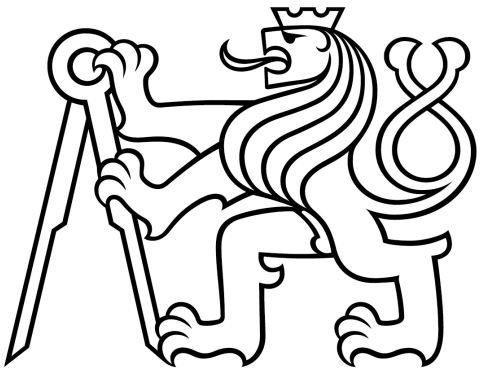
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics



Bachelor's thesis

Interpretation of Positive Detection of Out-of-Control State for Statistical Process Control

Ondřej Mísař

Supervisor: Ing. Macaš Martin, Ph.D

Study Programme: Cybernetics and Robotics

May 2022

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Mísař Ondřej** Personal ID number: **492368**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Cybernetics and Robotics**

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Interpretation of Positive Detection of Out-of-Control State for Statistical Process Control

Bachelor’s thesis title in Czech:

Interpretace pozitivní detekce statisticky nezvládnutého stavu při statistické regulaci procesů

Guidelines:

In multivariate methods for statistical process control, one of the main goals is to detect an out-of-control state (or process shift) from multiple quality characteristics. The detection methods typically aggregate all measured characteristics into one score. If the process shift is detected, an essential and closely related topic is the interpretation, i.e. the determination of which characteristics caused the positive detection. This helps with the subsequent diagnosis and selection of the process control action. The student should
1. make a literature survey on solutions of interpretation of positive detection of out-of-control state (process shift),
2. choose and implement 2-3 methods of interpretation in combination with one or more methods of process shift detection (e.g. Hotelling control chart or a machine learning based anomaly detector),
3. generate suitable synthetic data with properties specified by the supervisor that will be used for experimental evaluation (if possible, data from a real world process provided by supervisor can be also prepared),
4. evaluate the methods from task 2 on data from task 3 and compare them for different process shifts using appropriate performance measures (e.g. accuracy).

Bibliography / sources:

[1] Bersimis, Sotiris, Aggeliki Sgora, and Stelios Psarakis. "A robust meta method for interpreting the out of control signal of multivariate control charts using artificial neural networks." Quality and Reliability Engineering International 38.1 (2022): 30-63.
[2] Montgomery, Douglas C. Introduction to statistical quality control. John Wiley & Sons, 2020.
[3] Runger, George C., Frank B. Alt, and Douglas C. Montgomery. "Contributors to a multivariate statistical process control chart signal." Communications in Statistics--Theory and Methods 25.10 (1996): 2203-2213.

Name and workplace of bachelor’s thesis supervisor:

Ing. Martin Macaš, Ph.D. Cognitive Neurosciences CIIRC

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **27.01.2022** Deadline for bachelor thesis submission: **20.05.2022**

Assignment valid until: **30.09.2023**

Ing. Martin Macaš, Ph.D.	prof. Ing. Tomáš Svoboda, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
Supervisor’s signature	Head of department’s signature	Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

. _____ Date of assignment receipt

_____ Student's signature

Acknowledgement

I would like to thank my supervisor Martin Macaš for fully supporting me and providing all the necessary help throughout the course of the work on this project. Also, I would like to thank Han and Charlotte, who were part of the project and also helped me with my thesis.

Author statement for undergraduate thesis

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, 20.05.2022 Ondřej Mísař

Abstrakt

Důležitým předpokladem mnoha aplikací metod strojového učení v reálném světě je vysvětlitelnost výsledků těchto metod. Proto je trendem vytvářet metody, které nejen fungují dobře, ale jsou také vysoce interpretovatelné. Tato práce se zabývá interpretací modelu One-Class Support Vector Machine aplikovaného v oblasti kontroly kvality. Konkrétně se problém interpretace zaměřuje na určení správných charakteristik kvality (QCs), které jsou příčinou statisticky nezvládnutého stavu stavu (OOC). Byly vybrány tři interpretační metody a porovnány pomocí tří navržených měr výkonnosti.

Novinkou této práce je použití metody LIME, protože dosud nebyla použita na problém interpretace One-Class SVM. Nevýhodou této metody je, že uživatel musí určit, kolik charakteristik kvality způsobilo OOC, a LIME pak vybere ty, o kterých si myslí, že to jsou. Tento problém jsem vyřešil pomocí jednoduché heuristiky. Vyhodnocení ukázalo, že výsledky metody LIME jsou výrazně horší než výsledky zbývajících dvou metod, které jsou již pro problém interpretace v kontrole kvality používány. Nicméně toto bylo způsobeno mnou navrženou heuristikou nikoliv samotnou metodou LIME. To potvrdila i druhá sada experimentů, v níž byla interpretační metodě LIME poskytnuta informace o tom, kolik QC je třeba určit pro daný vzorek. V tomto případě byly výsledky mnohem slibnější i v kontextu zbylých metod.

Metoda LIME se jeví perspektivně pro interpretaci v oblasti kontroly kvality, nicméně je třeba nahradit zmíněnou heuristiku, která určuje kolik QCs je třeba najít, aby byla metoda LIME použitelná v praxi.

Klíčová slova: interpretace, statistická kontrola procesů, detekce statisticky nezvládnutého stavu, charakteristika kvality

Abstract

An important assumption for many real-world applications of machine learning methods is the explainability of the outcomes of such methods. Therefore, the trend is to create methods that not only perform well but are also highly explainable. This thesis deals with the interpretation of the One-Class Support Vector Machine model applied to quality control. Specifically, the interpretation problem focuses on determining the correct quality characteristics (QCs) that are responsible for a positive detection of an out-of-control state (OOC). Three interpretation methods were selected and compared using three proposed performance measures.

The novelty of this thesis is the use of the LIME method, as it has not been applied before to the One-Class SVM interpretation problem. The disadvantage of this method is that the user has to determine how many quality characteristics caused the positive OOC detection, and then LIME estimates those that caused the OOC. I solved this problem using a simple heuristic. The evaluation showed that the results of the LIME method are significantly worse than the results of the two other methods that are already in use for the interpretation problem in quality control. However, this was due to the proposed heuristic, not the LIME method itself. This was confirmed by a second set of experiments, in which the number of QCs responsible for OOC detection was known and provided to the LIME method. Then, the LIME gave much more promising results even in the context of the other methods.

The LIME method is promising for interpretation in quality control. However, a more sophisticated approach to finding the correct number of shifted QCs needs to be devised to make the interpretation method applicable in practice.

Keywords: interpretation, statistical process control, out-of-control state detection, quality characteristic

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In multivariate methods for statistical process control, one of the main goals is to detect an out-of-control (OOC) state (or process shift) from multiple quality characteristics (QCs). Detection methods typically aggregate all measured characteristics into one score. If the process shift is detected, an essential and closely related topic is the interpretation, i.e. the determination of which characteristics caused the positive detection. This helps with the subsequent diagnosis and selection of the process control action. This thesis focuses on the last part of the process control, i.e. the interpretation of positive detection of an out-of-control state.

1.1 Motivation

Only production process with limited variability can be a source of products that meet the customer's expectations. Therefore, almost any process has to be monitored to ensure an outcome (or product) of expected quality. This is the role of Statistical process control (SPC) or Multivariate statistical process control (MSPC), when there is more than one quality characteristic that needs to be observed. Control charts are one of various MSPC tools and methods. If an out-of-control state is detected, which means that some unexpected and undesirable change has occurred in the process, action is needed as soon as possible. Thus, the search for assignable causes begins. Common practice is that engineers and experts on particular process step in at this point and, after they find the assignable cause, an action is taken to adjust or correct the process. Interpretation methods can help experts find the assignable cause. Because the process can be complex, it may not be easy to identify the assignable cause. Since taking a correct action is time-consuming, any clue "where to look for" can be crucial.

From one perspective, the interpretation falls into the problem of explainable Artificial Intelligence (XAI), which has been developing a lot lately and is getting more and more

attention. Its importance is more than obvious given the increasing number of applications of AI methods in critical sectors.

1.2 Goals

The overall goal of this thesis is to test and compare 2-3 methods for interpretation of positive detection of out-of-control state. This main goal is, of course, divided into several smaller parts. First, it is necessary to conduct a literature survey and find out which interpretation methods already exist. The second step is to implement two or three of them. For the comparison itself, it is necessary to generate benchmark out-of-control data. Then, in combination with the use of an out-of-control state detector (One-class SVM), the interpretation methods can be tested and compared on the generated data. In the last step, it is important to apply appropriate performance measures to compare the results of different interpretation methods.

1.3 Structure

This thesis is divided into five chapters. This first chapter, which briefly introduced the scope of this thesis, its goals, and why interpretation is important, is followed by a chapter with an overview of the basic SPC and MSPC methods. The second chapter also introduces the One-Class SVM classifier, which is used as an out-of-control state detector in this thesis. The third chapter covers the interpretation itself, which is explained in detail. Furthermore, the interpretation methods that I found in the literature are also mentioned there, especially the three that I decided to use and test. The final part of this chapter defines the performance measures used to compare the results of the interpretation methods. The final chapter describes the generation of out-of-control data, the experiments scenario, and the results of each method with discussion. The conclusion summarizes the entire work and points out possibilities for future research.

