The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2019

Volume 5, Pages 1-13

ICRETS 2019: International Conference on Research in Engineering, Technology and Science

Applied Machine Learning: Reconstruction of Spectral Data for the Classification of Oil-Quality Levels

Marco STANG
Karlsruhe Institute of Technology

Martin BOHME
Karlsruhe Institute of Technology

Eric SAX
Karlsruhe Institute of Technology

Abstract: In modern complex systems and machines - e.g., automobiles or construction vehicles - different versions of a "Condition Based Service" (CBS) are deployed for maintenance and supervision. According to the current state of the art, CBS is focusing on monitoring of static factors and rules. In the area of agricultural machines, these are for example operating hours, kilometers driven or the number of engine starts. The decision to substitute hydraulic oil is determined on the basis of the factors listed. A data-driven procedure is proposed instead to leverage the decision-making process. Thus, this paper presents a method to support continuous oil monitoring with the emphasis on artificial intelligence using real-world spectral oil-data. The reconstruction of the spectral data is essential, as a complete spectral analysis for the ultraviolet and visible range is not available. Instead, a possibility of reconstruction by sparse supporting wavelengths through neural networks is proposed and benchmarked by standard interpolation methods. Furthermore, a classification via a feed-forward neural network with the conjunction of Dynamic Time Warping (DTW) algorithm for the production of labeled data was developed. Conclusively, the extent to which changes in hyper-parameters (number of hidden layers, number of neurons, weight initialization) affect the accuracy of the classification results have been investigated.

Keywords: Machine learning, Neural networks, Spectral analysis

Topic

The total cost of machines – e.g., automobiles or construction vehicles – is composed of the acquisition costs and the costs during operation. The operation costs can further be partitioned in maintenance and repair cost [2]. For construction vehicles, maintenance costs are decisive and therefore one of the significant factors of cost estimation. To make cost estimates more reliable and accurate, "Condition Based Service" (CBS) is employed. According to the state of the art, static parameters such as operating hours, mileage and number of engine starts are mainly used in the CBS [3]. The parameters are acquired by suitable sensors. For oil condition sensors have been the subject of numerous research projects since the 1980s [4]. The development and application of these sensors can be divided into the areas of hydraulic [5], engine [6] and lubricating oils. A general distinction of the measuring systems are oil sensors, which measure only one oil quality parameter and so-called multi-parameter based oil sensors.

Multi-parameter sensor technology is mainly used in a laboratory environment for determining oil quality. Laboratory analysis provides a variety of meaningful chemical and physical parameters. For example, the electrical conductivity, viscosity and dielectric constant of the oil to be investigated is considered. An overview of these laboratory testing methods can be seen in [1]. The advantage of a laboratory analysis is the exact determination of the oil quality. Laboratory analysis, however, causes additional costs, as oil samples must first be taken from the machines which are operated far away from any laboratory in a rough environment. Due to

this environment a replacement for media such as fluids is mandatory because slowly dirt, particles, air, water or other matters that influence the life time and the correct behavior of the machine.

Currently there is a strong trend in online and on-site sensor technology. The deployment of an online sensor for detecting the oil quality and thus the optimal replacement time of hydraulic oil , would not only eliminate the need for expensive laboratory analysis, but would also have the following benefits [5]:

- Increasing the life time of the machine
- Strategic planning of oil change and maintenance work
- Saving in oil and disposal costs
- Reduction of machine downtimes and maintenance costs

However, when is the oil replacement optimum exactly reached? Machine learning, especially neural networks, are designed to find these optima based on the provided data. Thus, this paper presents a method to support continuous oil monitoring with the emphasis on machine learning using real-world spectral oil-data. Thus, this paper presents a method to support continuous oil monitoring with the emphasis on machine learning using real-world spectral oil-data, Concrete one oil quality parameter is analyzed, namely the transmission of light in the visible and near infrared range. Laboratory-like analysis technologies on a much smaller scale and at reduced costs can be achieved.

Technical Approach

CRISP-DM (CRoss-Industry Standard Process for Data Mining) has been developed since 1996 within the context of an EU-funded project (Participants include DaimlerChrysler and SPSS). Since then it has become the dominant process framework for data mining. The individual sub-steps of the process are explained in general using Figure 1. Then the steps Data Understanding, Data Preparation, Modeling and Evaluation for the use case oil analysis are considered.

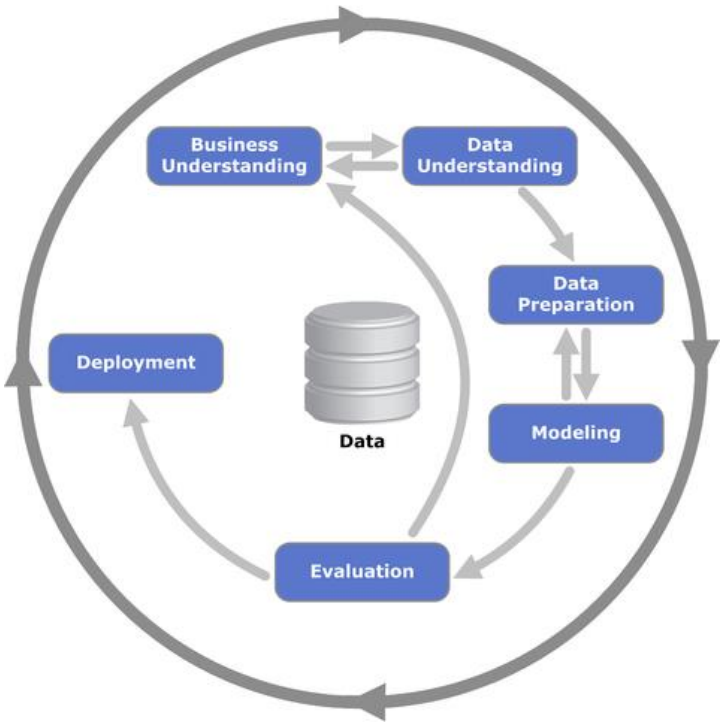


Figure 1. The CRISP-DM life cycle from

1. **Business understanding** – The initial phase focuses on understanding the project and derives requirements from this. In the beginning, the problem is viewed from a business perspective and then converted into a data mining problem. From these specifications, a preliminary project plan is formulated.

