MLA Data Challenge – Group 37

TECHNISCHE UNIVERSITÄT DARMSTADT

Anomaly Detection in Pneumatic Cylinder Production

Motivation and Goals

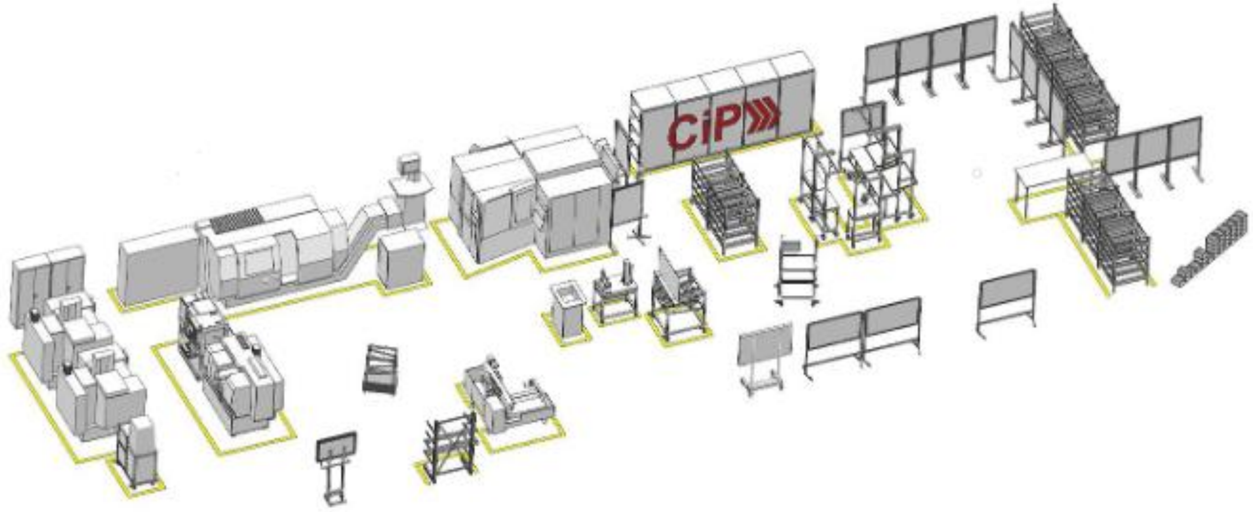
TECHNISCHE UNIVERSITÄT DARMSTADT

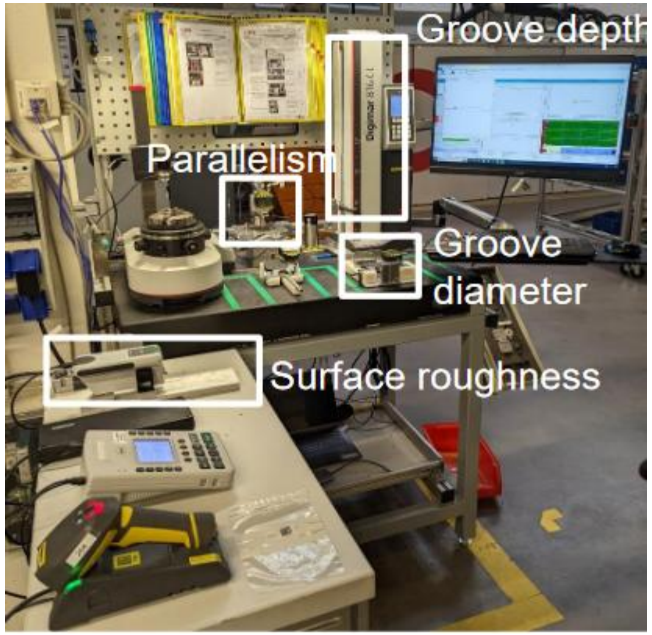
- Background is a manufacturing process of a pneumatic cylinder by a CNC-Milling machine
- Task is to develop a machine learning model, that uses internal and external machine data to classify the bottom parts into:

False: Anomaly, **True**: No Anomaly

Motivation and Goals

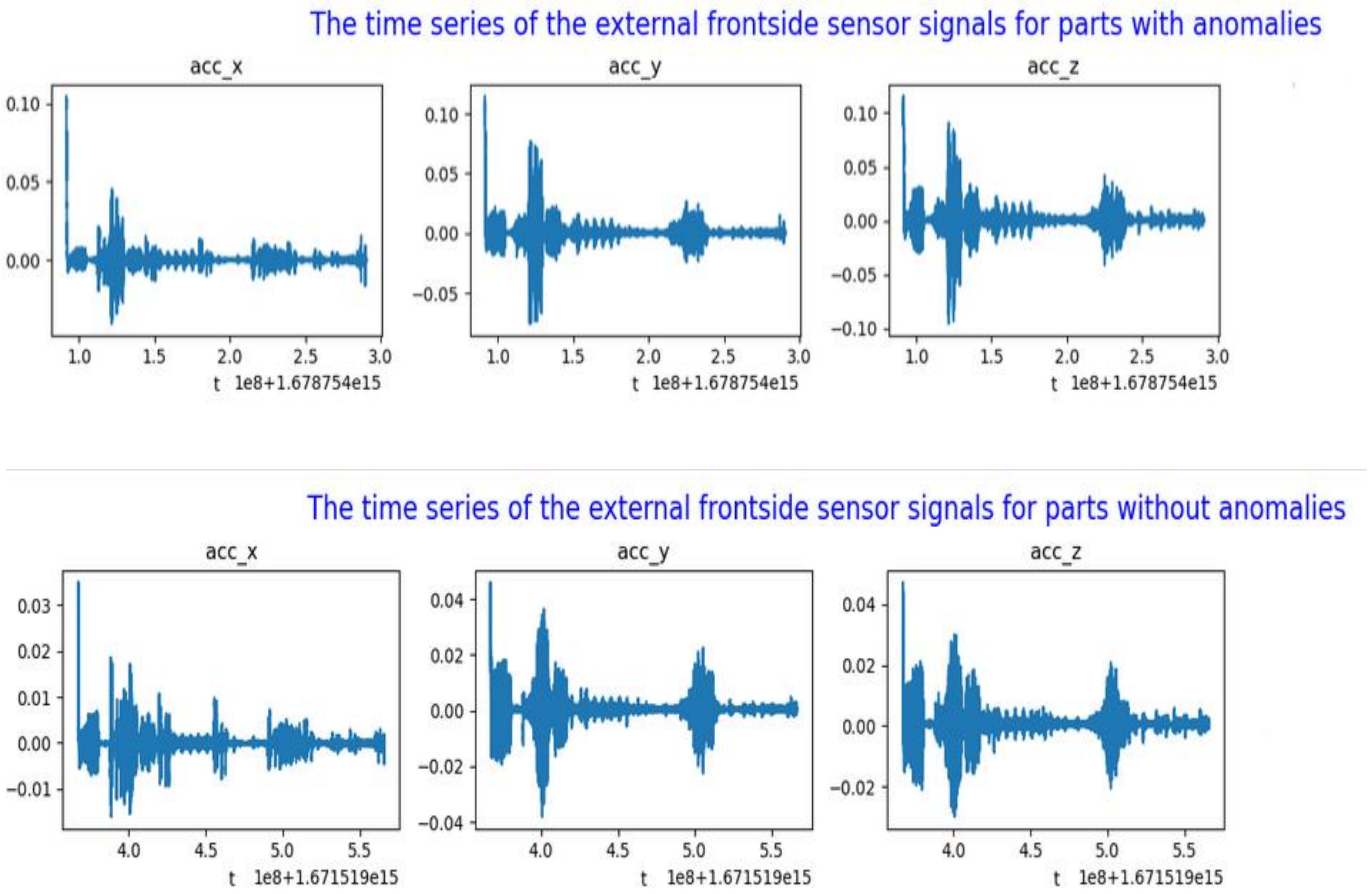
- In this way, the required quality can be ensured before the next production steps and the functionality of the product can be guaranteed at an early stage