Chapter 2

MSPC

Multivariate Statistical Process Control (MSCP) can be described as a set of methods for controlling a continuous process which leads into its stability and reduction of variability. In general, p quality characteristics (correlated random variables), x_1, x_2, ..., x_p characterize the quality control problem. For example, it can be the inner and outer diameter of a bearing that together define the usefulness of the part [1].

2.1 SPC

Historically MSPC evolved from Statistical Process Control (SPC). As the name suggests, MSPC considers multiple quality characteristics together while SPC considers only one quality characteristic, which is measured and controlled.

One of the SPC tools is the well-known Shewhart control chart, which can be seen in Fig. 2.1. The basic control chart graphically represents the measured value of the quality characteristic as a function of time or sample number. It contains three lines, upper control limit (UCL), lower Control limit (LCL) and center line, which reflects the average value of the measured quality characteristics. UCL and LCL are choosen based on the measured data and so that if the process is in control, then almost all points (usually, the limits are set as $\pm 3\sigma$ *standard deviations*, which results in ≈ 99.7 %) will fall between them. As long as the points plot within the control limits, the process is assumed to be in control, and no action is necessary [1].

2.2 Hotelling T^2 control chart

Naive idea could be to use Shewhart control chart for each monitored QC individually and only if all of p QCs are between their control limit, than the process is considered to be in-control. We can plot this using a 2d graph, where we plot the values of one QC on

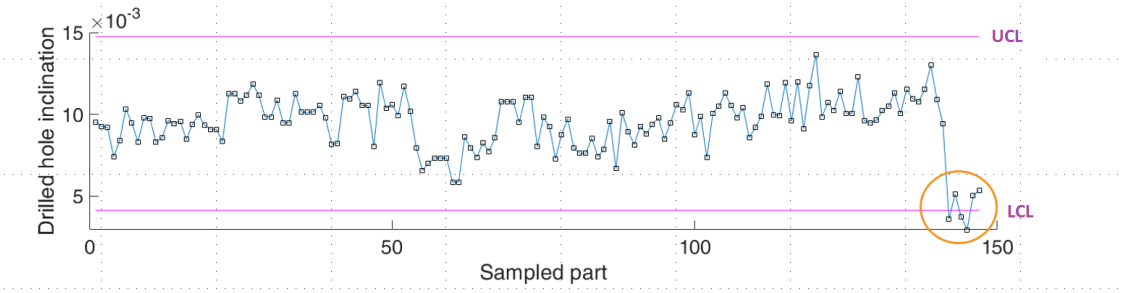


Figure 2.1: Basic example of typical Shewhart control chart

the x-axis and the other on the y-axis. Limits then create a square around the measured points.

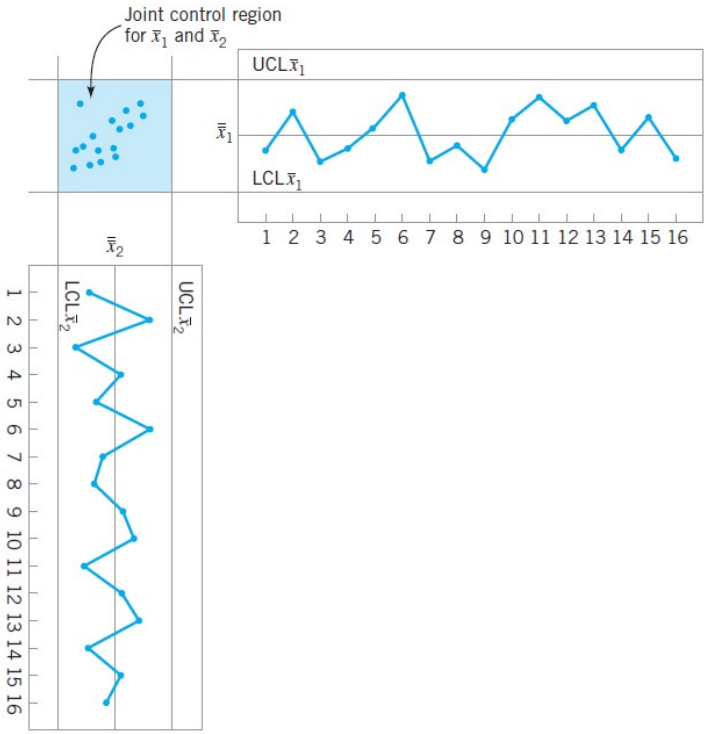


Figure 2.2: Control region using independent control limits (adopted from [1])

But as we can see in Fig. 2.2, one point (observation) appears in unusual distance from others, but is still inside the control limits. It is not possible to detect this suspiciously distant observation by univariate control charts, because the shape of the arrangement of the other points is caused by the correlation between the QCs. It is common in practice, that QCs are usually not independent, especially when they relate to the same product. Thus Hotelling T^2 control chart is considered as direct analog of univariate Shewhart control chart in MSPC.

In univariate statistics, considering normal distribution, we have formula for probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < \infty \tag{2.1}$$

If we ignore the constant, the term in the exponent of the normal distribution can be written as follows:

$$(x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{2.2}$$

This can be interpreted as standardized distance from x to the mean μ. By "standardized" we refer to distance, which is expressed in standard deviation units. We can apply the same approach to multivariate normal distribution case. Let's suppose, that we have p variables. Now the $\mathbf{x}$ and $\boldsymbol{\mu}$ aren't scalars, but vectors and instead of σ we have $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. So the squared standardized distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ from Eq. 2.2 changes to

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \tag{2.3}$$

We use T^2 hotelling distance definitions for individual observations, since in our case our sample size n is always equal to 1. Thus, it can be stated that in the context of this thesis, observation and sample are the same thing. Suppose that m samples are available and that p is the number of quality characteristics observed in each sample. Let $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ be the sample mean vector and covariance matrix, respectively, of these observations.

$$T^2 = (\boldsymbol{x}-\bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{2.4}$$

Note that Eq. 2.4 differs from Eq. 2.3 only by substituting $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. In practice, we usually do not know the exact values of the mean and covariance matrix, so we have to estimate them from the measured samples. This gives us a basic intuition on how to understand the value of T^2. If we satisfy the condition of sufficiently large m (depending on the size of p), for example $m \geq 250$ or the covariance matrix and mean vector are known, we can use the Chi-Square value with p degrees of freedom at the α significance level, which can be found in tables, to determine the UCL. For $p = 2$, we can visualize the measured samples together with the control ellipse, which determines the UCL threshold, this is depicted by a green line in Fig. 2.3.

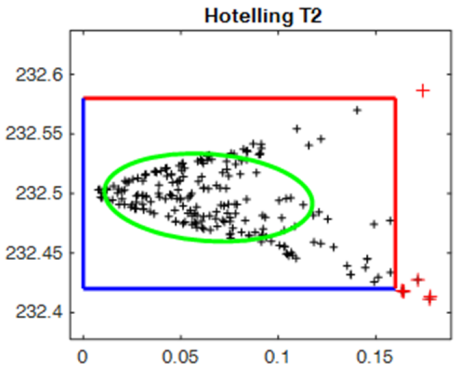


Figure 2.3: Control ellipse example

The shape of the ellipse is determined by the correlation between the quality characteristics, in the extreme case where they are independent it becomes a circle. On the other hand the greater the dependence between quality characteristics the "narrower" the ellipse. Another visualization option is the so-called chi-square control chart. Its advantage is that it can be used for more than two quality characteristics, because we plot only one value for each measured sample, which in our case is the value of T^2, and it preserves time sequence of the plotted points. Other variants of the MSPC control chart are, for

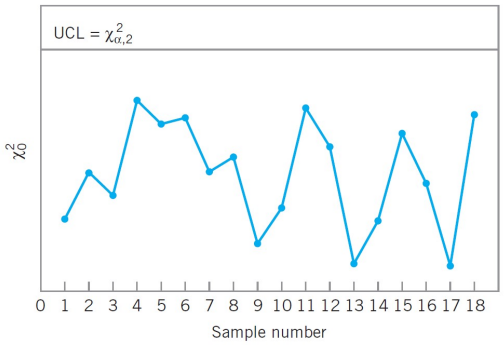


Figure 2.4: Chi-square control chart for two quality characteristics (adopted from [1])

example, the multivariate exponentially weighted moving average chart (MEWMA) or the multivariate cumulative sum chart (MCUSUM), but since they are not relevant for the type of data used in this thesis, they will not be further described here [1].

2.3 One-class SVM

In addition to the standard MSPC methods mentioned above, machine learning methods are also used in the field of out-of-control state detection. Specifically, Support Vector Machine (SVM). SVM is a supervised learning algorithm that creates one or more optimal separation hyperplanes that separate classes in a multidimensional space. SVM creates an ideal hyperplane in a binary classification environment as a linear classifier by maximizing the margin, or the distance between two classes. By solving a quadratic optimization problem that maximizes the margin while attempting to keep the training error low, the SVM is able to perform well on both linearly separable and linearly non-separable datasets in this situation. Despite the fact that the SVM was designed as a linear classifier, Boser et al. suggested a way to create non-linear SVM classifiers based on a kernel trick on the original hyperplane-maximizing SVM [2]. Single-class classification issues are also solved using a modified version of SVM, where the goal is to represent a positive class without taking data from a negative sample. The model is normally trained using only the training set created by the data points of one specific class because the purpose is to find the class entry in the middle of all classes. In this example, One-Class SVM creates a minimum

hyperplane, which is optimal. made up of all data points. The hyperplane boundary then encircles the data considered to be in-control [3].

