Monitoring data quality for AI models in industrial glass production

Tom Röger, Fabian Steinle, Johannes Schilp

Angaben zur Veröffentlichung / Publication details:

Röger, Tom, Fabian Steinle, and Johannes Schilp. 2024. "Monitoring data quality for AI models in industrial glass production." *Procedia CIRP* 130: 282–87.
https://doi.org/10.1016/j.procir.2024.10.088.

Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 130 (2024) 282–287

www.elsevier.com/locate/procedia

57th CIRP Conference on Manufacturing Systems 2024 (CMS 2024)

Monitoring Data Quality for AI Models in Industrial Glass Production

Tom Röger[a,*], Fabian Steinle[b], Johannes Schilp[a]

[a]*University of Augsburg, Universitätsstr. 2, 86159 Augsburg, Germany*
[b]*University of Applied Science Aalen, Beethovenstraße 1, 73430 Aalen*

* Corresponding author. Tel.: +4917647334196; *E-mail address:* tom.roeger@web.de

Abstract

As artificial intelligence (AI) becomes more and more important in industrial settings, the quality and consistency of the data fed into these AI systems is becoming crucial. The success of AI models heavily relies on good data quality. That's why this thesis introduces a new system for monitoring the data going into AI models as it happens, this system is tested with historical data to make sure it works well. In industries where machines and production lines need to run efficiently and reliably, having high-quality data is a big challenge. Sometimes, the data quality changes or isn't good enough, which can make the AI models give wrong results. This can make people lose trust in how well AI models work. Our paper tackles this issue by using a mix of methods - statistical, clusters and classes - to check the data in real time. We apply this to the data from an AI model designed to predict when cutting tools for glass manufacturing need to be replaced. We rate each checking method and come up with an overall score. This way, we can accurately and efficiently evaluate both the data we already know about and new, unseen data. This overall score even helps someone who's not an AI expert to quickly figure out if the AI model they're using can be trusted right now, and it also points out when something might be off with the data. Looking ahead, we plan to fine-tune how we balance the importance of each method based on different situations. This will help make our monitoring system work well for all kinds of data going into AI models, not just in glass manufacturing.

Keywords: Artificial Intelligence (AI), Data Quality Monitoring, Anomaly Detection Methods, Industrial Production

1. Introduction

1.1. Background and Motivation

Artificial Intelligence (AI) is increasingly recognized as a pivotal technology and is experiencing widespread adoption, particularly within the industrial sector [1,2]. AI models are deployed for various purposes including customer interaction, process automation, and data analytics [3, 4]. The effectiveness of these models critically depends on the quality of the underlying data [5, 6]. When faced with inferior input data, these models often generate inaccurate or erroneous predictions [7, 8]. Frequently, such flawed data go undetected without technical intervention, adversely affecting the AI's performance [9] and reducing the trust of users [10]. Considering these challenges, this paper proposes an initial concept for the sensor-based monitoring of input data, aiming to improve the reliability of AI models by ensuring higher data quality. Production data from float glass production is used as an example.

1.2. Objective and Research Questions

To support the development of anomaly detection methods, precise monitoring of input data is essential. For this purpose, historical data from a period of 16 months, which include measurements of various parameters, are used. These serve as input for an AI model aimed at predicting the remaining

2212-8271 © 2024 The Authors. Published by Elsevier B.V.

10.1016/j.procir.2024.10.088

lifespan of cutting wheels in glass cutting and breaking technology.

The analysis focuses on parameters such as glass thickness and conveyor speed. Since glass thickness varies from 3 mm to 12 mm and conveyor speeds differ accordingly, the data concerning glass thickness are examined in detail. The goal is to reduce variances and detect anomalies more effectively by considering specific subranges of glass thickness. After dividing the data into training and test datasets, various analysis methods are applied, based on statistical, distance-based, cluster-based, and class-based approaches.