2. **Data understanding** – The data understanding phase proceeds with the collection of data and continues with activities that contribute to the knowledge of the data. The data is examined for its quality (noise, statistical significance). First insights or interesting subsets of a large dataset should be identified. The steps business and data understanding are closely linked and iteratively run in parallel. For the formulation of the data mining problem, it is necessary to have at least a fundamental comprehension of the data.

3. **Data preparation** – The data preparation phase is used to construct the ready-to-work data set. The data set defines the data used to train the model. Data preparation tasks are most likely applied several times and in no prescribed order.

4. **Modeling** – In this phase different model techniques are selected and applied. The parameters of the methods have to be optimized. However, there are several techniques to solve the same data mining problem. Some methods require different data formats. Therefore a step back to data preparation can be necessary.

5. **Evaluation** – Before moving on to the final phase of the project, it is essential to evaluate the selected model thoroughly. The analysis also includes the previous steps that resulted in the model. A key objective is to determine if crucial business problems have not been adequately addressed.

6. **Deployment** – The creation of a model is generally not the end of a project. The knowledge that has been generated through the implementation must be collected and organized in a way that is understandable to the customer. Depending on the requirements, the deployment phase can be a report but can also be an implementation of a repeatable data mining process. [8]

The structure of the CRISP-DM serves as a template for this paper. Subsequent the items Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation are discussed for the concrete task of hydraulic oil monitoring and maintenance

Business Understanding

In order to measure the oil quality, the hydraulic oil to be tested is withdrawn from the circulation and tested in an external laboratory. This offline process is time-consuming and costly. The timing of the inspection is based on an estimate by means of static factors. The primary objective of the project is the continuous online analysis of the oil quality using a limited number of LEDs with different peak wavelengths. No complete spectrometer can be used on the vehicle for space and cost reasons. This results in the following central question the data mining process is supposed to answer:
How well do neural networks perform for the reconstruction of a hydraulic oil spectrum and the classification of hydraulic oil in quality classes?

Data Understanding

The data-understanding step covers the data collection and the review of available data. In this step, promising wavelengths were identified using different hydraulic oils and how real changes in the hydraulic oil are expressed in the spectral data of the hydraulic oils.

Data acquisition was done using a breadboard model. The breadboard model is designed to provide a spectral analysis (based on polynomial smoothing according to Savitzky and Golay [9]) of selected oils and serves as a proof of concept. Furthermore, real data of aged oils can be recorded, and on this basis, algorithms for further analysis can be employed.

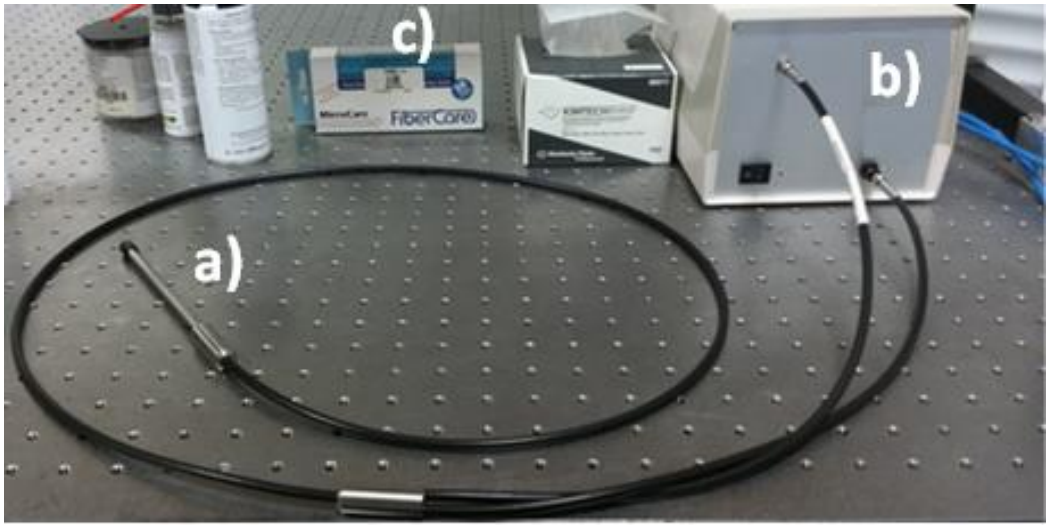


Figure 2. Breadboard model (b) with dipping probe (a) and cleaning utensils (c)

The data acquisition components are shown in Figure 3. The breadboard model (b) consists of two specific spectrometers covering different frequency ranges. The model is optimized for the measurement range from λ = 300 to 1650 nm. One of these spectrometers operates in the VIS range (λ = 360 to 600 nm), the other in the NIR range (λ = 600 to 1650 nm). The internal broadband light source of the breadboard model emits light that is coupled into a multimodal fiber optic cable and connected to a submersible probe. The light reflected by the immersion probe is then directed onto the two internal spectrometers, which generate a transmission wavelength spectrum. Depending on the conditions of the oil sample, the transmission changes at specific wavelengths and thus indicates the quality of the oil sample. The breadboard model has the advantage that it can be used to test oil under laboratory conditions. Ergo it is possible to generate and store comparative data of old oil, particle contaminated oil and fresh oil. The spectral data can then be further processed on this basis by other third-party software. By the use of the breadboard model, measurements were performed to investigate the oil quality. Figure 3 shows the spectral sensitivity of a selection of oils from different aging stages. It can be seen that oils of different ages also generate different spectral sensitivities, especially in the VIS range (up to 1200nm).