Since they proved that One-Class SVM has better classification results than Hotelling T^2 control chart in the out-of-control state detection, this thesis will mainly use One-Class SVM as a classifier and investigate the role of interpretation on the outliers classified by the SVM [3].

Chapter 3

Interpretation methods

The problem of interpretation falls into the deeper category of explainable AI (XAI), which has received increasing interest recently. This is due to the fact that, with the increasing possibilities of using AI methods, many times the output of the black-box model alone is no longer enough. The user also needs to know the reason (explanation) why the model made the decision that it did. This is usually needed, especially in critical areas such as defense, medical, or financial systems. It is often the case that the most accurate methods (e.g., SVM or DL NN) are the least explainable, and, in contrast, the most explainable methods (e.g., decision trees) achieve the lowest accuracy. The trend and challenge of the time are to develop methods that meet both high-performance and explainability requirements. The main requirement of interpretation is that its output, describing the outcome or decision of a model, should be primarily human-understandable [4].

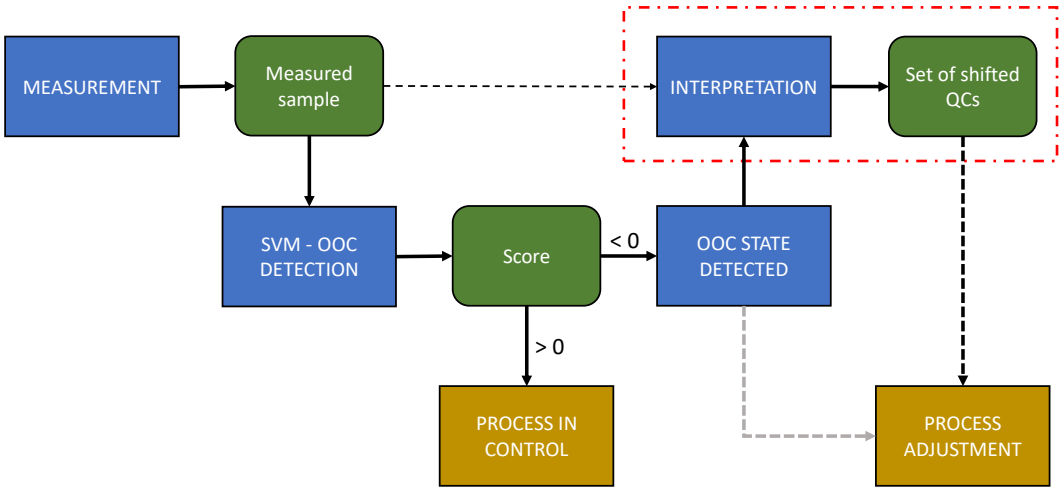


Figure 3.1: Flowchart of process control with added Interpretation

In our case, we have a trained One-Class SVM model, which is able to evaluate a validation sample whether it is an in-control (IC) state or an out-of-control (OOC) state. This method only returns one value for a measured sample, called the score.

This thesis focuses on the interpretation of positive detection of the out-of-control state by One-Class SVM classifier. This means that if an out-of-control state is detected, an interpretation method is used on that particular measured sample. It interprets why the classifier decided that the sample is out-of-control, by choosing a set of quality characteristics (QCs) in which a shift has occurred. The entire course of actions, from measuring a sample to adjusting a manufacturing process, when an out-of-control state is detected, is shown in the flow chart in Fig. 3.1. Our interpretation part is highlighted by a red dashed rectangle.

The importance of interpretation is pretty clear from the flow chart. When an out-of-control state occurs and is detected, without interpretation, a supervisor of a manufacturing process is provided only with the information that the process is out-of-control and has to find the cause manually himself. On the other hand, when an interpretation is available, he also receives additional information about which quality characteristics are shifted. This can directly save time to repair or adjust the manufacturing process, thus save money.

3.1 State of the art

Historically, around the beginning of the century, analytical interpretation methods based primarily on various statistical heuristics were developed. Some of them used T^2 Hotelling value decomposition [5], other projection methods (mainly principal components analysis) [6] and other approaches [7], [8], [9]. All of the analytical methods mentioned above are compared and briefly explained in [10].

The development of computation methods followed. These are characterized by the fact that, unlike analytical methods, they require some training. First of all, these were different types of neural networks. And they have still been investigated to date. They have in common that they are mainly feedforward fully connected networks with 3 or 4 layers. There are papers that confirm the suitability of this architecture for the MSCP interpretation problem [11]. There was still some agreement on the results, with most of the neural networks having p outputs, each indicating a shift in one of the quality characteristics. In other studies, the parameters of NNs differ from implementation to implementation [12]–[15].

In addition to classical neural networks, some authors have also chosen to use SVMs [15], an ensemble of neural networks with different parameters, and then combine the outputs [16]. One of the last approaches was also a combination of four analytical interpretation

methods and a neural network, where the neural network had as input the outputs of the analytical interpretation methods. The task of the neural network was to choose which of the analytical methods to trust and, therefore, to use its output [17].

Despite all the progress made in this area so far, none of the interpretation methods has been proven to perform the best in general.

3.2 MYT method

First of implemented interpretation methods is the Mason, Young, and Tracy decomposition method (MYT) [5]. It is one of the analytical methods based on the T^2 Hotelling statistic. The main idea behind this algorithm is that we can decompose the T^2 Hotelling statistic into orthogonal components [10]. These components then show the effect of each quality characteristic on the resulting T^2 statistic, i.e. the shift. The algorithm works in the following steps:

Step 1: Computes T_i^2 for each variable and removes quality characteristics with significant T_i^2 value.

Step 2: Check whether the T^2 statistic for the remaining subvector (originally measured vector of quality characteristics without removed ones) is in control, if yes, all of the shifted quality characteristics were found. Algorithm ends.

Step 3: Removes all pairs that have significant $T_{i,j}^2$ and again recomputes T^2 for remaining subvector.

Step 4: In the next steps, the algorithm removes all significant triplets(quadruplets,...) until the remaining subvector is in-control

T_1^2 denotes the Hotelling's T^2 statistic for the first variable, and it is computed as:

$$T_1^2 = \left(\frac{x_1 - \mu_1}{\sigma_1} \right)$$

where μ_1 and σ_1^2 denote the mean and variance of the first quality characteristic of the vector $\mathbf{x}$, where the components of $\mathbf{x}$ are all quality characteristics.

The general formula for a set q of k quality characteristics is the following.

$$T_q^2 = (\boldsymbol{x}_q - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1} (\boldsymbol{x}_q - \boldsymbol{\mu}_q),$$

where $\boldsymbol{x}_q$, $\boldsymbol{\Sigma}_q^{-1}$ represents subvector, respectively covariance submatrix obtained from $\mathbf{x}$ respectively $\boldsymbol{\Sigma}$ by selecting only variables from q.

3.3 ANN method

As second interpretation method, I chose the ANN method. In implementing the neural network, I was inspired by the architectures already used for this problem that I found

during a literature research mentioned in Section 3.1. Since the number and selection of inputs, as well as the number of neurons in the hidden layer varies from one implementation to another, I did some small tests at the beginning when designing the network and chose the configuration with the best results.

3.3.1 Architecture

Implemented ANN is Multilayer Perceptron (MLP), which is a fully connected class of feedforward artificial neural networks (ANN), since it has been proven to perform well in the MSCP domain [11]. It contains 3 layers input, hidden, and output. The number of inputs and outputs is identically p corresponding to the number of quality characteristics of the measured sample. The number of neurons in the hidden layer is 16. The sigmoid function is used as the activation layer for the hidden and output layers.

3.3.2 Inputs

Since the form of the measured data did not allow sample sizes larger than one, the input options for the neural network were very limited. Thus, the input vector for the neural network contained p measured values of the quality characteristics of a given sample. I also tried adding a statistic T^2 or SVM score as input $p+1$, but it did not have a positive effect on network performance, rather the opposite. For better network performance, input values were normalized to the interval [0, 1].

3.3.3 Outputs

The choice of outputs was relatively simple, each output corresponded to one quality characteristic. Since a sigmoid activation layer was included with the output layer, the output values were in the interval [0, 1]. Thus, the value in a particular output represented a kind of probability that a shift in that quality characteristic had occurred. In the case where the ANN had information on how many k quality characteristics were shifted, it chose the k quality characteristics with the highest value in the output (probability) as the interpretation of the particular OOC. Otherwise, when the ANN did not have this information, it rounded the output values to 0 or 1 and then all quality characteristics with an output of 1 were chosen as the interpretation of the particular OOC.

3.3.4 Training

The training data generation is shown in Fig. 3.3. The context can be seen in Fig. 4.3. The original unshifted X_A dataset (1000 samples) was generated from the same distribution as the validation data used for the comparison of the method. Similarly, the shifted dataset X_{trn} was created in the same way as the validation data. See Sections 4.1 and 4.2 for more details. The Levenberg-Marquardt algorithm and the MSE loss function were used for training.