Each data point is individually evaluated and assigned a score reflecting data quality. An average value from the different methods is calculated to combine their strengths. As a result, each data point receives a rating between 0% and 100%, with lower values indicating poor data quality, while higher ratings indicate good input data.

This refined presentation makes the data analysis process and the choice of methods more transparent and understandable without overwhelming the reader with too many technical details.

The primary goal of this study is to develop a software-based monitoring system for the input data of an AI model used in glass manufacturing plants. This system will automatically check large volumes of data quickly to support operational processes. It provides operators with an overview of current process parameters during operation and detects anomalies or abnormal data in real time. Additionally, an indicator of the current data quality is calculated and displayed directly on the plant's interface. This feature also makes it easier for inexperienced workers to recognize inaccurate or erroneous predictions of the AI model caused by poor input data. Such insights enable the quick identification of issues, such as faulty sensors, and the prompt implementation of corrective measures. The central research question of this work is: "How can input data from AI models in production be effectively monitored for anomalies?"

2. Fundamentals and State of the Art

In production applications, systems are usually sensor-based [11]. The literature primarily recommends anomaly detection methods that use statistical, distance-based, cluster-based, class-based, or deep learning approaches [12]. Anomalies can be classified into several categories: point anomalies are individual data points that significantly deviate from the rest of the data mass; contextual anomalies are data points that would normally be considered normal but are deemed abnormal in a specific context; collective anomalies refer to groups of data that exhibit abnormal behavior compared to the entire dataset.

Statistical methods assume that normal data are in densely populated regions, whereas abnormal data are often isolated. Gaussian methods are advised for these scenarios, assuming a normal distribution (parameter-based), as well as histograms and kernel-based methods that do not require known data relationships (non-parametric) [13, 14]. Distance-based methods often use the position or data density in the immediate vicinity of a point as a criterion for assessment, utilizing techniques such as the Nearest Neighbor method or various density procedures.

Cluster-based methods group data points into clusters that display similar patterns. Abnormal data either do not fit well into these clusters or are found in clusters with few values [12, 15]. Class-based methods learn patterns within the dataset, identify deviating data points, and classify them as anomalies. Recommended techniques include Bayesian Networks, Neural Networks, Support Vector Machines (SVM), and rule-based methods [12, 15].

Deep learning approaches transform data into other dimensionalities, through which anomalies are identified by different data mappings. This can be achieved using Principal Component Analysis (PCA), Fuzzy logic, or Neural Networks [14, 17].

The literature has described specific cases where input data were monitored for anomalies. Königs and Brecher [18] used a Gaussian method from the statistical approaches to detect anomalies occurring during the prediction of remaining tool life in milling processes. A distance-based method that calculates the angle between points yielded the best results in simulated datasets augmented with anomalies [19]. For anomaly detection in real production data during steel machining on a lathe, a statistical Gaussian method proved most effective [21]. Liu et al. [20] also used a Gaussian method but incorporated a prior classification based on the signal length of the monitored parameter. Alimohammadi and Nancy Chen [22] found that the Nearest Neighbor method and a class-based approach were most effective for production data from the oil and gas industry.

It is evident that there is no universally optimal method for detecting anomalies in sensor-supported systems. The choice of method and specific technique depends heavily on the application area in production and the parameters being monitored. Often, combinations of multiple methods are employed.

3. Methodology

3.1. Integration and Detailing of Data

As a foundation for developing anomaly detection methods, a structured dataset is essential. In this process, parameters of an AI model are correlated, which predicts the remaining lifespan of a cutting tool and thus the timing for replacing the cutting wheel. Data integration allows parameters read from the database to be compiled into a dataset. In this study, the glass thickness, which is measured in millimeters, and the conveyor speed of the production facility, given in millimeters per minute, are used. These are combined in the dataset titled "All Glass Thicknesses." Figure 1 displays the glass thicknesses long with their corresponding conveyor speeds over a period of six months.

