TCM Dataset: a benchmark for anomaly detection in Industry 4.0

Jakub Jakubowski [1, 2]

October 21, 2023

[1]AGH University of Science and Technology

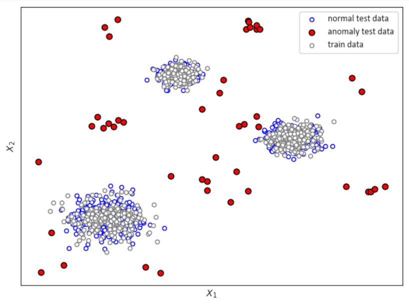
[2]ArcelorMittal Poland

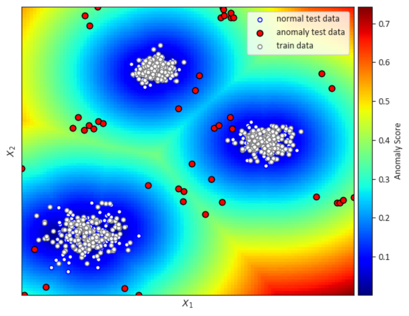
Table of Contents

Anomaly Detection

Anomaly Detection

- Observations which do not fit to a distribution of a given process.
- A small fraction of all observations
- Diverse and sparse





Anomaly Detection

Supervised methods	Unsupervised methods
• Requires labels for training • Needs balanced dataset • Learns patterns based on the provided labels • Better control over learning procedure	• No labels for training • Does not require dataset balancing • Learns general patterns in the data • Little control over learning procedure

Anomaly Detection

Unsupervised methods

- **Isolation Forest**: It identifies anomalies by isolating them into shorter paths in a binary tree structure.
- **One-Class SVM**: It learns a boundary that encompasses the majority of data points, classifying outliers as anomalies.
- **Autoencoders**: Neural networks are used to compress and reconstruct data, and anomalies are detected by measuring reconstruction error.
- **Local Outlier Factor**: It assesses the local density of data points, identifying those with significantly lower density.
- **HBOS**: It creates histograms for individual features and combines their outlier scores to identify anomalies.

Motivation

Motivation

Real-world datasets

- represents a real-life scenario
- they are often of low quality, which makes AD tideous task.
- the number of actual anomalies might be very limited.
- the data might miss labels or be poorly annotated (no ground-truth)
- data might be affected by concept drifts.
- require expert knowledge to understand the source of anomalies.

Motivation

Synthetic datasets

- access to the ground-truth
- might be adapted to the specific needs
- can be based on artificial dependencies between features
- lack real-world variability
- difficult to model real anomalies

Available benchmarks

Available benchmarks

Numenta Anomaly Benchmark (NAB)

- Artificial and real-world datasets
- Includes industrial examples
- Univariate time series

CMAPSS

- Run-to-failure simulations of aircraft engine
- Desgined for remaining useful life estimation
- 4 datasets with different level of difficulty
- Up to 2 failure modes and 6 operating conditions

Available benchmarks

Credit Card Fraud Detection

- 30 features and ¿200k observations
- Relatively low number of anomalies (0.17%)

KDD Cup 1999 Data

- Intrusion detection data
- 40 features and ¿4M observations

Available benchmarks

Yahoo Webscope Anomaly Detection Dataset

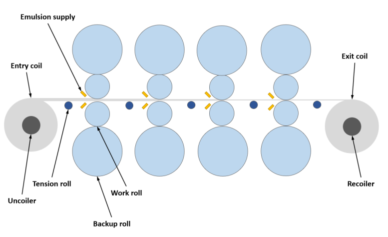
- 367 real and synthetic time series data
- Includes e.g. production traffic in computer networks

MetroPT dataset

- Real-world data from a metro train with digital and analog signals
- Developed for the anomaly detection and failure prediction

Tandem Cold Mill

Tandem Cold Mill



Schme of 4-high rolling mill with 4 stands [2]

Actual cold rolling mill (Kraków, Poland)

Tandem Cold Mill

Most important rolling parameters

- Reduction (draft) of the thickness
- Rolling force
- Rolling torque
- Rolling speed
- Forward and backward tensions
- Mechanical properties of the materials
- Power consumption

Tandem Cold Mill

Friction coefficient

- Friction coefficient affects the rolling force and torque.
- Difficult to precisely estimate due to complexity of the phenomemnon.
- Is not constant along the roll-strip arc of contact.