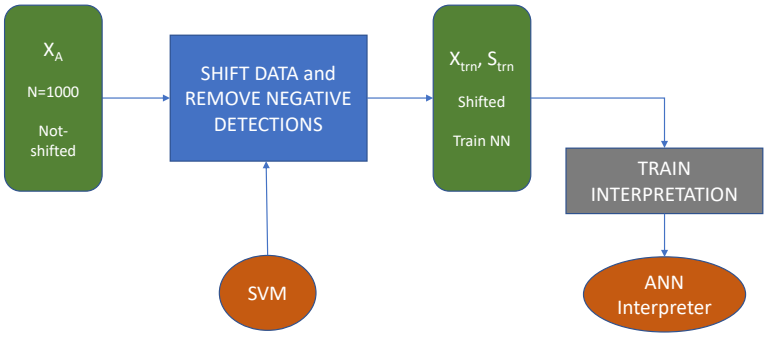


Figure 3.2: Flow chart of training the ANN interpretation method

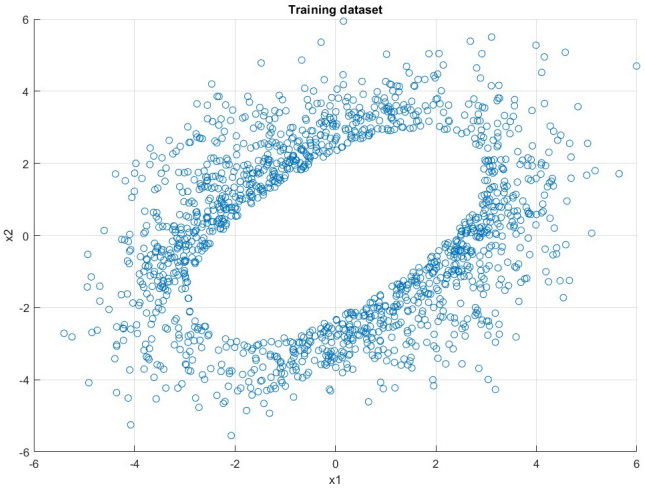


Figure 3.3: An example of the training dataset for ANN interpretation method. Two quality characteristics and a particular experiment setup is assumed here.

3.4 LIME method

Local Interpretable Model-agnostic Explanations (LIME) [18] is a relatively new method that offers a solution to the XAI problem. As the name suggests, it can explain the predictions of classifiers and regressors using a local approximation by an interpretable model. It also treats the original model as a black box, which means that it can be applied to almost any model. The paper illustrates applications, for example, to text

classification with SVM classifiers or to image classification with deep networks. Since this method looks very promising based on the number of citations, I decided to apply it to our case of interpreting positive OOC detection by a One-Class SVM classifier. This has not been tried before, to the best of my knowledge.

3.4.1 Algorithm

In the following steps, I describe how the algorithm works:

Step 1: A measured sample is set as the query point to be explained.

Step 2: Several points X_s are generated from the multivariate normal distribution around the query point, for each point the OOC detection is performed by One-Class SVM classifier and all OOC predictions form the vector Y_s.

Step 3: The weights w_q for each point x_s are calculated based on the Euclidean distance from the query point.

Step 4: An explainable linear model mapping the generated points X_s to the predictions Y_s using the weight values w_q is fitted.

Step 5: The absolute value of the coefficients of the linear model determines the importance of each variable (quality characteristic). The larger the absolute value of the coefficient, the more LIME thinks that there has been a shift in this quality characteristic.

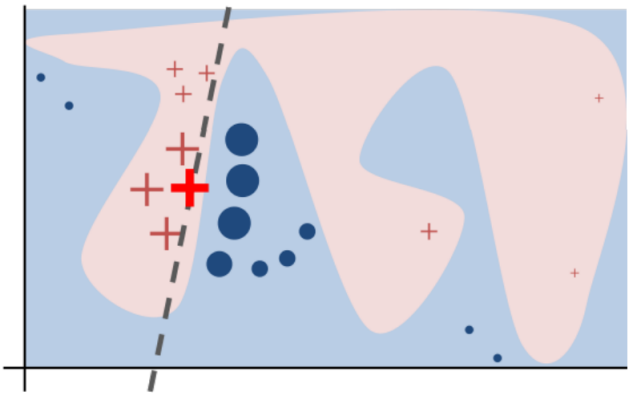


Figure 3.4: An example to present intuition for LIME (adopted from [18])

The intuition for LIME can be seen in Fig. 3.4. The complex black-box model is represented by the background color. The blue color is used for the area where the model predicts one class, and the pink is used for the second. The query point is depicted as a bold red cross. Around the query point, several points are generated and labeled. LIME then fits a linear model, shown as a dashed line, which is locally faithful for the query point.

3.4.2 Determination of the number of shifted QCs

One of the main disadvantages of the LIME method is that it returns only a coefficient for each quality characteristic in the output. The absolute value of the coefficient expresses how much LIME thinks that a given QC is shifted and therefore caused the OOC. However, the values are only meaningful relative to each other; there is no threshold that determines whether or not a particular QC is shifted. For this reason, I created a simple heuristic that sequentially selects the QCs and checks whether the sample is still in the OOC state. The heuristic works as follows:

Step 1: Sort the coefficients of the linear model by absolute value.

Step 2: Select the QC with the largest coefficient and replace its measured value by the mean component from the original in-control data in the available data set.

Step 3: Use One-Class SVM classifier on the updated sample (run OOC detection).

Step 4: If the updated sample is in-control all of the shifted QCs were found. Otherwise, select the QC with the largest coefficient and repeat steps 2-3.

3.5 Performance evaluation

This section describes the criteria for the evaluation and comparison of the interpretation methods. It should be pointed out that I evaluate the performance of the interpretation, where the output is a set, and not the performance of OOC detection, where the output is the binary class label. However, similarly to the evaluation of common classification, interpretation can also be evaluated using accuracy accompanied by sensitivity and specificity.

Let $\mathbf{x}_i$ be the ith vector of quality characteristics assigned to the positive class (OOC) by the OOC detector. Let S_i denote the set indices of quality characteristics that are truly shifted (i.e. ground truth). Interpretation predicts a set $\hat{S}_i$ of indices of quality characteristics that are shifted (i.e. are responsible for the alleged out-of-control state). Note that our criterion must compare those two sets.

3.5.1 Accuracy

First performance measure is accuracy. It is defined as follows:

$$Accuracy = \frac{\sum_i^n \mathbb{1}_{S_i=\hat{S}_i}}{n} \cdot 100\ [\%], \tag{3.1}$$

where the indicator function is defined as

$$\mathbb{1}_f = \begin{cases} 1, & \text{if formula } f \text{ is true,} \\ 0, & \text{otherwise,} \end{cases}$$

and n is the number of samples in the evaluation data set.

The advantage of this performance measure is the output in the form of a single value that directly corresponds to the ratio of successful interpretations, i.e. the number of cases in which correct shifted quality characteristics were determined in a single experiment.

On the other hand, accuracy assumes only the correct or incorrect result of the interpretation (0/1 penalty). It does not take into account the severity of the error. As an example, consider a particular data instance $\mathbf{x}_i$ where a shift has occurred in the first, third, and fifth QC, i.e., its true label is $S_i = \{1, 2, 5\}$. The output of interpretation method A would be the set $\hat{S}_i = \{1, 5\}$ and the output of method B would be $\hat{S}_i = \{3, 4\}$. It is obvious that method A has only forgotten one QC, and its result can be helpful, but method B was completely wrong and its result is useless. For this particular data instance, the accuracy evaluates the responses of methods A and B as equally wrong, although A performs better. Moreover, the accuracy does not say whether the method tends to erroneously output indices of non-shifted quality characteristics or forgets characteristics that are truly shifted. For these reasons, the following additional criteria are used.

3.5.2 Sensitivity and Specificity

To resolve the problem described above, I used two other complementary performance measures. The first one captures "how well" an interpretation method "does not forget" QCs if they are included in the label. It is called sensitivity (or true positive rate) and is defined as follows:

$$TPR = \frac{TP}{P} = \frac{\sum_{i=1}^{n} |S_i \cap \hat{S}_i|}{\sum_{i=1}^{n} |S_i|} \cdot 100\ [\%], \tag{3.2}$$

where the cardinality of the intersection of the target set S_i and the predicted set $\hat{S}_i$ defines the number of true positives for the ith data sample. It is the number of shifted quality characteristics that are correctly predicted as shifted. The sum over all n samples aggregates the true positives over the whole dataset. The cardinality of S_i then defines the number of positives, and its sum gives the number of all occurrences of the shifted quality characteristic in the dataset. An interpretation with small TPR tends to forget some shifted characteristics and gives only a subset of shifted characteristics. However, the interpretation method that always answers that all QCs are shifted will reach the highest TPR value equal to 100%. Therefore, one must add another criterion.