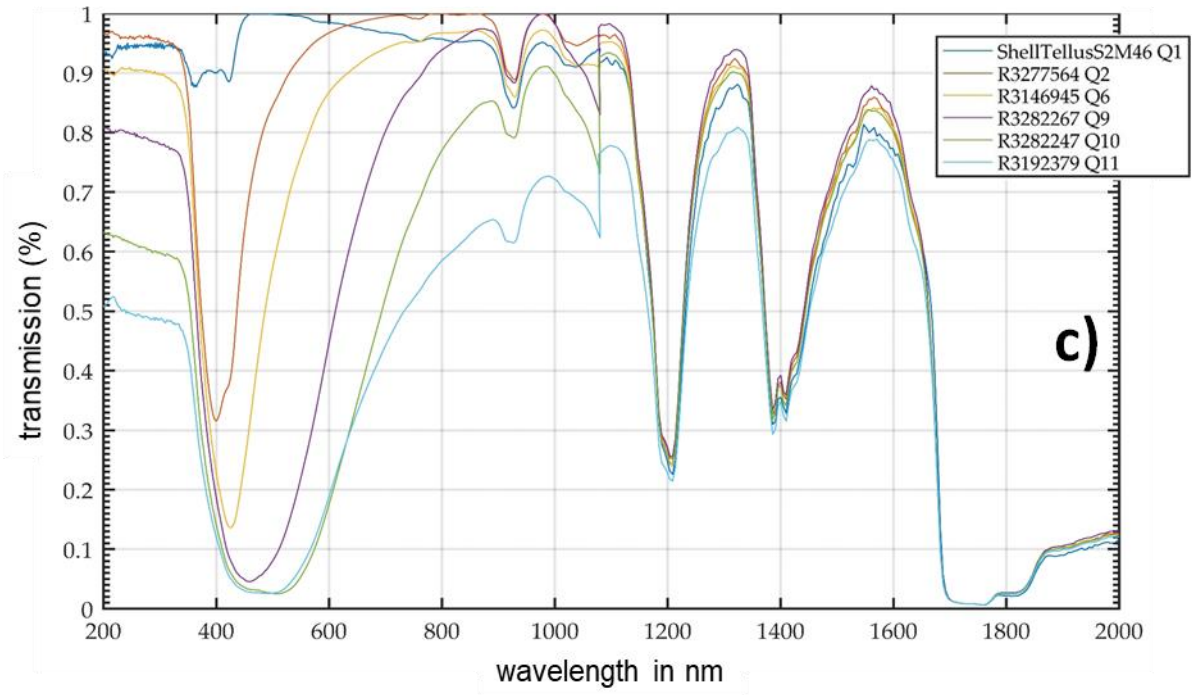


Figure 3. Spectral responses of oil samples of the same type with varying stages of aging (Q2-Q11) compared to fresh oil (Q1)

In Figure 4 the differential or reference measurements of two oils are shown. The analysis implies that one of the oils is the reference compared to the second oil. In the measurement, the oil marked with a) represents the reference signal and b) represents the old oil. The analysis can be related to aging (oxidation) of hydraulic oil, which is indicated by a change in the color of the oil. The reviewed oil types can be distinguished visually and by the spectral data at a wavelength in the range of 450 nm to 600 nm (**Hata! Başvuru kaynağı bulunamadı.**). In this wavelength range, a peak is discernible, which shows the correlation of the different colors. Therefore, it is possible to detect different oil color tones, which can be related to the aging status of hydraulic oil by using the breadboard model.

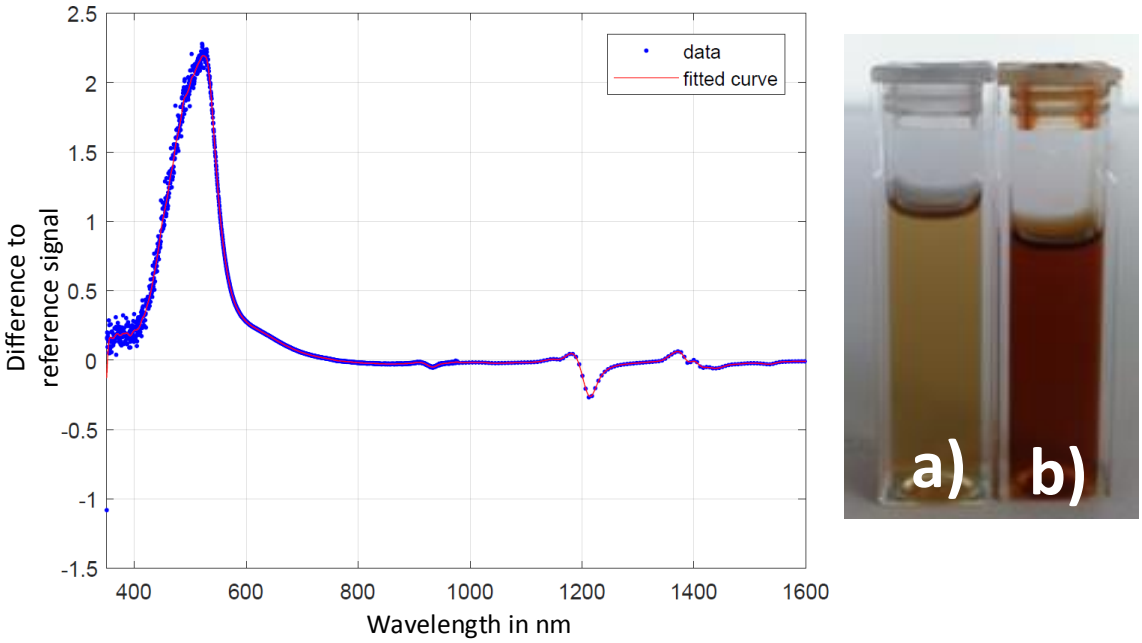


Figure 4.Differential signal of Shell Spirax (a) versus Panolin HLP Synth (b)

Data Preparation

Before we start with the actual Data Preparation, the term Machine Learning will be considered at first. The idea of Machine Learning was pioneered by Arthur Samuel in 1959 [10]. This idea was adopted by Tom Mitchell [11] and extended to the following formal definition:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

In this paper, the program is a tool to separate hydraulic oils into classes of different quality. Hence the task T is of the type classification. The program will learn from examples of good, medium or bad hydraulic oils – the experience E. The performance P is defined as the ratio of correctly classified oil samples. This performance score is called accuracy and is often used in classification tasks, as it is a supervised learning approach. The main tasks T are shown in Figure 5.

