51	**Participant**	But it already it already threw out the A-node – all of them. So I am a bit confused because I expected something there.
0:37:01	**Participant**	So basically all of them are bad then.
0:37:26	**Participant**	But I don't see any A-node here.
0:37:33	**Participant**	So this smells bad. They have ??? precision and recall.
0:37:45	**Participant**	So it sucks basically, A-node. (laughs)
0:38:01	**Participant**	Ok this is not filtering anything (turns of filters). So I want to understand..
0:38:10	**Participant**	So let me go back to evaluated – the A-node not evaluated. Why not?
0:38:34	**Participant**	Ok because I didn't evaluate them. Right? (laughs)
0:38:41	**Participant**	Ok let me see now.
0:38:53	**Participant**	Ok so things start to happen.

0:39:01	**Participant**	Ok now I am going back then so “Hide...”
0:39:08	**Participant**	Again there’s nothing so all of them the A-node results are bad then. Must be very bad.
0:39:21	**Participant**	Did I do something wrong here? Did I evaluate the wrong thing or not?
0:39:25	**Moderator**	No everything is fine.
0:39:28	**Participant**	Ok if everything is fine, basically if I don’t set any filtering. It’s basically hidden for sure automatically from this one (“Hide non-optimal conf.”) which means they are already filtered out.
0:39:46	**Participant**	So if they’re here, I can see they’re here and when I hide them if they don’t appear they are worse than the set threshold.
0:40:03	**Participant**	So no matter what I do no matter how I change this stuff basically it’s not gonna come. So this algorithm has low accuracy, low precision, recall than the Wgng. Isn’t it? Hm?
0:40:18	**Moderator**	Ye.
0:40:20	**Participant**	That’s it?
0:40:21	**Moderator**	Ye.
0:40:23	**Participant**	Ok I thought there was... I was expecting here to see one number so that I can then compare.
0:40:29	**Moderator**	Ye.