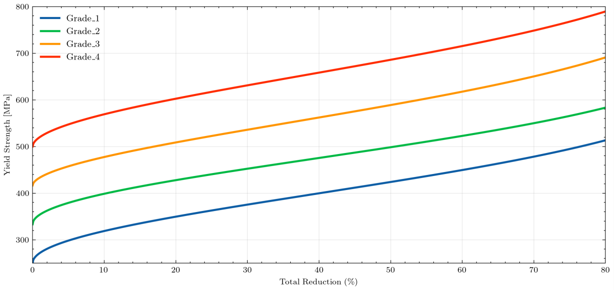
Mechanical properties of steel

- Depends mainly on the chemical composition of the material and its pretreatment.
- Determines the energy required to deform the material.

Tandem Cold Mill

$$\varepsilon = \log \frac{h_0}{h_1} \tag{1}$$

$$\sigma = \sigma_0 + K\varepsilon^b \tag{2}$$



Mechanical characteristics of the steelgrades

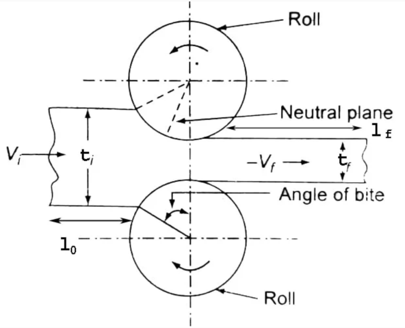
Bland-Ford model

$$F_r = 2R' \left[\int_0^{\phi_n} \frac{kh}{h_o}(1 - \frac{\sigma_o}{k_o})^{(\mu H)}\, d\phi + \int_{\phi_n}^{\phi_i} \frac{kh}{h_i}(1 - \frac{\sigma_i}{k_i})^{(\mu(H_i - H))}\, d\phi \right] \quad (3)$$

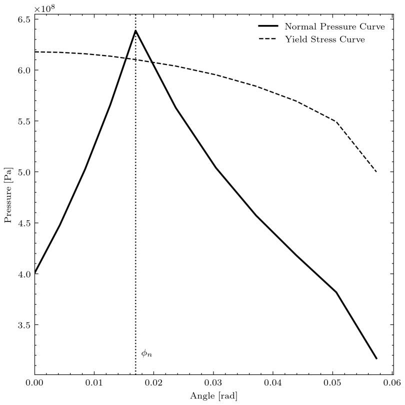
$$T_r = \mu R R' \left[\int_{\phi_n}^{\phi_i} \frac{kh}{h_i}(1 - \frac{\sigma_i}{k_i})^{(\mu(H_i - H))}\, d\phi - \int_0^{\phi_n} \frac{kh}{h_o}(1 - \frac{\sigma_o}{k_o})^{(\mu H)}\, d\phi \right] \quad (4)$$

where F_r - rolling force [N]; T_r - rolling torque [Nm] R'-deformed roll radius [m]; ϕ - contact angle [rad]; k - strip yield stress [Pa]; σ - strip tension [Pa]; h - strip thickness [m]; μ - friction coefficient [-] ; H - dimensionless thickness [-]; i - entry; o - exit, n - neutral point.