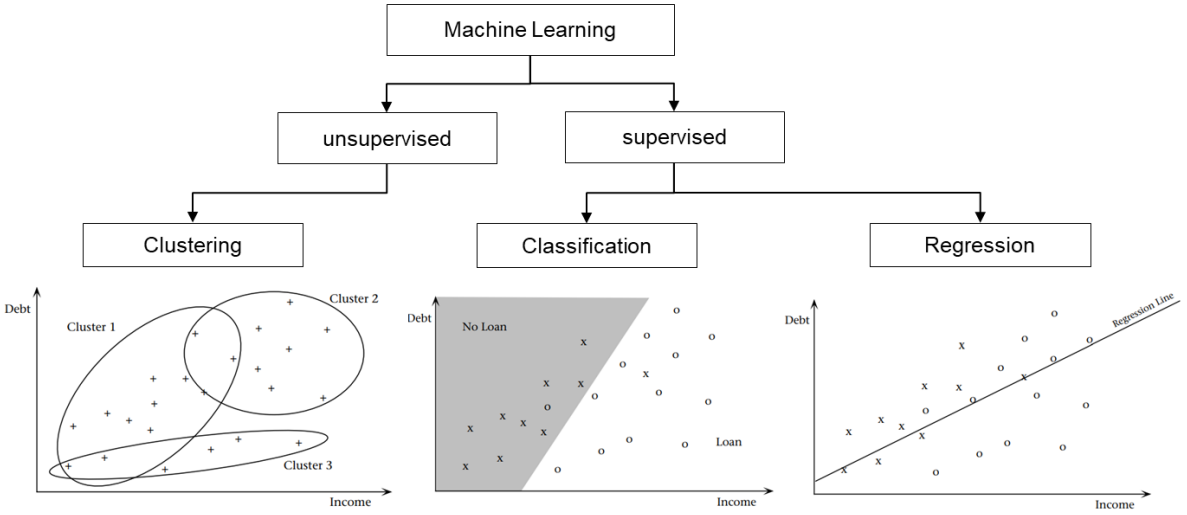


Figure 5.Categorisation of machine learning tasks

The tasks of machine learning are divided into two subgroups of learning - unsupervised and supervised. Occasionally a third subgroup reinforcement learning is mentioned, but since there is no relevance for the topic of the paper, this subgroup is neglected.

Supervised learning is based on data sets consisting of data points, but additionally, each data point is associated with a label or target. Neural networks are widely used in this category because they need a goal for the optimization of their weights. The supervised learning method can be further separated into the learning tasks classification and regression. Classification determines the class into which the dependent member belongs, based on one or more independent variables. Classification will produce discrete relationships. Related to the oil analysis this means: only membership in one of the classes good, medium or bad is possible. The second group of tasks is the Regression analysis; it is a form of predictive modeling technique that explores the interaction between a target and independent variables (predictor). In the present paper, the regression is utilized to predict the course of a spectral oil curve with only a reduced number of variables.

Unsupervised learning implies the analysis of unlabeled data. The learning task is to recognize useful new patterns in a data set. Typically, a dataset has too many characteristics or features for a human to perceive or the underlying structure is too complicated for human understanding. Frequently unsupervised learning is used to perform clustering, which divides the data set into clusters of similar data points. Dynamic Time Warping (DTW) is used as the clustering algorithm in this chapter and will be described below.
For the purpose of using neural networks for classification and regression it is necessary to label the data. DTW is a well-known algorithm for detecting the optimal match between two time-dependent curves which makes it highly applicable for our use case. The sequences are stretched in a non-linear way to match each other. DTW was initially invented to synchronize different speech patterns automatically. In this case, the DTW is utilized to compare the similarity between a defined reference oil curve and a measured oil curve. The curve is not scaled over time but over the frequency. The extent of stretching can be used to create a distance metric which is used to label the oil data set automatically. Three reference curves are specified (good, medium, bad) and the measured curves are compared by the DTW algorithm. The pair of curves with the smallest distance value is assigned to the matching category (Figure 6).

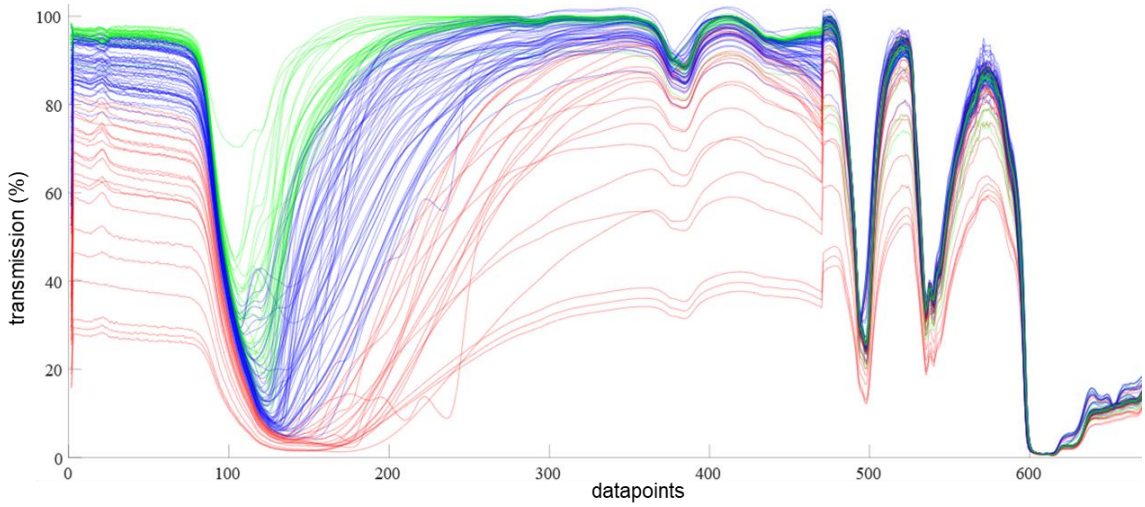


Figure 6. DTW labeled hydraulic Oil DataSet (red=bad Quality, blue=medium Quality, green=good Quality)