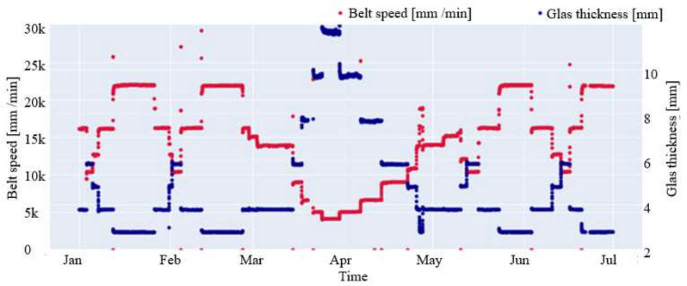

Fig. 1. Dataset „All Glass Thicknesses“

Correlations in the data show that conveyor speed decreases as glass thickness increases, with speeds ranging from 5,000 mm/min at a nominal glass thickness of 12 mm to 22,000 mm/min at 3 mm. Notably, during transitions between nominal glass thicknesses, sensor data for both parameters may be sparse or absent, as glass manufacturers often turn off measuring devices during these adjustments (indicated by a green marker in Fig. 1). This variability makes anomaly detection in the 'All Thicknesses' dataset unfeasible.

To manage data variance, creating subgroups based on glass thickness is effective. For instance, by focusing on a specific nominal glass thickness, all irrelevant data entries are excluded, forming a streamlined dataset named 'Nominal Glass Thickness 4 mm', covering thicknesses from 3.50 mm to 4.49 mm with corresponding conveyor speeds.

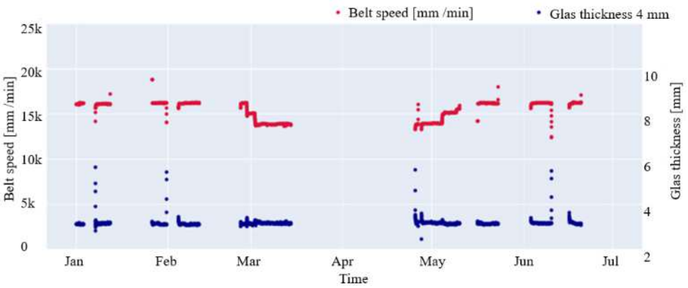

Fig. 2. Dataset „Nominal Glass Thickness 4 mm”

In the 'Nominal Glass Thickness 4 mm' dataset, measurement variance is reduced. Data patterns at the series' start and end are highlighted with green in Fig. 2, indicating thickness adjustments that affect the variance. To pinpoint anomalies unrelated to glass thickness changes and to minimize measurement scatter, further detailing of the data is useful. Fig. 3 illustrates the 'Glass Thickness 3.8 mm' dataset, displaying conveyor speed and glass thickness over six months, with thicknesses ranging from 3.749 mm to 3.850 mm.

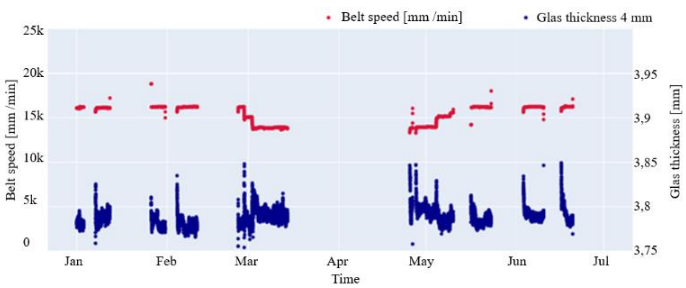

Fig. 3. Dataset „Nominal Glass Thickness 3,8 mm”

In the "Glass Thickness 3.8 mm" dataset, there are only a few data points remaining that indicate the beginning or end of a thickness adjustment. Further detailing would offer minimal benefit, as the dataset still contains 76,258 data points, accounting for 99.5% of the "Nominal Glass Thickness 4 mm" dataset. By detailing the "Glass Thickness 3.8 mm" dataset to one decimal place, the variances within the dataset are minimal, making it a suitable basis for developing input data monitoring. This approach can be applied across all thickness ranges.