Tandem Cold Mill

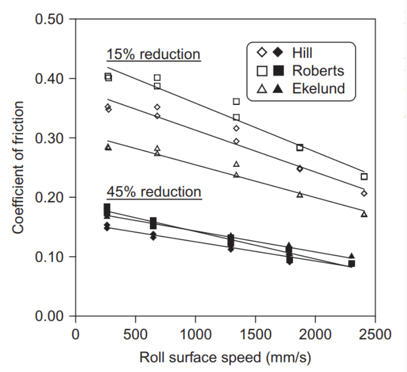


Roll bite [1]

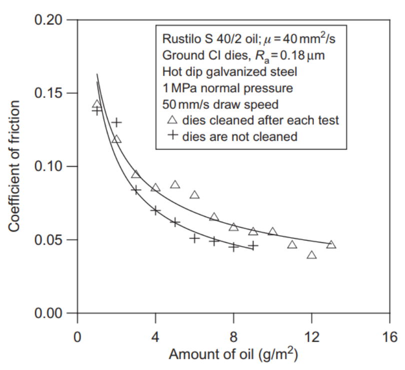


Pressure along the contact arc

Tandem Cold Mill

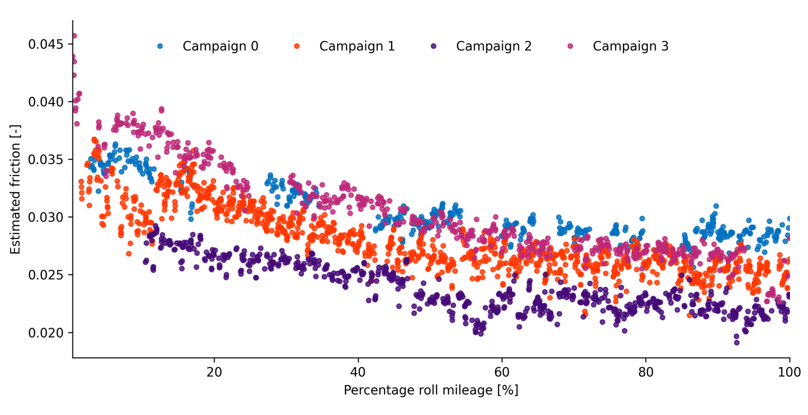


Influence of speed and reduction on friction [3]



Influence of lubricant on friction [3]

Tandem Cold Mill



Influence of roll mileage on friction [2]