Modeling

Deep Learning describes neural networks with a large number of layers and neurons. Deep neural networks have emerged from the field of machine learning in recent years and appear to be a promising approach for this data science problem. For the reconstruction of the spectral transmission waveforms of hydraulic oils and the classification of oil, a Multi-Layer Perceptron (MLP) was used.

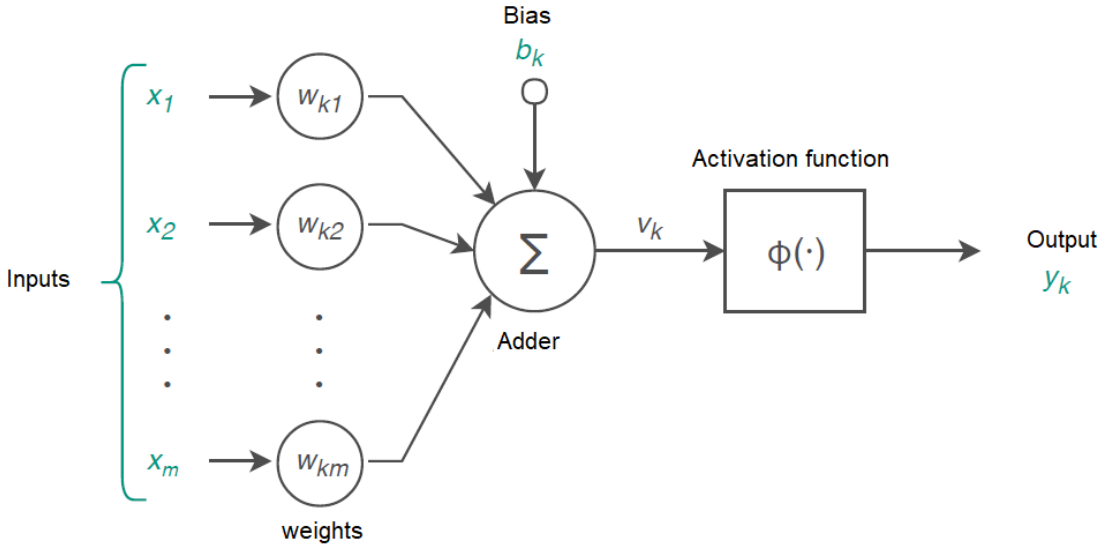


Figure 7. Illustration of an artificial neuron (based on [13])]

An artificial neuron is considered as an information-processing unit calculating an output value from one or more input values. Instead of specifying this mapping function, it is acquired by the neural network from data. The basic structure of a neural network consists of at least three parts:

1. A series of inputs multiplied by weights. Each input value x_j is multiplied by the weight w_{kj}.
2. An adder is accumulating the weighted inputs.
3. An activation function that restricts the output value range and implements a nonlinear property to the system.

The weighted sum of the input values is referred to as u_k and is defined as follows.

$$u_k = \sum_{j=1}^{m} w_{kj}\, x_j \tag{1}$$

Also, the artificial neuron shown in Figure 7 contains a bias b_k. The bias b_k can be defined as a positive or negative fixed value to increase or decrease the input value v_k of the activation function. The value v_k results from the weighted sum of the input signals xj, added with the bias b_k.

$$v_k = \sum_{j=1}^{m} w_{kj}\, x_j + b_k \tag{2}$$

The output value of the neuron y_k is the result of the nonlinear activation function φ applied to v_k.

$$y_k = \varphi(v_k) = \varphi(u_k + b_k) \tag{3}$$

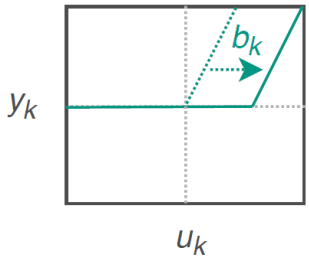


Figure 8. ReLU function

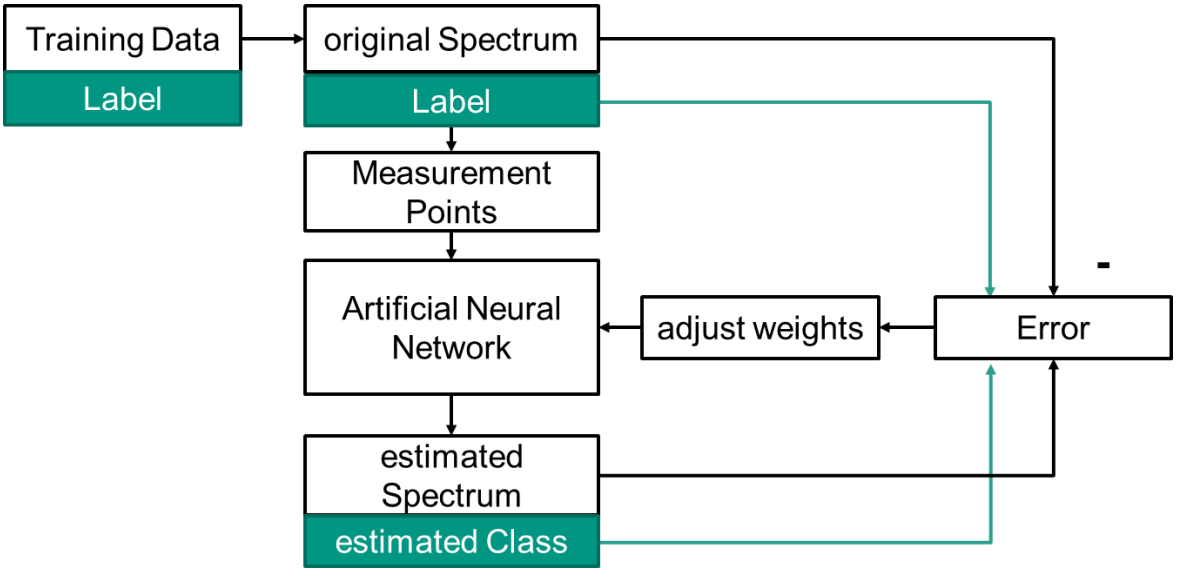
Currently, the most preferred non-linear activation function is the ReLU function [14] (see Figure 8). The abbreviation stands for "Rectified Linear Unit" and describes a function that accepts the value 0 for input values < 0 (inactive) and behaves according to the identity function (active) for values > 0. The threshold value can be adjusted in the neuron via the bias b_k (see Fig. 1). Consequently, the output of the neuron is only activated from a specific value of the weighted sum u_k. Fig. 8 shows a negative bias value bk. Due to the negative bias, the neuron will need a higher weighted sum u_k to be active.