3.2. Development of Anomaly Detection Methods

At the beginning of the development process, the data are divided into training and test datasets. The training dataset "Training 3.8 mm" is used to train methods or to determine reference values (e.g., median, standard deviation), covering data from one year. The test dataset "Test 3.8 mm" is used to test the methods on unknown data, simulating a production environment, and includes data from four months. The "Training 3.8 mm" dataset is utilized for development, where methods are applied to identify anomalies and abnormal data.

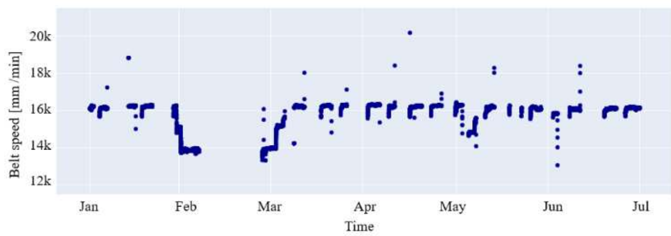

Fig. 4. Dataset „Training 3,8 mm”

In the subsequent development process of the methods, binary labeling is used in the plots. Additionally, an Outlier Score is calculated, which will be utilized in Chapter 4. From the statistical methods, a Gaussian approach is employed. This method determines the median and standard deviation from the training data. The range between the median value plus or minus the standard deviation (possibly multiplied by a factor) is recognized as normal. Data points outside this range are classified as abnormal. Therefore, a fixed range around the median value is established within which the normal data lie. This approach is designed to detect both point anomalies and collective anomalies that are far from the median value. In the "Training 3.8 mm" dataset used, the median value is 16,137 mm/min and the standard deviation is 773 mm/min. Thus, data points outside the range of 15,363 to 16,910 mm/min are identified as abnormal.

The Z-Score method also employs the median and standard deviation. A measurement value X has the median $\overline{X}$ subtracted and is then divided by the standard deviation S. The absolute result is referred to as the Z-Score Z. For normal data, this score ranges between zero and one. Values greater than one are classified as abnormal. The Z-Score thus represents the percentage deviation of the measurement from the median value, relative to the standard deviation. This method aims to detect both point anomalies and collective anomalies that are far from the median value.

For these two parameter-based methods using statistical approaches, the comparable results are shown in Figure 5.

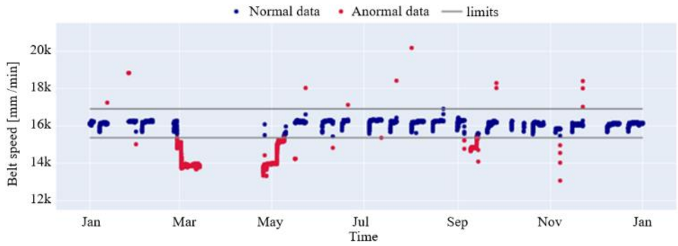

Fig. 5. Gaussian Method

These methods identify approximately 19% of the dataset as abnormal. Data points within a fixed boundary around the median are considered normal, while those outsides are flagged as abnormal, effectively detecting point anomalies. The period between March and May is identified as a collective anomaly. However, abnormal data are always detected using the median and standard deviation, even when variance around the median is low, such as 500 mm/min.

For the next two non-parametric statistical methods, changes in data points relative to their predecessors are constrained. The difference between consecutive data points is calculated, and if this exceeds a threshold set at the 95th percentile, it is deemed abnormal. These methods aim to identify significant changes from the dataset norm. One method limits absolute change, while the other restricts relative change by normalizing the difference by the preceding value, giving more weight to shifts at lower measurement values. Both methods produce similar results.