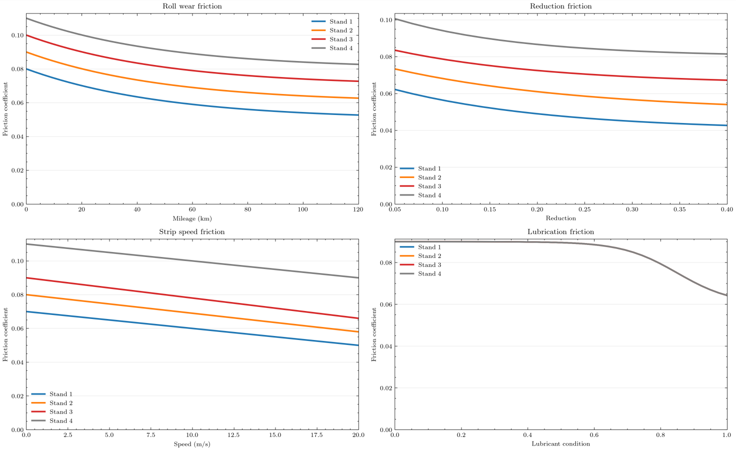
Dataset characteristics

Dataset characteristics

Main assumptions

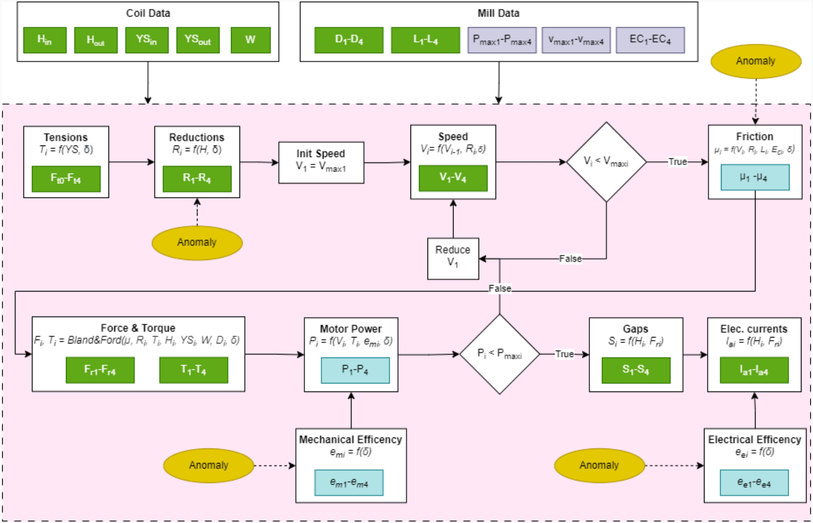
- We consider a four stand rolling mill, which reduces the thickness of steel strips.
- We have predefined family of products, which might be processed. These products differ by the mechanical properties, thickness, reduction and width.
- The rolling mill have defined characteristics like the motor power, speed limit, reduction range.
- The parameters of rolling process are dependent on the physics-based models and correlations adapted from scientific papers.
- After each rolled product, the parameters of the mill (work rolls characteristics, lubrication) are updated.

Dataset characteristics



Assumed relation between different features and friction coefficient

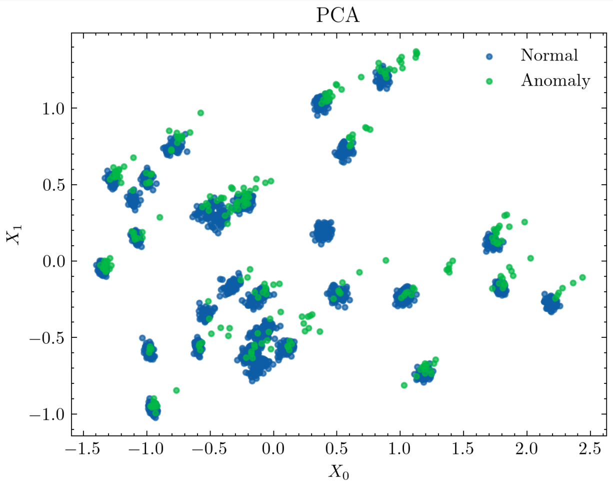
Dataset characteristics



Data generator pipeline

Preliminary results

Preliminary results



PCA visualization of the example dataset

Preliminary results

Exemplary results of Unsupervised Anomaly Detection

Model	Precision	Recall	F1	PR AUC
Autoencoder	**0.773**	**0.630**	**0.695**	**0.782**
Half-Space Trees	0.293	0.239	0.263	0.298
Isolation Forest	0.480	0.391	0.431	0.497
LODA	0.547	0.446	0.491	0.543
LOF	0.507	0.413	0.455	0.526
One-class SVM	0.533	0.435	0.479	0.565

Summary

Summary

- We present synthetic data generator based on cold-rolling process.
- Includes 4 different types of anomalies.
- Allows to generate concept drifts.
- Ground-truth for evaluation of ML and XAI methods.
- Possibility to customize the dataset to the specific needs.
- Research ideas: Anomaly Detection, Remaining Useful Life prediction, Concept Drift detection, Domain Adaptation, Explainable AI

References

[1] Anon. **Rolling -(A Brief Guide To Rolling And Rolling Mills) — industrialsafetyguide.com.** `https://industrialsafetyguide.com/rolling/`. [Accessed 21-10-2023].

[2] Jakub Jakubowski et al. **"Roll Wear Prediction in Strip Cold Rolling with Physics-Informed Autoencoder and Counterfactual Explanations".** In: IEEE, Oct. 2022, pp. 1–10. DOI: 10.1109/DSAA54385.2022.10032357.

[3] John G. Lenard. **"9 - Tribology".** In: *Primer on Flat Rolling (Second Edition)*. Second Edition. Oxford: Elsevier, 2014, pp. 193–266. DOI: `https://doi.org/10.1016/B978-0-08-099418-5.00009-3`.