Training Process

The development process of a neural network is subdivided into a learning phase and the inference phase - the application of the neural network. Figure 9 illustrates the flowchart of the learning process used for a neural network. It is divided into the learning process for reconstruction and the learning process for classification (green).

Regression

The regression task aims to learn the structure of the spectral oil data and then to estimate this curve with far fewer measuring points (in our example 6 measuring points). The reference spectral data consists of approx. 1700 measuring points, which were captured by the breadboard model (Figure 2). Each of these measuring points represents a wavelength between 200 nm – 1900 nm. Out of these points, 6 measuring points were randomly selected and presented to the neural network as an input. Each of the measuring points represents a peak wavelength of the LEDs. The neural network serves for the approximation of the original spectrum, containing 1700 measuring points, out of the limited inputs. For this purpose, the estimation of the network is iteratively compared with the original range. In the first iterations, the error (deviations between the frequencies) will be significant. The error is minimized by adjusting the weights of the neural net. The number of weights to be adjusted can be very large. For that reason, a Scaled Conjugate Gradient training algorithm was used. Once the training is finalized, the artificial neural networks can approximate oil spectra from six data points. Note: the exact number of input data points is not fixed, but can be adjusted



.

Figure 9. Flow chart of reconstruction and classification (green) using a neural network

Classification

The learning process of classification is similar to the reconstruction task just described. However, the difference is that the neural network does not estimate oil spectra but rather the oil's assignment to one of the labeled oil classes (good, medium, bad). In the learning phase, spectral oil data is presented as input to the neural network. The neural network estimates the class affiliation of the oil sample data and iteratively adjusts the weights in the case of an error. At the end of such a training process, the neural network must be able to categorize unfamiliar spectral oil data into the given set of classes.

Topology

A particular class of a feedforward artificial neural network, a **M**ulti-**L**ayer **P**erceptron (MLP), was used for the classification and reconstruction task. The exact topology is described below and shown in Figure 10. The neural networks consist of seven layers: one input and one output layer, two drop-out layers and three fully connected (dense) layers. A widely known problem with neural networks is overfitting: meaning the network does not learn the general structure of the data but memorizes an exact representation of the data. To avoid overfitting drop out layers were introduced. In addition, a classical division of the data into training, testing, and validation with 10-fold cross validation was applied. As an activation function, the ReLu function shown in Figure 8 was used. Adam optimizer was utilized for optimizing the weights of the net.

Input Layer, 6 dim	Input Layer, 6 dim
Dense Layer, 128 dim, ReLu	Dense Layer, 128 dim, ReLu
Dropout Layer	Dropout Layer
Dense Layer, 128 dim, Sigmoid	Dense Layer, 128 dim, Softmax
Dropout Layer	Dropout Layer
Dense Layer, 128 dim, Sigmoid	Dense Layer, 128 dim, Softmax
Output Layer, 725 dim, Sigmoid	Output Layer, 3 dim, Softmax

Figure 10. Topology of the regression neural network (left) and classification neural network (right)

Evaluation

The evaluation step of the CRISP-DM model is divided into two parts. As described in the chapter Business Understanding, the objective of the Evaluation is to answer whether neural networks perform well for the classification and reconstruction of oil data. Ensuing these two questions will be answered separately.

Reconstruction

The results of the regression of one curve are presented in **Hata! Başvuru kaynağı bulunamadı.** and Figure 12. **Hata! Başvuru kaynağı bulunamadı.** displays the reconstruction with a neuronal net with three fully connected layers. The number of neurons in each layer is respectively 80, 50 and 20. The blue curve is the output of the neural network; in orange, the original curve is drawn (target). We can see that the neural network already delivers weak results, due to flaws in terms of noisy behavior. This characteristic can be seen even more clearly if several curves are contemplated as seen at the bottom of Figure 11. Examining the area from 250nm to 1200nm an exact reconstruction is necessary by reason of high variance of input data. For this reason, the number of neurons was successively increased.

Figure 12 visualizes the modified neural network, with an increased number of neurons per layer. The output of the neural network (blue curve; top) is more even, and the noisy behavior has disappeared. In addition, the output is very close to the target. This behavior can also be confirmed by looking at several different oilsamples (bottom). A clear distinction can be made between different types of oil.