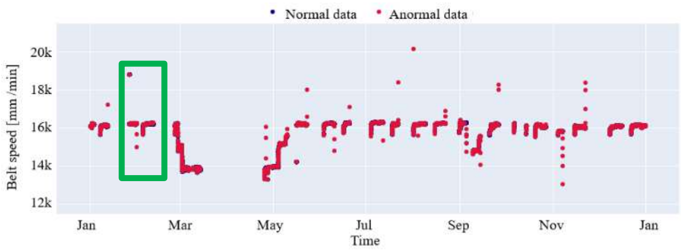

Fig. 6. Limitation of the Absolute Change (a)

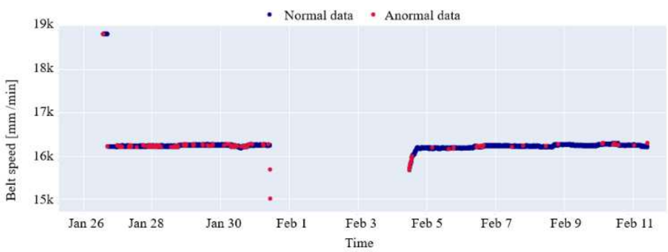

Fig. 7. Limitation of the Absolute Change (a)

Due to the high data volume and low resolution, Fig. 6 may mislead about the quantity of abnormal data. On average, only 3.5% of the dataset is classified as abnormal using these methods. This is clearer when focusing on a smaller timeframe, as shown in Fig. 7, which zooms into a 15-day period within the green-marked area of Fig. 6, clearly displaying the method's results.

By restricting changes between consecutive data points, significant deviations are flagged as abnormal, aiding in the detection of point anomalies. If a collective anomaly is distant from other data, only the first affected data point is marked as abnormal.

Distance-based methods, specifically the Nearest Neighbor and Local Outlier Factor, utilize Euclidean distances for anomaly detection. These methods calculate distances between a data point and all others in the dataset. In the Nearest Neighbor method, these distances are summed and compared to a reference value (95th percentile) to classify data points as abnormal or normal. The Local Outlier Factor averages these distances, with the same classification criteria. Both methods, adjustable by the variable k for the number of nearest neighbors, aim to identify anomalies based on deviations from neighboring data points, yielding comparable results.

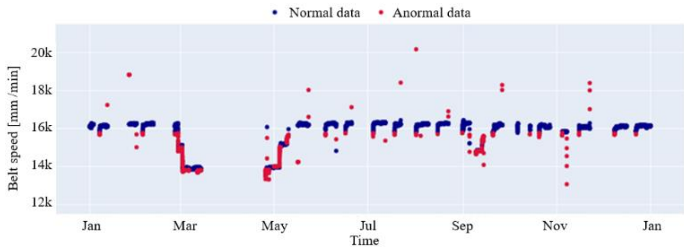

Fig. 8. Nearest Neighbor Method

Using these methods, around 5% of the dataset is identified as abnormal. Data points in areas with sparse measurements, particularly those affected by point anomalies or thickness adjustments (like those in early November), are classified as abnormal. For the 14,000 mm/min range between March and May, the classification is mixed, with data assessed as both normal and abnormal.

A density-based method, which combines cluster-based and distance-based approaches, is employed from the cluster-based strategies. This method divides the range of data variation into 100 clusters, each representing approximately 250 mm/min of conveyor speed. It assesses what percentage of the dataset falls within each cluster. Data points in a cluster with density below a reference value are deemed abnormal, while those with higher density are considered normal. This approach aims to identify ranges with lower data density compared to other areas as abnormal.

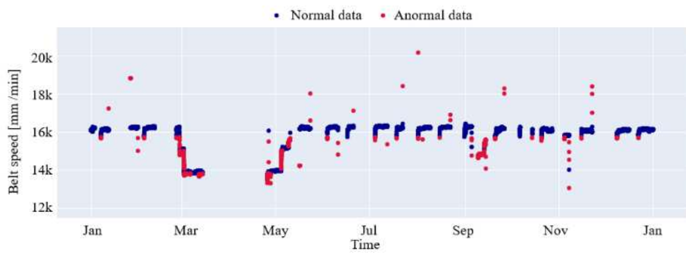

Fig. 9. density-based method

With this method, 8.2% of the dataset is identified as abnormal. The density-based method categorizes areas with low data density as abnormal, thus detecting point anomalies. Additionally, data points that vary due to thickness adjustments are classified as abnormal, as observed in early May and early November. By segmenting into subsections, measurements with minor deviations can be classified differently, and possibly recognized as abnormal, such as in early January through February.

