**Anomaly Detection in Industrial Control Systems using Machine learning approach leveraging Novelty Detection methods.**

***Abstract***: Attack detection problems in Industrial Control Systems (ICSs) are usually known as a network traffic monitoring scheme for detecting abnormal activities. However, a network-based intrusion detection system can be deceived by attackers that imitate the system’s normal activity. The proposed approach is called measurement intrusion detection system (MIDS), which enables the system to detect any abnormal activity in the system even if the attacker tries to conceal it in the system’s control layer. A supervised machine learning model is generated to classify normal and abnormal activities in an ICS to evaluate the model’s performance. We leverage anomaly detection methods which is the process of finding abnormalities in data. Abnormalities are defined as the rows that deviate significantly from the general behavior of the data. Novelty detection is a method when the training data is not polluted by outliers, and we are interested in detecting whether a new observation is an outlier. In this context an outlier is also called a novelty. In the context of novelty detection, novelties/anomalies can form a dense cluster as long as they are in a low-density region of the training data, considered as normal in this context.

***Index Terms***: Industrial control systems (ICS), Anomaly Detection, Intrusion Detection System, Novelty detection, Supervised Machine Learning, SVM, Autoencoders, Multi-layer Perceptron, HAI datasets.

***Problem statement***: The objective of the project is to find out if it is possible to use a semi-supervised anomaly detection methods such as One class SVM, Autoencoder, IsolationForest and Deep Learning Multi-Layer Perceptron to find outliers that could indicate an alert for an attack in measurement-based intrusion detection system.

***Introduction***: The industrial control system (ICS) consists of devices, networks, and controllers to automate industrial processes. ICS contains several types of control systems, such as supervisory control and data acquisition (SCADA) systems, and distributed control systems (DCSs). ICSs are widely used in different critical infrastructures such as smart grids, power distribution, transportation systems, water treatment plants, and manufacturing. In the power plants, ICS’s key role is evident, and a multitude of automated systems are operating in a SCADA framework. The automated systems’ entanglements could endanger the entire system’s performance, where a small fault or malfunction would lead to a cascade failure. Thus, fault detection in ICSs, especially in critical infrastructures such as large-scale power plants, has attracted much attention in recent years.   
 Generally, communication between ICS components is based on an information technology stack (ITS) and remote connectivity. The reliance on communication networks to transmit measurements could increase the possibility of intentional attacks against physical plants. Owing to the design and structure of the system it is very vulnerable against malicious activities such as insider sabotages, spoofing, and stealthy attacks. One of the proposed solutions to this issue it MIDS (Measurement based Intrusion Detection Systems). This fault detection approach can find any deviation from normal performance caused by malicious activities such as changing the sensors setpoints or injecting fake data measurements into the ICS network levels.

Basically, due to the difficulties in generating a labeled dataset, which indicates different types of attacks in an ICS, most studies apply a normal activity dataset for training machine learning models. Therefore, the MIDS could only compare a set of normal data with the incoming data and detect any deviation from the normal activity. This is one of the crucial reasons we apply Novelty detection approach owing to the absence of labeled dataset in the training data. However, this strategy would fail while a stealthy attack that imitates a normal behavior intrudes into the system.

The detection process requires being able to decide whether a new observation belongs to the same distribution as existing normal observations (which is an inlier without an attack) or should be considered as different (it is outlier) or an anomaly. Often, this ability is used to clean real data sets where two important distinctions must be made whether the new unseen data is an anomaly (attack) or Normal.

The investigation procedure includes pre-processing the data, fitting supervised learning models, and evaluating each model’s classification accuracy. The standard methods of assessing models’ effectiveness are the confusion matrix, the area under the curve (AUC), and the receiver operating characteristics (ROC) curve.

***Methodology:***

***Dataset:*** The dataset used in this paper is from a HIL-based augmented ICS security (HAI) available at <https://www.kaggle.com/datasets/icsdataset/hai-security-dataset/>. The testbed dataset is built by collecting measurements of 86 sensors continuously for several days. During these days, 58 sophisticated attacks are injected into the system. These attacks are a combination of 14 process control loop (PCL) primitive attacks which are affecting four points in the system: setpoints, process variables, control output, and control parameters. The attacks are stealthy type and cannot be detected easily by the conventional network intrusion detection systems.