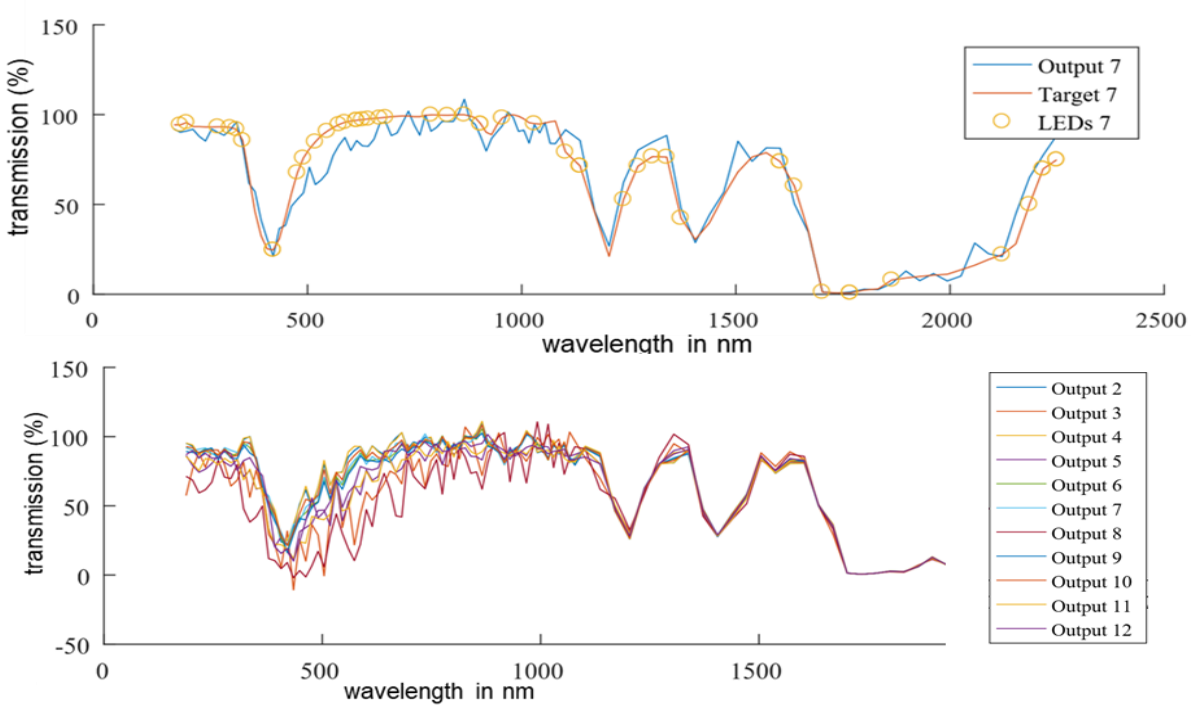


Figure 11. Neural Net with 80, 50, 20 neurons in the hidden layers (top) and superposition of 11 reconstructed spectral curves (bottom)

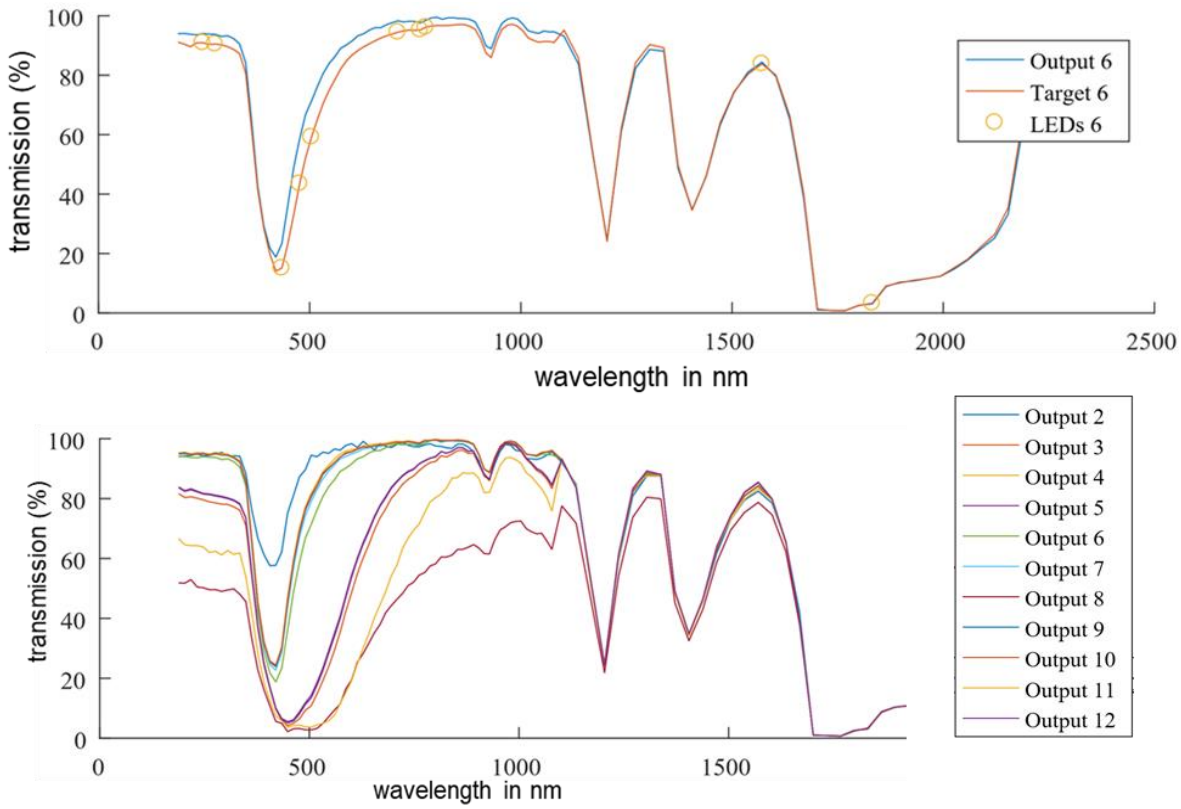


Figure 12. Neural Net with 200, 200, 200 neurons in the hidden layers superposition of 11 reconstructed spectral curves (bottom)