From the class-based methods, the Isolation Forest, also known as a decision tree, is used. It is trained using the training data to recognize patterns in the dataset and autonomously sets various criteria with threshold values. These criteria are sequentially checked for each data point, and if any are not met, the point is classified under that criterion. The fewer criteria a data point meets, the higher the likelihood it is considered b normal. Normal data are labeled with 1, abnormal data with -1. The algorithm requires a predetermined percentage of the dataset to be recognized as abnormal [23]. In Fig. 10, 5% was selected.

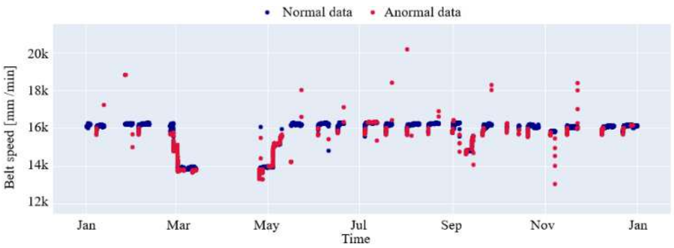

Fig. 10. Isolation Forest

With this method, as predetermined, 5.0% of the dataset is classified as abnormal. The Isolation Forest is particularly effective at detecting point anomalies. It also recognizes values that vary due to thickness adjustments. However, with this method, it must be specified how much of the dataset should be categorized as abnormal. This percentage will always be identified by the method, regardless of the dataset's quality.

4. Evaluation and Results

Since the developed methods each have different strengths and weaknesses, they are combined using the Outlier Scores from each method, which indicate how well a data point fits within a given method. A score close to 0% suggests an anomaly, while a score closer to 100% indicates a lower probability of anomaly. For example, in the Gaussian method, a data point that matches the median value receives an Outlier Score of 100%. The further the value deviates from the median, the lower its Outlier Score. This provides a meaningful evaluation for each data point on a scale from 0% to 100%. Subsequently, an average score, called the Score, is calculated from the individual Outlier Scores. This averaging combines the strengths of the various methods. There is also the option to weight certain methods more heavily. In Fig. 11, the Score corresponds to the color of the data points.

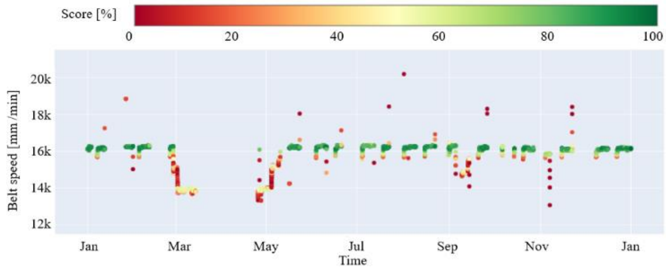

Fig. 11. Dataset „Training 3,8 mm”

Data points around 16,100 mm/min consistently receive scores near 100%, reflecting the highest density in that region. As data points approach or diverge from this value, color gradients appear, particularly noticeable in November and December. From March to May, many data points around 14,000 mm/min are scored around 50%, indicating deviation from the norm but also concentration in this area. Anomalies and variations due to thickness adjustments are scored close to 0%. These results demonstrate that input data monitoring effectively identifies and rates deviations, yielding promising results for the 'Training 3.8 mm' dataset.