***Anomaly:*** Anomaly detection consists of various domains, such as intrusion detection, fault detection, and event detection in sensor networks. Any deviation from a normal performance could be considered an anomaly in an ICS. It could happen due to several reasons, including a malfunction in a system’s component, insider sabotage, or an intentional cyberattack. In this context the concept of anomaly detection based on the MIDS refers to fault detection and intrusion detection. When a malfunction or insider sabotage occurs, the system tries to detect faults in the system. In addition, when an attacker attempts to intrude in the system, it is known as intrusion detection. Some of the anomaly detection algorithms are,

* Local Outlier Factor
* Isolation Forest
* Connectivity Based Outlier Factor
* KNN based Outlier Detection
* One class SVM
* Autoencoder

***Novelty Detection***:

Consider a data set of n observations from the same distribution described by features. Consider now that we add one more observation to that data set. Is the new observation so different from the others that we can doubt it is regular? (i.e., does it come from the same distribution?) Or on the contrary, is it so similar to the other that we cannot distinguish it from the original observations? This is the question addressed by the novelty detection tools and methods. This is also called one class learning or cost-effective learning since the entire training set is labelled only as a single class.

In general, it is about to learn a rough, close frontier delimiting the contour of the initial observation’s distribution, plotted in embedding -dimensional space. Then, if further observations lay within the frontier-delimited subspace, they are considered as coming from the same population than the initial observations. Otherwise, if they lay outside the frontier, we can say that they are abnormal with a given confidence in our assessment.

***Outlier Detection vs Novelty Detection:*** In outlier detection, the training data consists of both anomalies and normal observations whereas in novelty detection the training data consists only of normal observations rather than having both normal and anomalous observations. In this project, we’re going to see a use case of novelty detection.

***Feature Engineering***: The goal of feature selection is to find the most effective features that lead to training more accurate models and less computation time. Hence, we have applied random forests dimensionality reduction algorithm not only to lower the computation cost, but also to oppose the problem of overfitting and multicollinearity. Random forests is a tree-based model which is widely used for regression and classification tasks on non-linear data. It can also be used for feature selection with its built-in attribute which calculates feature importance scores for each feature based on the ‘gini’ criterion (a measure of the quality of a split of internal nodes) while training the model. As the result of dimensionality reduction process, based on the features’ importance scores in Figure 1, we selected first 30 features with the highest importance on the target. Figure 2 shows the correlation matrix of the training set before and after applying dimensionality reduction.

Chart

Description automatically generated with medium confidence

**Figure 1**. Importance scores of features

A picture containing text

Description automatically generatedChart, treemap chart

Description automatically generated

**Figure 2**. Correlation matrix before dimensionality reduction (left) and after (right)

After selecting the features with the most importance scores, in the pre-processing step, the input data should be scaled. This could result in a sustainable learning process. . In this project, we used

Data standardization to scale non-Boolean feature values. Standardizing the dataset involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1. The equation below describes standardization process, where X’ is standardized form of feature X, μ is the mean value for the feature, and σ is the standard deviation of the feature.



***Machine Learning Models and techniques***:

Supervised anomaly detection in ICS generally uses normal activity data to build a predictive model of normal class as well as anomaly class. Then, any unforeseen data are compared with the generated model to detect its class. Several algorithms are applied in this study to train a machine learning model for detecting anomalies by the MIDS. Having access to a labeled dataset allows for applying supervised learning strategies by considering two classes of attack and normal activities. In this study, the most accurate supervised learning algorithms are chosen that are One Class SVM, Autoencoder, Local Outlier Factor, and Multi-Layer perceptron.

* [***One-class SVM***](https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection) is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set. The [Support Vector Method For Novelty Detection by Schölkopf et al.](http://scholar.google.nl/citations?view_op=view_citation&hl=en&user=DZ-fHPgAAAAJ&cstart=400&pagesize=100&sortby=pubdate&citation_for_view=DZ-fHPgAAAAJ:GFxP56DSvIMC) basically separates all the data points from the origin (in feature space F) and maximizes the distance from this hyperplane to the origin. This results in a binary function which captures regions in the input space where the probability density of the data lives. Thus, the function returns +1 in a “small” region (capturing the training data points that are said to be normal) and −1 elsewhere (that are anomalies.
  + One class SVM requires the choice of a kernel and a scalar parameter to define a frontier. The RBF kernel is usually chosen although there exists no exact formula or algorithm to set its bandwidth parameter. The nu parameter, also known as the margin of the One-Class SVM, corresponds to the probability of finding a new, but regular, observation outside the frontier.