The second criterion captures "how well" an interpretation method "does not add" QCs that are not shifted. It is called specificity or true negative rate and is defined as follows:

$$TNR = \frac{TN}{N} = \frac{\sum_{i}^{n} |(S_i \cup \hat{S}_i)'|}{\sum_{i}^{n} |S_i'|} \cdot 100\ [\%], \tag{3.3}$$

where A' denotes the relative complement of a set A with respect to the set of indices of all quality characteristics $U = \{1, 2, \ldots, p\}$, which is defined as

$$A' = U \backslash A.$$

The cardinality of the complement of the union of label set S_i and the result set $\hat{S}_i$ defines the number of true negatives for ith sample $\mathbf{x}_i$ for a particular interpretation method. The cardinality of complement of S_i' then defines the number of negatives. Similarly to TPR, an interpretation method that always responds that no QC is shifted would have 100 % TNR and therefore one must consider both TPR and TNR.

TPR and TNR are evaluated on all samples in each experiment, thus the factor and denominator in Eq. 3.2 and 3.3 are summed over all samples. In general, the higher TPR and TNR, the better a given interpretation method, but both must be taken into account.

Four variables (P, N, TP, TN) are used in Eq. 3.2 and 3.3. These variables represent well-known categories: Positive, Negative, True Positive, True Negative. However, in our case of interpretation evaluation, they refer to individual quality characteristics. Positive (P) QCs are those that are included in the label of a particular sample, whereas negative (N) QCs are those that are not. Using the result of an interpretation method to the identical sample, two remaining categories can be explained. True positive (TP) QCs are those that are present in both the label and the result, and similarly true negative QCs (TN) are those that are not present in either.

For a better understanding, suppose the example from the previous Subsection 3.5.1, i.e. a sample with a shift present in the first, third, and fifth quality characteristics, and the results of two interpretation methods A and B. The label of this sample is {1, 3, 5}, the result of the interpretation method A is {1, 5} and result of interpretation method B is {2, 4}. Since Positive and Negative QCs depend only on the label, they are the same for both methods. As the label indicates, Positive (P) QCs are {1, 3, 5} and Negative (N) are {2, 4}.

The results of both methods are shown in the following tables, where the QCs are categorized together with the number of QCs in each category.

	TP	TN
QCs	{1, 5}	{2, 4}
Σ	2	2

Table 3.1: Interpretation method A

	TP	TN
QCs	{}	{}
Σ	0	0

Table 3.2: Interpretation method B

If we wanted to evaluate the Eq. 3.2 and 3.3 equations with only this one sample, we would get these values for interpretation method A:

$$TPR = \frac{2}{3} \cdot 100 \approx 66\ \%$$

$$FPR = \frac{2}{2} \cdot 100 = 100\ \%$$

and for interpretation method B:

$$TPR = \frac{0}{3} \cdot 100 = 0\ \%$$

$$FPR = \frac{0}{2} \cdot 100 = 0\ \%$$

According to this, we can say that interpretation method A is better than interpretation method B, because it has higher TPR and TNR at the same time.

It should be noted that all performance measures contain a multiplication by 100 in their definition formulas. This is because I have decided to display the comparison of interpretation methods as percentages for the sake of clarity.

Chapter 4

Experimental comparison of interpretation methods

Since this work is part of a larger project, I tried to keep the real parameters as much as possible in the practical part, which, for example, specify the form of the measured data or the limitations of the measurement itself. First, I focused on generating synthetic data and then processing it, which mainly consisted of shifting it in different directions. Then I implemented the selected interpretation methods, applied them to the prepared data, and compared their results. All the implementation and testing were done in MATLAB.

4.1 In-control data generation

I used the Matlab function *mvnrnd* to generate random points from the specified normal distribution.

4.1.1 Normally distributed data

I primarily used the multivariate normal distribution to generate the data, as it closely approximates the real data distribution, and other papers have also used it for their experiments. The multivariate normal distribution has two input parameters. The first is the p-dimensional vector of means, and the $p \times p$ covariance matrix, where p is the number of quality characteristics (QCs). Without loss of generality, I chose the mean vector as the zero vector. Much more interesting is the choice of the covariance matrix. It determines the correlation between quality characteristics. Since the number of quality characteristics and the correlations between them vary in real data, I tried several different covariance matrices that define the generated data on which I tested the interpretation methods. The covariance matrix must be a symmetric positive definite matrix. I used two types of

covariance matrices inspired by [10]

$$\Sigma_1 = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix},$$

where ρ is from the interval [-1, 1]. In this example, the covariance matrices for $p = 3$ quality characteristics is used, the matrices for different p are generated analogically. It is worth mentioning that the Type I covariance matrix Σ_1 is identical to Type II covariance matrix Σ_2 in case of $p = 2$.

Figure 4.1: Shape of bivariate normal distribution in dependence on ρ

4.1.2 Not-normally distributed data

Other synthetic data was also created using various distributions. First, I created a bivariate normal distribution, where the two variables (QCs) were correlated. The first variable has a gamma distribution with parameters of 3 and 0.5, while the second variable has parameters of 1 and 1. The normal cumulative distribution function (cdf) was then applied to a conventional normal random variable, producing a uniform random variable in the interval [0,1]. Applying the inverse cdf of any distribution F to a random variable $U(0,1)$ results in a random variable whose distribution is exactly F, according to the theory of univariate random number generation. When a two-step transformation is applied to each variable in a standard bivariate normal distribution, dependent random variables with arbitrary marginal distributions are created (gamma distributions defined above) [3]. This dataset is named Copula because of the way it is generated, and this name will also be used to refer to the results that belong to such data.

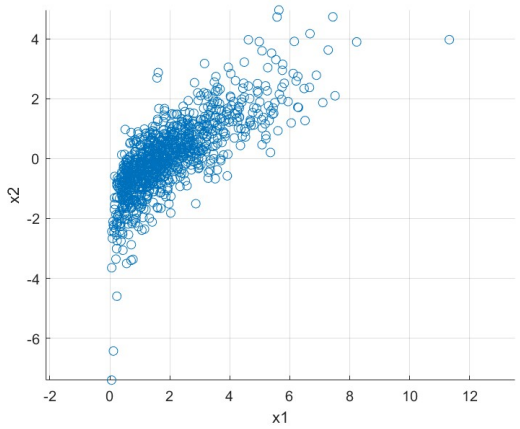


Figure 4.2: Shape of not-normally distributed data

4.2 Data shifting

An out-of-control state of a process is typically manifested by a change of statistical properties of post-process quality characteristics. Both the shift in the mean and the change in the variance can occur. It was not possible to determine from the data or from the knowledge of the production process what the most common possible shifts are and their sizes in the production process if it gets into an out-of-control state. Because of that, I decided to consider a set of a shifts of range of different sizes and in different directions. Moreover, I represent the shift in the manufacturing process as a change in the mean vector μ and do not consider a shift in the variance (i.e. the change of the covariance matrix). Thus, the samples that represent the out-of-control process come from a normal or non-normal distribution with a mean parameter different from those that represent the in-control process.

The shift was implemented by adding a p-dimensional vector $\mathbf{s}$ to the points in the generated in-control dataset X. The vector $\mathbf{s}$ is defined as

$$\mathbf{s} = (s_1, s_2, ..., s_p)^T, \tag{4.1}$$

where

$$s_i = \begin{cases} r, & \text{if there is a shift in } i\text{th QC,} \\ 0, & \text{otherwise,} \end{cases}$$

and r is a random number from the interval $[-3\sigma, -1\sigma] \cup [1\sigma, 3\sigma]$.

Each sample from the dataset X is shifted separately, i.e. a unique shift vector $\boldsymbol{s}$ is generated for each sample. It should be noted that the random values r are selected from a uniform distribution from the interval mentioned above.

I tried to be as general as possible, so I considered all possible sets of shifted quality characteristics. For p quality characteristics there are

$$a = \sum_{k=1}^{p} \binom{p}{k} \tag{4.2}$$

possible sets of shifted quality characteristics in total. The simplest example is for $p = 2$, where a shift could occur in the first, the second or in both quality characteristics. Also the size of shift in each quality characteristics is important, because it defines a specific one shift. Because of that, I used all the points in the X for each possible set of shifted quality characteristics and each point from the set X was shifted individually, i.e. a unique vector $\mathbf{s}$ was generated for each point to perform the shift. So the final shifted set $X_{shifted}$ was a times bigger than the original dataset X. This lead to maximal diversity and generality of the shifts contained in the shifted set $X_{shifted}$. Also during the shifting a label describing particular shift was added to each point in $X_{shifted}$. This approach simulates a situation in which each data sample comes from a different process shift. Although this does not correspond to reality, where the shift is not so variable, it allows us to provide a more robust and more general evaluation and comparison of the methods.