To evaluate these methods on unfamiliar and smaller datasets, input data monitoring is applied to the 'Test 3.8 mm' dataset, covering January to April 2023. This simulates the method's application in production. In Fig. 12, results for the 'Test 3.8 mm' dataset are displayed, showing conveyor speed over four months with scores visualized by the color of the data points.

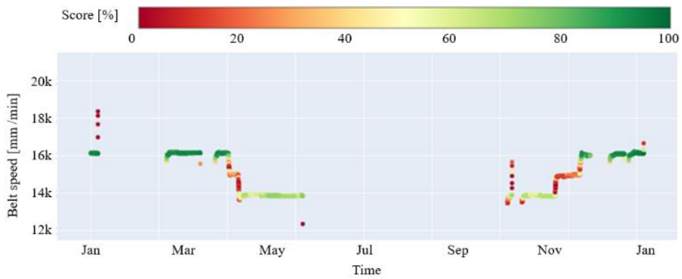

Fig. 12. Dataset „Test 3,8 mm“

In the 'Test 3.8 mm' dataset, values around 16,100 mm/min also receive high scores, while areas like around 14,000 mm/min, with many data points, are rated slightly lower. Outliers and variations due to thickness adjustments are scored near 0%. Notably, around January 29th and April 9th, scores fluctuate around 20%.

These findings affirm that input data monitoring effectively identifies and evaluates deviations in a production environment, delivering promising results even with unfamiliar datasets.

5. Conclusion and Future Work

In this study, each data point of a dataset is evaluated by eight different methods using statistical, distance-based, cluster-based, or class-based approaches, with scores ranging from 0% to 100%. An average score, termed the Score, is then calculated from these individual evaluations, describing the data quality of the data point on a scale from 0% to 100%. By averaging these scores, the strengths of the methods are combined. Displaying the Score on the equipment makes it clear, even to inexperienced workers, what the input data quality of the AI model is. Lower scores indicate poor data quality, while higher scores suggest good input data.

The input data monitoring was tested on data from an AI model that predicts the remaining life of a cutting tool, and the results were visualized in color plots. These plots showed that areas with the highest data density received very high scores, typically around the median value. Areas with fewer data points received lower scores. Point anomalies or outliers were detected and assigned very low scores. Thus, the input data monitoring identifies and evaluates deviations from the rest of the dataset. Significant results were achieved for the conveyor speed. The results of the input data monitoring were validated by experts and rated as a functional solution for anomaly detection in the input data of an AI model. This thesis demonstrates that the developed concept can monitor input data of AI models for anomalies or abnormal data in production settings.

The concept developed in this Paper could be adapted to other parameters of the glass manufacturing process with some modifications. Additionally, the weightings of the methods for anomaly detection could be adjusted to optimize the results, potentially through an AI model. Furthermore, more methods with different approaches could be developed and implemented, or existing methods could be replaced. To use the concept in production in the future, it must be implemented, tested, and validated on the equipment. In the long term, the tasks of input data monitoring are intended to be undertaken by an AI model, which will also handle the management of anomalies or abnormal data.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT4 to improve readability and grammar. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] V. Brühl, "Artificial Intelligence, Machine Learning and Big Data: Basics, Market Potentials and Economic Policy Relevance," WiSt - Business Studies, vol. 48, no. 11, pp. 34–41, 2019.

[2] Data Scientist. "Machine Learning: Definition, Functioning, Applications." Accessed on: April 19, 2024. [Online.] Available: https://datascientest.com/de/machine-learning-definition-funktionsweise-anwendungen

[3] C. Rammer, I. Bertschek, B. Schuck, V. Demary, and H. Goecke, "Use of Artificial Intelligence in the German Economy: Status of AI Usage in 2019,". Accessed on: April 19, 2024. [Online]. Available at: http://hdl.handle.net/10419/222374

[4] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. (Prentice-Hall Series in Artificial Intelligence). Upper Saddle River, NJ: Prentice-Hall, 2010. Accessed on: April 19, 2024.