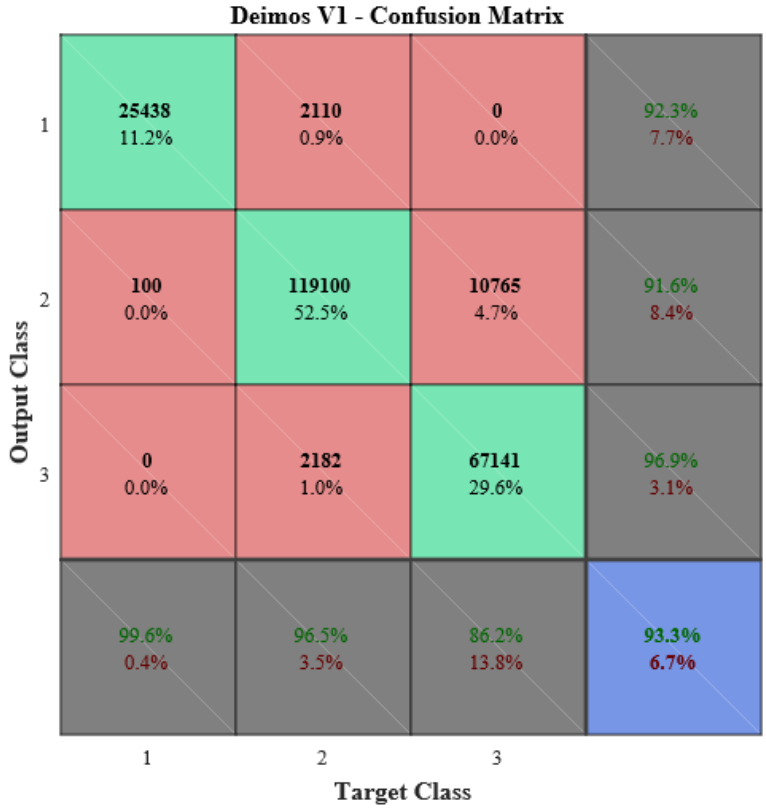


Figure 13.Confusion Matrix for classification in oil classes (good, medium, bad)

Classification

A confusion matrix (Figure 13) is used to describe the performance of a classification model. Since we have three different classifiers (good, medium, bad) the table consists of nine different combinations of predicted and actual values. The accuracy for detecting Target 1 (good) is at 99,6%, shown in the grey box in the first column. The overall accuracy of all targets can be concluded at 93,3% shown in the blue box at the corner.

Concluding Remarks

This paper presents a machine learning approach for the analysis of oil data. The focus of the paper is to determine whether a reconstruction and a classification inferring Multi-Layer-Perceptrons is viable. The proposed supervised approach requires labeled data. The labeled data was created using the Dynamic Time Warping algorithm. Two different neural networks were trained for classification and reconstruction. The subsequent performance evaluation has shown that an accurate reconstruction is already possible with a neural network consisting of three layers with 200 neurons each. A further increase in the number of neurons or layers does not result in a significant improvement.The classification in separate oil classes based on a confusion matrix ended up showing an overall accuracy of 93,3 % with low percentages in false positives and negatives.Future research will be devoted to the application of the proposed method on further liquids (e.g. wine, beer or tap water) in order to assess its generalization capabilities.

Acknowledgments

The results of this paper have been produced in cooperation with the project OSA4Consens, funded by the Federal Ministry of Education and Research (FKZ:02P16K112).

References

K. Pöpping, “Das Betriebs- und Alterungsverhalten biologisch schnell abbaubarer Hydrauliköle" Dissertation, April, 2012.

G. E. Newell, “Oil analysis cost-effective machine condition monitoring technique,” *Ind. Lubr. Tribol.*, vol. 51, no. 3, pp. 119–124, Jun. 1999.

A. K. S. Jardine, D. Lin, and D. Banjevic, “A review on machinery diagnostics and prognostics implementing condition-based maintenance,” *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006.

A. D. Stuart, S. M. Trotman, K. J. Doolan, and P. M. Fredericks, “Spectroscopic measurement of used lubricating oil quality,” *Appl. Spectrosc.*, vol. 43, no. 1, pp. 55–60, January 1989.

S. Paul, W. Legner, A. Hackner, V. Baumbach, and G. Müller, “Multi-parameter monitoring System für hydraulische Flüssigkeiten in Offshore-Windkraftgetrieben,” *Tech. Mess.*, vol. 78, no. 5, pp. 260–267, 2011.

A. Agoston, C. Oetsch, J. Zhuravleva, and B. Jakoby, “An IR-absorption sensor system for the determination of engine oil deterioration,” in *Proceedings of IEEE Sensors, 2004.*, pp. 463–466.

“CRISP-DM: Ein Standard-Prozess-Modell für Data Mining – Statistik Dresden.” [Online]. Available: https://statistik-dresden.de/archives/1128. [Accessed: 06-Apr-2019].

R. Wirth and J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining.”

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, “Chapter 14. Statistical Description of Data. 14.9 Savitzky-Golay Smoothing Filters,” pp. 766–772, 2007.

A. L. Samuel,“Some studies in machine learning using the game of Checkers,” *IBM J. Res. Dev.*, pp. 71--105, 1959.

T. Mitchell, *Machine Learning*. .McGraw-Hill Education Ltd., 1997

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, pp. 37–37, March 1996.

S. S. Haykin and S. S. Haykin, *Neural networks and learning machines*. Prentice Hall/Pearson, 2009.

P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for Activation Functions,” October 2017.

Author Information

Marco Stang
Karlsruhe Institute of Technology (KIT)
Kaiserstraße 12, 76131 Karlsruhe
Contact E-mail: *marco.stang@kit.edu*

Martin Bohme
Karlsruhe Institute of Technology (KIT)
Kaiserstraße 12, 76131 Karlsruhe
Contact E-mail: *martin.boehme@kit.edu*

Eric Sax
Karlsruhe Institute of Technology (KIT)
Kaiserstraße 12, 76131 Karlsruhe
Contact E-mail: *eric.sax@kit.edu*