4.3 Experimental scenario

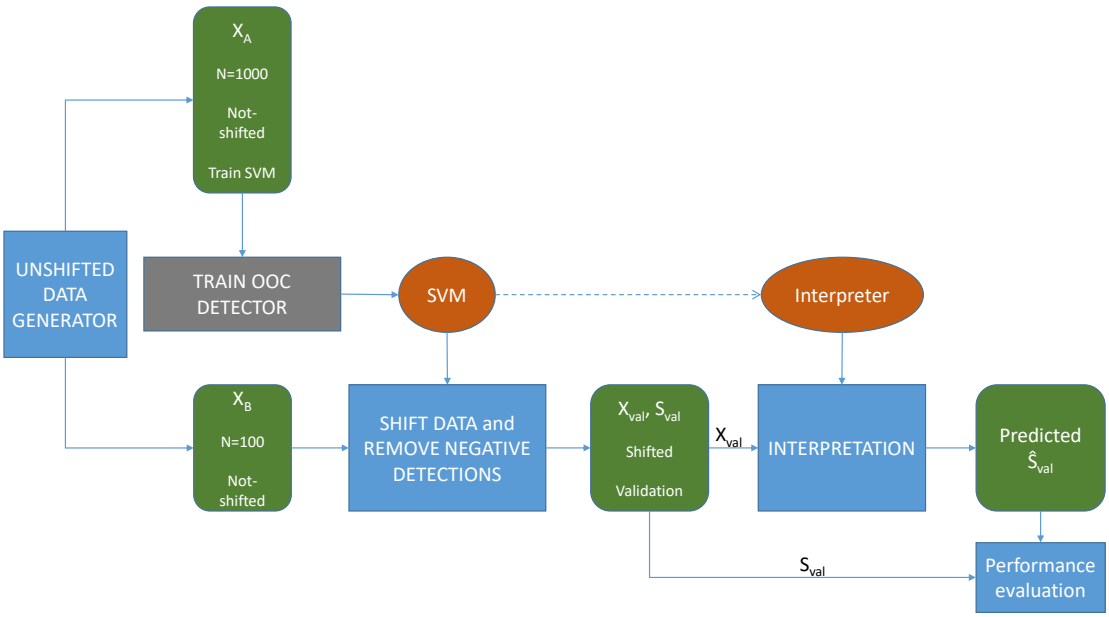


Figure 4.3: Flow chart of the experiment

The main purpose of the experiments is to evaluate and compare different interpretation methods (i.e., determination of the predictor's inputs that are responsible for the current OOC prediction). It is important to note that the goal is not to evaluate the OOC detection, which has been performed in other works ([3], [19]).

4.3.1 Unshifted data generator

First, two sets of in-control data were generated X_A and X_B with the same distribution. The dataset X_A is used to train the out-of-control state detector, which is the One-Class SVM classifier in all experiments.

4.3.2 Creating validation data

In practice, the interpretation of an OOC prediction will be performed only in case of positive outcome of the prediction (regardless of whether the prediction is correct or not). If the positive prediction is incorrect (false positive prediction), it will be interpreted, but the interpretation will be neither needed nor meaningful because the process will actually be in an in-control state. Thus, the validation of the interpretation must be performed only for correct prediction of OOC state, i.e. on data that are true positive OOC predictions.

The dataset X_{val} is used for the validation of the interpretation method. Therefore, I added a shift to the original in-control data X_B to artificially simulate the OOC data. Since the shift is pre-defined and known, I labeled each instance of X_{val} using the set of all shifted quality characteristics S_{val} that is also used in the further validation.

Subsequently, I removed the data which were classified as in-control. This ensures that the resulting X_{val} dataset consists only of true positive OOC predictions.

4.3.3 Validation

In next step X_{val} set is passed to interpretation. An interpretation method estimates which quality characteristics are responsible for positive OOC detection. Thus, for each data instance, the label is estimated in the form of a set of quality characteristics that are shifted. For the whole X_{val} dataset, the interpretation returns the predicted labels $\hat{S}_{val}$, which are afterwards compared with the true labels stored in S_{val}. In the last step, the evaluation metrics are calculated and used for the comparison of the interpretation methods.

The original generated and validation datasets can be seen in Fig. 4.4. The visible gap in the middle of the validation dataset is caused by the aforementioned removal of false negative OOC predictions.

Since the interpretation method is tested on data from shifted process only, all points predicted as in-control (assigned to the negative class) are actually false negative OOC detection. There is no point in interpreting a negative detection of an out-of-control state, since it would not be detected in a real process anyway. The purpose of our experiments is to evaluate interpretation rather than out-of-control state detection.

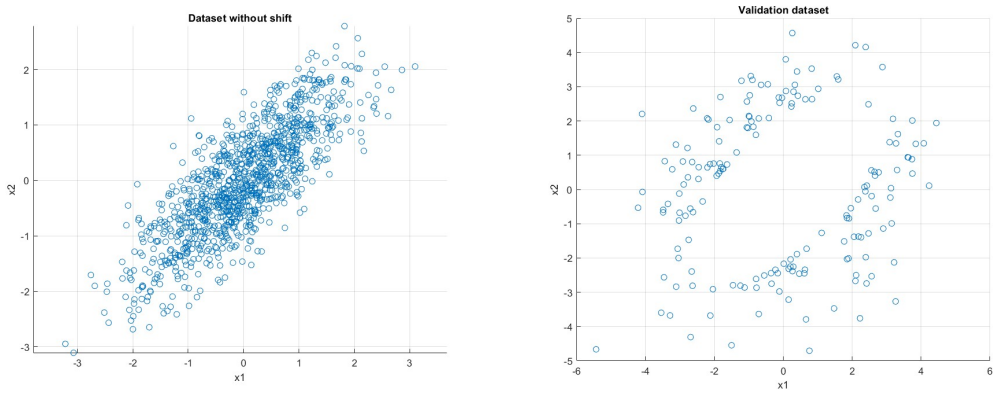


Figure 4.4: Visualization of in-control $X_A(X_B)$ dataset generated from normal distribution in comparison with shifted X_{val} validation dataset

4.4 Results

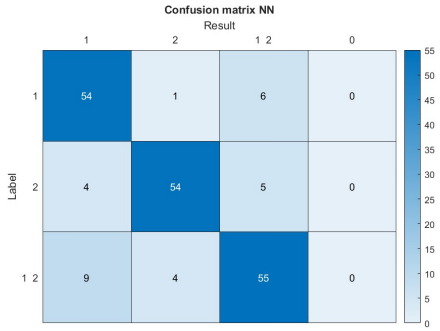
The results for the data generated from the different types of distribution are shown in the tables, selected tests are shown in more detail in the confusion matrices. Since the LIME method only determines how much each quality characteristic contributes to the out-of-control state, but, by default, does not answer the question of how many are shifted, I divided the experiments into two main groups.

4.4.1 Unknown number of shifted QCs

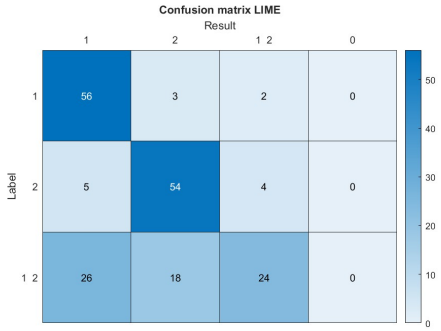
The following results refer to a group of experiments without information on the number of shifted quality characteristics.

Performance measure	Method	ρ -0.7	-0.3	0.1	0.5	0.9	Copula
Accuracy [%]	ANN	**76**	70	68	70	**85**	**71**
	LIME	55	49	52	58	70	41
	MYT	67	**72**	**73**	**71**	73	58
TPR [%]	ANN	**89**	88	**90**	**89**	**93**	**94**
	LIME	70	67	68	72	80	63
	MYT	83	**89**	**90**	87	86	83
TNR [%]	ANN	80	72	65	75	87	48
	LIME	**93**	**91**	**94**	**96**	**89**	**86**
	MYT	78	72	77	79	81	53

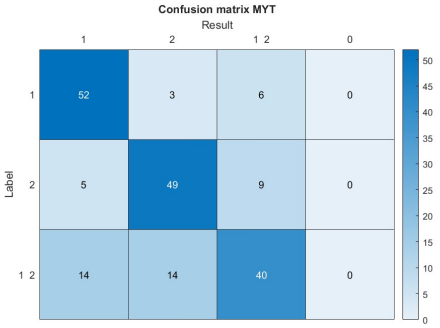
Table 4.1: Comparison of interpretation methods for $p = 2$ and unknown number of shifted QCs



(a) ANN results



(b) LIME results



(c) MYT results

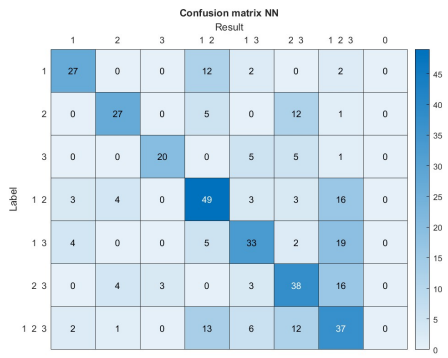
Figure 4.5: Comparison for bivariate ($p = 2$) normally distributed data with $\rho = 0.9$ in confusion matrices

From the results, we can see that the LIME interpretation method has the worst accuracy of all the methods in almost all experimental scenarios. On the other hand, it has the best specificity in all cases and at the same time a not so bad sensitivity, indicating a tendency to select the correct but only a subset of shifted QCs. This is also evident from confusion matrix in Fig. 4.5b.

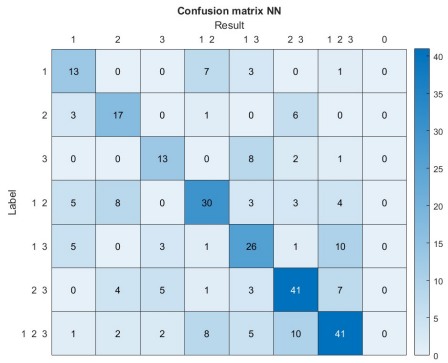
The ANN method performs slightly better than the MYT method when there is a higher correlation between the quality characteristics; however, when the correlation is lower, the results of the MYT method appear to be better. An even greater effect of the correlation between QCs on accuracy is seen in the LIME method, particularly the biggest difference appeared in Fig. 4.6, where is a comparison between Type I and II covariance matrices, again increasing correlation has a positive effect on accuracy as in the ANN method.