[5] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine Learning in Manufacturing: Advantages, Challenges, and Applications," Production & Manufacturing Research, vol. 4, no. 1, pp. 23–45, 2016.

[6] G. Reggio and E. Astesiano, "Big Data/Analytics Projects Failure: A Literature Review," in the 2020 46th (SEAA), Portoroz, Slovenia, 2020, pp. 246–255, doi: 10.1109/SEAA51224.2020.00050.

[7] V. Wittpahl, Resilience: Life - Spaces - Technology. Springer Nature, 2023. Accessed on: April 19, 2024. [Online]. Available at: https://doi.org/10.1007/978-3-662-66057-7

[8] K. Arntz et al., "Artificial Intelligence in Single and Small Batch Production." Accessed on: April 2, 2024. [Online.] Available: https://doi.org/10.24406/ipt-n-640931

[9] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network Outlier Detection in Wireless Sensor Networks," Knowledge Information Systems, vol. 34, no. 1, pp. 23–54, 2013, doi: 10.1007/s10115-011-0474-5.

[10] A. Moring, AI in the Job. Berlin, Heidelberg: Springer Berlin Heidelberg, 2021. Accessed on: April 19, 2024.

[11] A. Mockenhaupt, Digitization and Artificial Intelligence in Production. Wiesbaden: Springer Fachmedien Wiesbaden, 2021. Accessed on: April 19, 2024.

[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[13] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," IEEE Communications Surveys & Tutorials, vol. 12, no. 2, pp. 159–170, 2010, doi: 10.1109/SURV.2010.021510.00088.

[14] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier Detection Approaches for Wireless Sensor Networks: A Survey," Computer Networks, vol. 129, pp. 319–333, 2017, doi: 10.1016/j.comnet.2017.10.007.

[15] B. Chander and G. Kumaravelan, "Outlier Detection Strategies for WSNs: A Survey," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5684–5707, 2022, doi: 10.1016/j.jksuci.2021.02.012.

[16] A. Patcha and J.-M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends," Computer Networks, vol. 51, no. 12, pp. 3448–3470, 2007, doi: 10.1016/j.comnet.2007.02.001.

[17] T. Ott, "Anomaly Detection with Machine Learning," BI-Spectrum, no. 2, pp. 26–30, 2018. [Online]. Available at: https://www.consist.de

[18] M. Königs and C. Brecher, "Process-parallel Virtual Quality Evaluation for Metal Cutting in Series Production," Procedia Manufacturing, vol. 26, pp. 1087–1093, 2018, doi: 10.1016/j.promfg.2018.07.145.

[19] H. Shekhar Jha, H. Mustafa Ud Din Sheikh, and W. J. Lee, "Outlier Detection Techniques Help Us Identify and Remove Outliers in Production Data to Improve Production Forecasting," Online, 2021, doi: 10.15530/ap-urtec-2021-208384.

[20] Y. Liu, J. Zhang, X. Hu, and S. Sun, "Sensor Data Anomaly Detection and Correction for Improving the Life Prediction of Cutting Tools in the Slot Milling Process," International Journal of Advanced Manufacturing Technology, vol. 119, nos. 1-2, pp. 463–475, 2022

[21] L. C. Brito, M. B. Da Silva, and M. A. V. Duarte, "Identification of Cutting Tool Wear Condition in Turning Using Self-Organizing Map Trained with Imbalanced Data," Journal of Intelligent Manufacturing, vol. 32, no. 1, pp. 127–140, 2021, doi: 10.1007/s10845-020-01564-3.

[22] H. Alimohammadi and S. Nancy Chen, "Performance Evaluation of Outlier Detection Techniques in Production Timeseries: A Systematic Review and Meta-analysis," Expert Systems with Applications, vol. 191, p. 116371, 2022, doi: 10.1016/j.eswa.2021.116371.

[23] S. Mazzanti, "'Isolation Forest': The Anomaly Detection Algorithm Any Data Scientist Should Know," Accessed on: April 19, 2024.