Chart, diagram

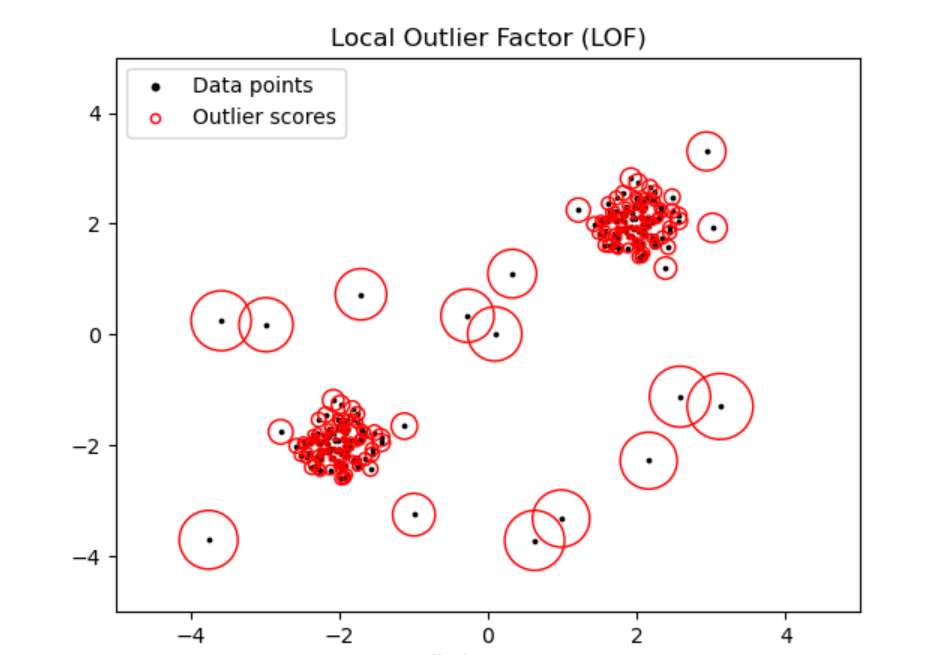
Description automatically generated

* ***Autoencoder*** is an unsupervised artificial neural network that learns how to efficiently compress and encode data then learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible.

Diagram

Description automatically generated

* + Autoencoders are applied to many problems, from facial recognition, feature detection, anomaly detection to acquiring the meaning of words. Autoencoders are also generative models: they can randomly generate new data that is similar to the input data (training data)
  + By learning to replicate the most salient features in the training data under some of the constraints, the model is encouraged to learn to precisely reproduce the most frequently observed characteristics. When facing anomalies, the model should worsen its reconstruction performance. In most cases, only data with normal instances are used to train the autoencoder; in others, the frequency of anomalies is small compared to the observation set so that its contribution to the learned representation could be ignored. After training, the autoencoder will accurately reconstruct "normal" data, while failing to do so with unfamiliar anomalous data. Reconstruction error (the error between the original data and its low dimensional reconstruction) is used as an anomaly score to detect anomalies.
* ***Local Outlier Factor***: The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors.
  + The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters. The "reachability distance" used by LOF has some subtle details that are often found incorrect in secondary sources.
  + Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbors.
  + The number of neighbors considered (parameter n\_neighbors) is typically set 1) greater than the minimum number of samples a cluster has to contain, so that other samples can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by samples that can potentially be local outliers.



* Multi-layer perceptron: A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network. MLP is a deep learning method.
  + An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.
  + MLPClassifier trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters.
  + It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.