Performance measure	Method	ρ (Type of covariance matrix) -0.5(1)	-0.5(2)	0.3(1)	0.3(2)	0.7(1)	0.7(2)
Accuracy [%]	ANN	**58**	**60**	50	51	**61**	**60**
	LIME	41	28	23	26	36	39
	MYT	55	56	**59**	**59**	55	57
TPR [%]	ANN	**90**	**87**	88	**89**	**90**	**90**
	LIME	60	56	54	56	63	61
	MYT	85	**87**	**89**	**89**	85	86
TNR [%]	ANN	74	78	63	66	74	74
	LIME	**94**	**96**	**97**	**95**	**94**	**94**
	MYT	77	73	77	76	70	73

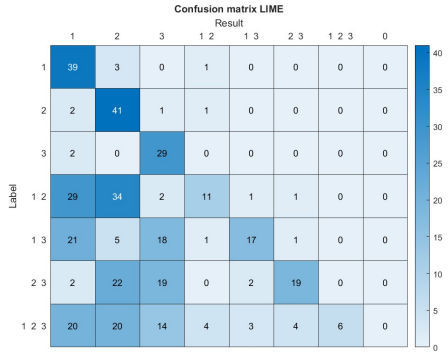
Table 4.2: Comparison of interpretation methods for $p = 3$ and unknown number of shifted QCs



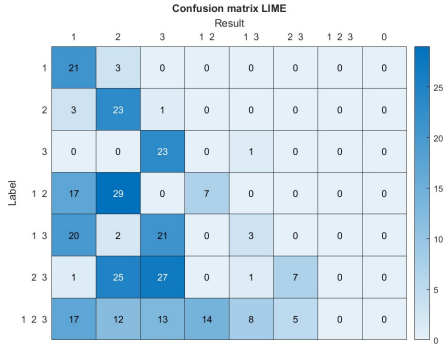
(a) ANN result for Type I covariance matrix



(b) ANN result for Type II covariance matrix



(c) LIME result for Type I covariance matrix



(d) LIME result for Type II covariance matrix

Figure 4.6: Comparison of ANN and LIME method for Type I and II covariance matrices ($p = 3$ and ρ = -0.5)

Performance measure	Method	ρ (Type of covariance matrix) 0.3(1)	0.3(2)	0.8(1)	0.8(2)
Accuracy [%]	ANN	**33**	15	16	17
	LIME	11	11	31	23
	MYT	31	**31**	**35**	**34**
TPR [%]	ANN	**83**	**89**	**92**	**92**
	LIME	48	45	67	58
	MYT	80	80	87	85
TNR [%]	ANN	75	46	47	48
	LIME	**97**	**98**	**92**	**95**
	MYT	76	78	57	66

Table 4.3: Comparison of interpretation methods for $p = 5$ and unknown number of shifted QCs

Performance measure	Method	ρ (Type of covariance matrix) 0.3(2)	0.8(1)	0.8(2)
Accuracy [%]	ANN	**33**	**52**	**46**
	LIME	8	28	22
	MYT	32	37	34
TPR [%]	ANN	**82**	86	**85**
	LIME	43	65	59
	MYT	**82**	**89**	**85**
TNR [%]	ANN	77	84	82
	LIME	**98**	**93**	**95**
	MYT	76	59	66

Table 4.4: Comparison of interpretation methods for $p = 5$ and unknown number of shifted QCs (4x larger datasets)

From Table 4.3 is obvious that the performance of ANN depends on the size of the training data set. For five or more quality characteristics, 1000 samples in the generated in-control dataset are not enough. Enlarging the in-control dataset four times solved this problem as can be seen in Table 4.4.

The sensitivity and specificity of ANN and MYT interpretation methods are similar and basically follow the accuracy values, i.e. where ANN had better accuracy, it typically has better TPR and TNR values and vice versa. It can be seen that TNR is generally 10 to 20

percent worse than TPR for both methods (ANN and MYT). This may have two causes. The first is the tendency of the method to rather evaluate QCs for which it is not quite sure as shifted, leading to a higher TPR but lower TNR. The second possible cause is the imbalance of the samples in the experiment, since one of the extremes (samples with no shift in any of the QCs) is not present in the sample dataset at all, but the other (samples with shift in all QCs) is. This leads to an average number of shifted QCs in the sample higher than $p/2$ (half of the total number of QCs) and therefore a higher probability of higher TPR compared to TNR. Unfortunately, to determine which cause is at fault in a given case, the results need to be examined more closely in the confusion matrix.

4.4.2 Known number of shifted QCs

The following results refer to a group of experiments with information about the number of shifted quality characteristics.

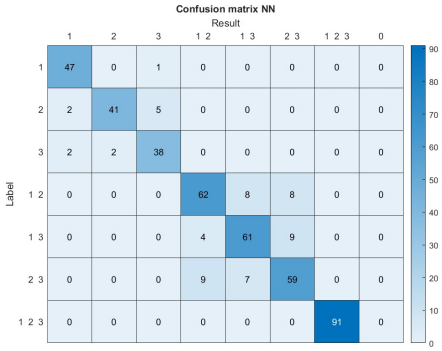
	ρ					
Method	-0.7	-0.3	0.1	0.5	0.9	Copula
ANN	**97**	**96**	**98**	96	**96**	80
LIME	**97**	**96**	97	**98**	**96**	**84**
MYT	75	82	82	81	81	75

Table 4.5: Accuracy [%] of the interpretation methods for $p = 2$ and known number of shifted QCs

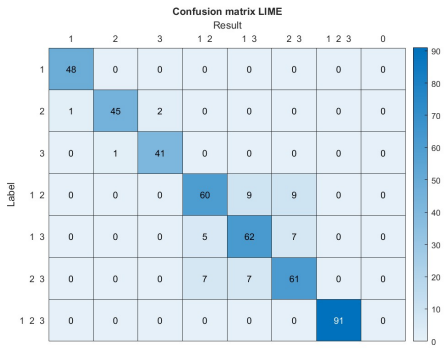
	ρ (Type of covariance matrix)					
Method	-0.5(1)	-0.5(2)	0.3(1)	0.3(2)	0.7(1)	0.7(2)
ANN	**86**	**89**	87	**89**	88	**92**
LIME	**86**	**89**	**88**	**89**	**89**	87
MYT	69	70	72	73	70	70

Table 4.6: Accuracy [%] of the interpretation methods for $p = 3$ and the known number of shifted QCs

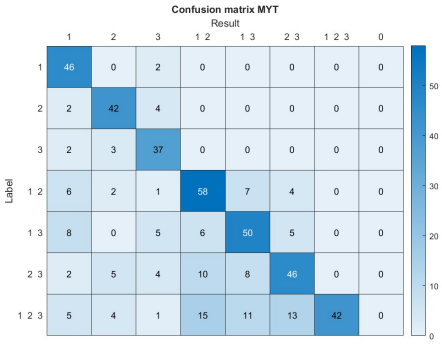
The results in the second group of experiments confirmed our expectations about the LIME method. Its accuracy is in most cases at a level comparable to the ANN method and in many cases even exceeds it which can be seen in Tables 4.5 and 4.6. This shows that the LIME method, presented as a new method of interpretation for One-class SVM classifier, can compete with current interpretation methods. The problem remains in resolving the question of how to identify the correct number of quality characteristics to be determined by the LIME method (my simple heuristic), which seems to not work well.



(a) ANN result



(b) LIME result



(c) MYT result

Figure 4.7: Comparison of methods, when the number of shifted quality characteristics is known ($p = 3$, $\rho = 0.7$, covariance matrix Type I)

The MYT method achieved significantly worse results in the second group of experiments compared to the other interpretation methods. The reason for this can be seen in Fig. 4.7. It is mainly due to the fact that the MYT method is the only one that can select a smaller number of quality characteristics in which a shift has occurred, even if it knows the correct number of them. However, since this set of experiments was primarily intended to verify that the poor results of the LIME method are mainly due to my imperfect heuristic for selecting the correct number of quality characteristics in which a shift has occurred, not the method itself, the comparison with the ANN method is fully sufficient.

