

## Kurzfassung

Die steigende Verfügbarkeit von Daten in Produktionsprozessen bietet neue Möglichkeiten in Bezug auf die Bewertung der Qualität von produzierten Bauteilen. Somit können bereits während des Produktionsprozesses datengetriebene Entscheidungen bezüglich der Qualität von Bauteilen getroffen werden, was gegenüber einer späteren Qualitätsprüfung hohe Kosten einsparen kann.

In dieser Arbeit werden Optimierungsmöglichkeiten für die binäre Klassifikation von fehlerhaften Ventilen im Rahmen der Qualitätsprüfung der hydraulischen Schlussprüfung von Wegeventilen untersucht, was dem letzten und kosten-intensivsten Produktionsschritt entspricht. Ein wichtiges Qualitätsmerkmal ist die gemessene Menge eines internen Öl-Volumenstroms im Bauteil. Durch Sensordaten aus den vorherigen Prüfschritten der Produktion soll die Qualität eines Bauteils klassifiziert werden, ohne dass eine Messung des Öl-Volumenstroms notwendig ist.

Diese Produktionsdaten weisen eine niedrige Qualität durch sich ändernde Prüfbedingungen, eine unbalancierte Zielvariable und eine geringe Anzahl vorhandener Messungen auf. Zusätzlich fehlen Informationen zur einer internen Spalthöhe im Ventil, die nicht durch Messungen erfasst werden kann. Deswegen existiert ein weiterer experimentell erstellter Versuchsdatensatz, der sowohl Informationen zu den Sensordaten, als auch zur Spalthöhe enthält.

Durch das Vorliegen dieser beiden Datensätze sind zwei Lösungsansätze zur Verbesserung der Datengüte zu untersuchen: Zum einen wird ein Meta-Regressionsmodell der Versuchsdaten mit der Spalthöhe als Zielvariable aufgestellt, um die Spalthöhe für den Produktionsdatensatz schätzen zu können. Dabei wird das unbalancierte Verhältnis von guten und fehlerhaften Bauteilen in den Versuchsdaten durch Vergleich und Anwendung von drei verschiedenen Oversampling-Methoden und drei Methoden zur synthetischen Datengenerierung ausbalanciert. Bei den Oversampling-Methoden handelt es sich um den Synthetic Minority Over-sampling Technique (SMOTE) Algorithmus sowie zwei Erweiterungen von diesem. Von den synthetischen Datengeneratoren basieren zwei auf Generative Adversarial Networks und eines auf Variational Autoencoders.

Zum anderen wird die Klassifikation der Produktionsdaten ohne die Verwendung der Versuchsdaten verbessert. Dies erfolgt durch die Wahl geeigneter Kreuzvalidierungssplits für die zeitabhängigen Produktionsdaten sowie durch die Anwendung von Methoden zur Feature Selection und das erneute Ausbalancieren des noch stärker unbalancierten Produktionsdatensatzes mit den zuvor genannten Methoden. Zur Klassifikation werden der tree-based Algorithmus Extreme Gradient Boosting Trees (XGBoost) und die auf neuronalen Netzen basierenden Deep Abstract Neural Networks (DANet) verglichen.

Es stellt sich heraus, dass die Anwendung eines Meta-Regressionsmodells nicht geeignet ist, um eine Verbesserung der Klassifikation des Volumenstroms in den Produktionsdaten zu erreichen. Für die synthetische Datengenerierung erweist sich jedoch das Verfahren, welches auf Variational Autoencoders basiert, gegenüber den Generative Adversarial Networks basierten Verfahren, als überlegen.

Die genannten Vorverarbeitungsschritte können die Klassifikation des Volumenstroms zwar verbessern, die erzielte Genauigkeit ist jedoch nicht ausreichend, um im Produktionsprozess angewandt zu werden. Als Klassifikationsalgorithmus ist XGBoost dem DANet für die gegebene Datenbasis deutlich überlegen.

## Abstract

The increasing availability of data in production processes offers new possibilities for the evaluation of the quality of produced components. Thus, data-driven decisions regarding the quality of these components can already be made during the production process, which can save high costs compared to a subsequent quality inspection.

In this work, optimization potentials for the binary classification of defective valves are investigated within the quality inspection of the hydraulic end-of-line testing of directional control valves, which corresponds to the last and most cost-intensive production step. An important quality feature is the measured quantity of an internal oil volume flow within the valve. Sensor data from the previous testing steps of the production should be used to classify the quality of a valve without the need to measure the oil volume flow. This production data has low quality due to varying test conditions, an unbalanced target variable, and a small number of existing measurements. In addition, there is a lack of information on an internal gap height in the valve, which cannot be captured by measurements. Therefore, another experimentally created data set exists, which contains information on sensor data as well as on the gap height.

Given these two data sets, two approaches to improve the data quality have to be investigated: First, a meta-regression model of the experimental data is set up to estimate the gap height for the production data set. This involves balancing the unbalanced ratio of good and defective valves in the experimental data set by comparing and applying three different oversampling methods and three synthetic data generation methods. The oversampling methods are the Synthetic Minority Over-sampling Technique (SMOTE) algorithm and two extensions of it. Two of the synthetic data generation methods are based on Generative Adversarial Networks, while the third one is based on Variational Autoencoders.

Second, the classification of the production data is improved without using the experimental data set. This is done by choosing appropriate cross-validation splits for the time-dependent production data, as well as by applying feature selection methods and rebalancing the even more unbalanced production data set with the previously mentioned methods. In regard to the classification methods, a comparison between the tree-based algorithm Extreme Gradient Boosting Trees (XGBoost) and the neural network-based Deep Abstract Neural Networks (DANet) takes place.

It turns out that the application of a meta-regression model is not suitable to achieve an improvement of the classification of the volume flow in the production data. In regard to the synthetic data generation, the method based on Variational Autoencoders proves to be superior to the Generative Adversarial Network based methods. The previously mentioned preprocessing steps can improve the classification of the volume flow, even though the accuracy achieved is not sufficient to be applied in the production process. As a classification algorithm, XGBoost is clearly superior to DANet for the given database.