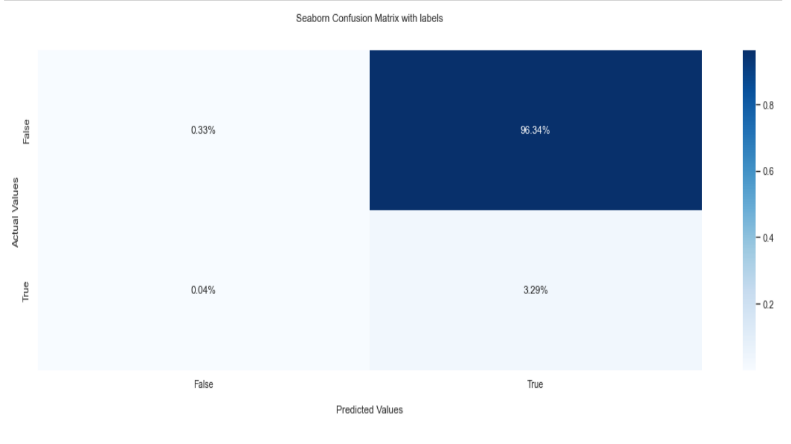
Diagram

Description automatically generated

Results and Conclusions:

The figures below compare the accuracy as well as confusion matrix of the four tested algorithms.

|  |  |  |
| --- | --- | --- |
| Technique | Classifier | Accuracy score |
| SVM | OneClassSVM | 0.95 |
| Autoencoders | Encoder and Decoder | 0.645 |
| LOF | OutlierClassifier | 0.036 |
| MLP | MLPclassifier | 0.9667 |

Chart

Description automatically generatedChart, treemap chart

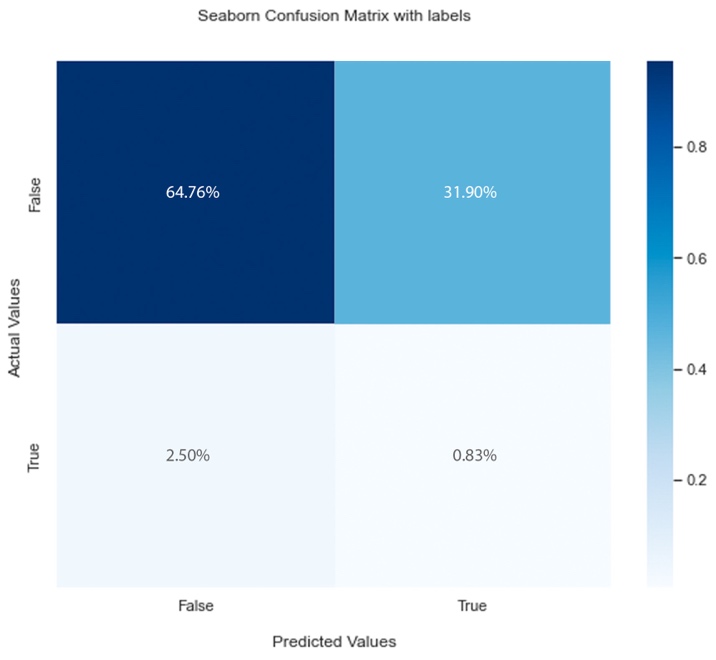
Description automatically generated

Figure X. Confusion matrix for Once Class SVM (top left), Autoencoder (top right), Local Outlier Factor (bottom left) and Multilayer Perceptron (bottom right) algorithms.

references:

1. Paridari, K.; O’Mahony, N.; Mady, A.E.D.; Chabukswar, R.; Boubekeur, M.; Sandberg, H. A framework for attack-resilient

industrial control systems: Attack detection and controller reconfiguration. Proc. IEEE 2017, 106, 113–128. [CrossRef]

2. A Machine Learning Approach for Anomaly Detection in Industrial Control Systems Based on Measurement Data by Sohrab Mokhtari 1,Alireza Abbaspour 2,, Kang K. Yen 1 and Arman Sargolzaei

3. Arafat, M.; Iqbal, S.; Hadi, M. Utilizing an Analytical Hierarchy Process with Stochastic Return on Investment to Justify Connected

Vehicle-Based Deployment Decisions. Transp. Res. Rec. 2020, 2674, 462–472. [CrossRef]

3. Abbaspour, A.; Mokhtari, S.; Sargolzaei, A.; Yen, K.K. A Survey on Active Fault-Tolerant Control Systems. Electronics 2020,