	ρ (Type of covariance matrix)			
Method	0.3(1)	0.3(2)	0.8(1)	0.8(2)
ANN	**66**	28	26	26
LIME	65	**65**	**75**	**70**
MYT	48	47	49	50

Table 4.7: Accuracy [%] of interpretation methods for $p = 5$ and the known number of shifted QCs

	ρ (Type of covariance matrix)		
Method	0.3(2)	0.8(1)	0.8(2)
ANN	**66**	**77**	**74**
LIME	64	**77**	68
MYT	49	52	51

Table 4.8: Accuracy [%] of the interpretation methods for $p = 5$ and the known number of shifted QCs (4x larger datasets)

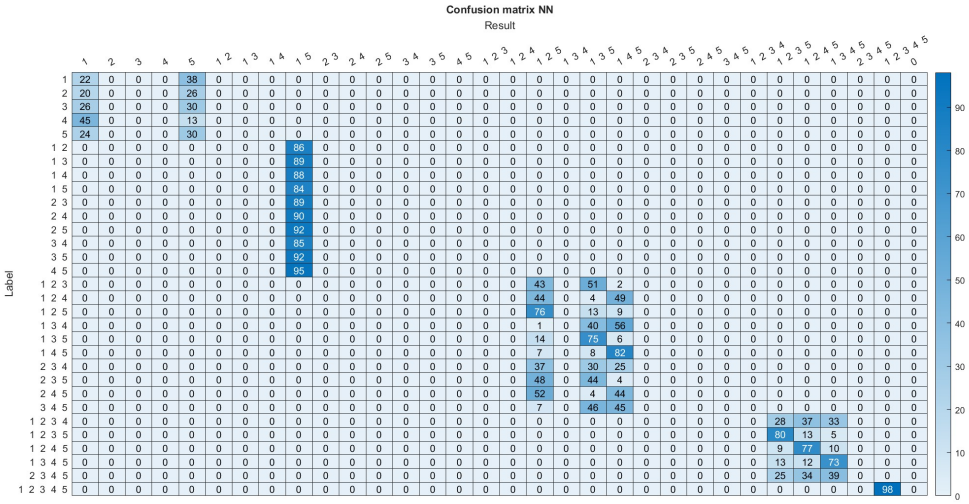


Figure 4.8: Result for ANN method ($p = 5$, $\rho = 0.8$, covariance matrix Type I)

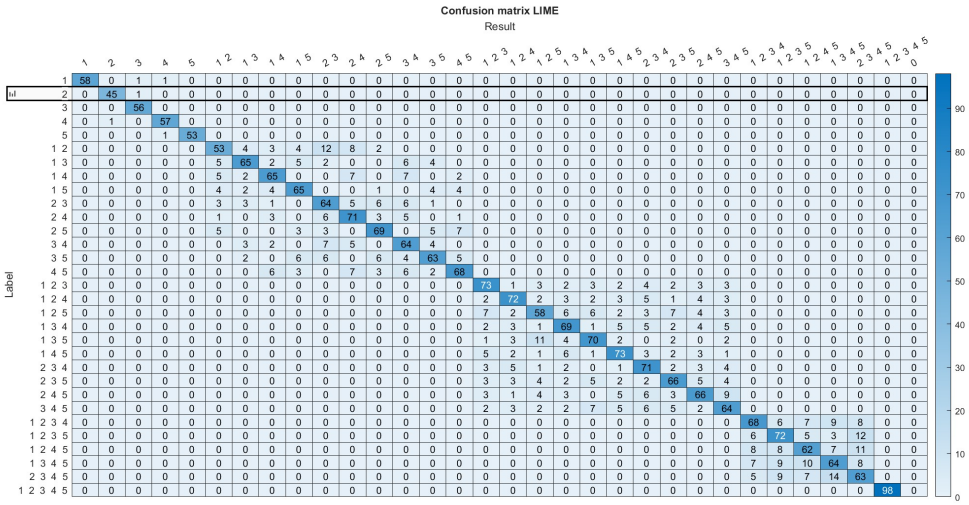


Figure 4.9: Result for LIME method ($p = 5$, $\rho = 0.8$, covariance matrix Type I)

Again, from Table 4.7 is obvious that the performance of the ANN has dropped in the case of five QCs, even in the case where the ANN is provided with information on the

number of shifted QCs. This problem is shown in detail in confusion matrix in Fig. 4.8. In comparison, in Fig. 4.9 the results of the LIME method are shown. Already from the color difference, it can be noticed that most of the samples are on the main diagonal, which indicates high performance of the interpretation method, because on the main diagonal there are samples for which the label fully matches the result of the interpretation method. Again, enlarging the in-control dataset four times solved this problem, as can be seen in Table 4.8.

Chapter 5

Conclusion

As part of this thesis, I conducted a literature research during which I found that the methods used for interpreting out-of-control state within MSPC are divided into two approaches, and these are analytical and computational. Currently, methods from the second approach, specifically different types of neural networks, are mainly being used and investigated. For comparison, I have implemented one analytical method, MYT, and one computational method, ANN. The most significant contribution of this thesis is the last LIME method I used. This method is designed to solve the explainable AI problem regardless of the blackbox model used for prediction. Furthermore, I used three performance measures to compare the results of the interpretation methods, particularly accuracy, sensitivity, and specificity. For the experiments, I generated synthetic in-control data from various distributions and then shifted these to create a validation dataset of samples. The biggest disadvantage of the LIME method is that it only returns coefficients for each quality characteristic in a sample. These coefficients determine how much particular QC is responsible for the out-of-control state, but the user has to define how many of them actually select. Thus, I created a simple heuristic to solve this problem and also divided the experiments into two groups, in first the interpretation methods did not have the information on how many QCs are shifted, in second this information was provided to them.

From the results is obvious that neither method was overall better than the others. In the first group of experiments, the LIME method was the worst in all of the tests, ANN and MYT results were similar according to accuracy. Nevertheless, sensitivity and specificity confirmed that poor LIME results were caused by the heuristic to estimate the number of shifted QCs, not the LIME method itself. These hypotheses were confirmed in the second group of experiments, where the accuracy of the LIME method was similar to the accuracy of ANN, and sometimes LIME even outperformed ANN.

5.1 Future work

Since the LIME method has already proven itself in many other areas, I believe that it has a future in this area as well. This hypothesis was partially confirmed by the results from the second group of experiments. On the other hand, an interpretation method that cannot determine the number of quality characteristics that are shifted is not very useful. Thus, the scope for future work is obvious. It is necessary to devise some smarter method to determine the number of quality characteristics that are shifted under the LIME method, so that this interpretation method is competitive with existing interpretation methods for real applications where there is no prior knowledge of the number of quality characteristics that are shifted.

Bibliography

[1] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2020.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers", in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[3] M. Macas, D. H. Nguyen, and C. Panuskova, "Support vector machines for control of multimodal processes", in *International Conference on Soft Computing and Pattern Recognition*, Springer, 2021, pp. 384–393.

[4] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence", *Science Robotics*, vol. 4, no. 37, eaay7120, 2019.

[5] R. L. Mason, N. D. Tracy, and J. C. Young, "Decomposition of t 2 for multivariate control chart interpretation", *Journal of quality technology*, vol. 27, no. 2, pp. 99–108, 1995.

[6] S. W. Choi, E. B. Martin, A. J. Morris, and I.-B. Lee, "Fault detection based on a maximum-likelihood principal component analysis (pca) mixture", *Industrial & engineering chemistry research*, vol. 44, no. 7, pp. 2316–2327, 2005.

[7] B. Murphy, "Selecting out of control variables with the t2 multivariate quality control procedure", *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 36, no. 5, pp. 571–581, 1987.

[8] N. Doganaksoy, F. W. Faltin, and W. T. Tucker, "Identification of out of control quality characteristics in a multivariate manufacturing environment", *Communications in Statistics-Theory and Methods*, vol. 20, no. 9, pp. 2775–2790, 1991.

[9] P. E. Maravelakis and S. Bersimis, "The use of andrews curves for detecting the out-of-control variables when a multivariate control chart signals", *Statistical Papers*, vol. 50, no. 1, pp. 51–65, 2009.

[10] S. Bersimis, A. Sgora, and S. Psarakis, "Methods for interpreting the out-of-control signal of multivariate control charts: A comparison study", *Quality and Reliability Engineering International*, vol. 33, no. 8, pp. 2295–2326, 2017.

[11] S. T. A. Niaki and B. Abbasi, "Fault diagnosis in multivariate control charts using artificial neural networks", *Quality and reliability engineering international*, vol. 21, no. 8, pp. 825–840, 2005.

[12] H. Hwarng, "Toward identifying the source of mean shifts in multivariate spc: A neural network approach", *International Journal of Production Research*, vol. 46, no. 20, pp. 5531–5559, 2008.

[13] L.-H. Chen and T.-Y. Wang, "Artificial neural networks to classify mean shifts from multivariate $\chi 2$ chart signals", *Computers & Industrial Engineering*, vol. 47, no. 2-3, pp. 195–205, 2004.

[14] F. Aparisi, G. Avendaño, and J. Sanz, "Techniques to interpret t 2 control chart signals", *IIE Transactions*, vol. 38, no. 8, pp. 647–657, 2006.

[15] C.-S. Cheng and H.-P. Cheng, "Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines", *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 198–206, 2008.

[16] J. Yu, L. Xi, and X. Zhou, "Identifying source (s) of out-of-control signals in multivariate manufacturing processes using selective neural network ensemble", *Engineering Applications of Artificial Intelligence*, vol. 22, no. 1, pp. 141–152, 2009.

[17] S. Bersimis, A. Sgora, and S. Psarakis, "A robust meta-method for interpreting the out-of-control signal of multivariate control charts using artificial neural networks", *Quality and Reliability Engineering International*, vol. 38, no. 1, pp. 30–63, 2022.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?"explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] S. He, W. Jiang, and H. Deng, "A distance-based control chart for monitoring multivariate processes using support vector machines", *Annals of Operations Research*, vol. 263, no. 1, pp. 191–207, 2018.