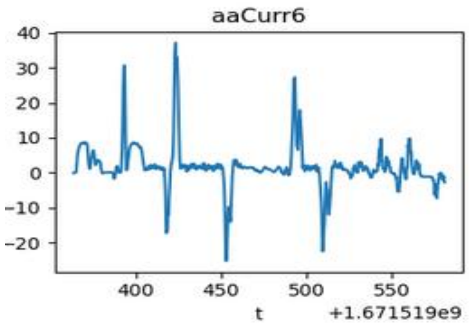
Data and Feature Exploration

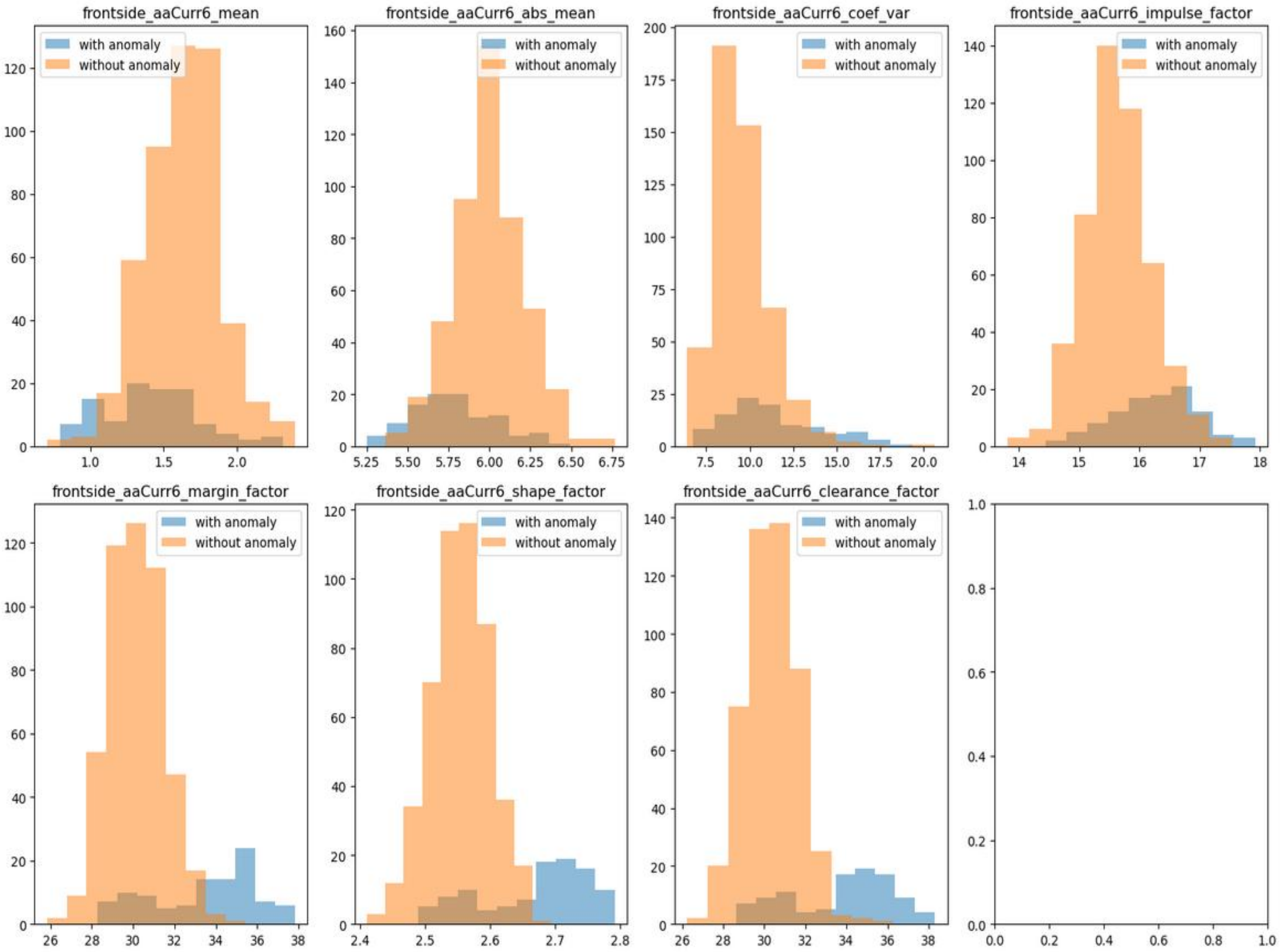
- Seperating true and false parts
- Investigation of time series for different sensors



Data and Feature Exploration

- Deriving features as input data for further processing





Concept and Methodology

TECHNISCHE UNIVERSITÄT DARMSTADT

- Data Preparation:

Feature Extraction → Feature Selection → Data Split → Class Imbalance → Feature Scaling

- Machine Learning Models:

MLP

SVM

RF

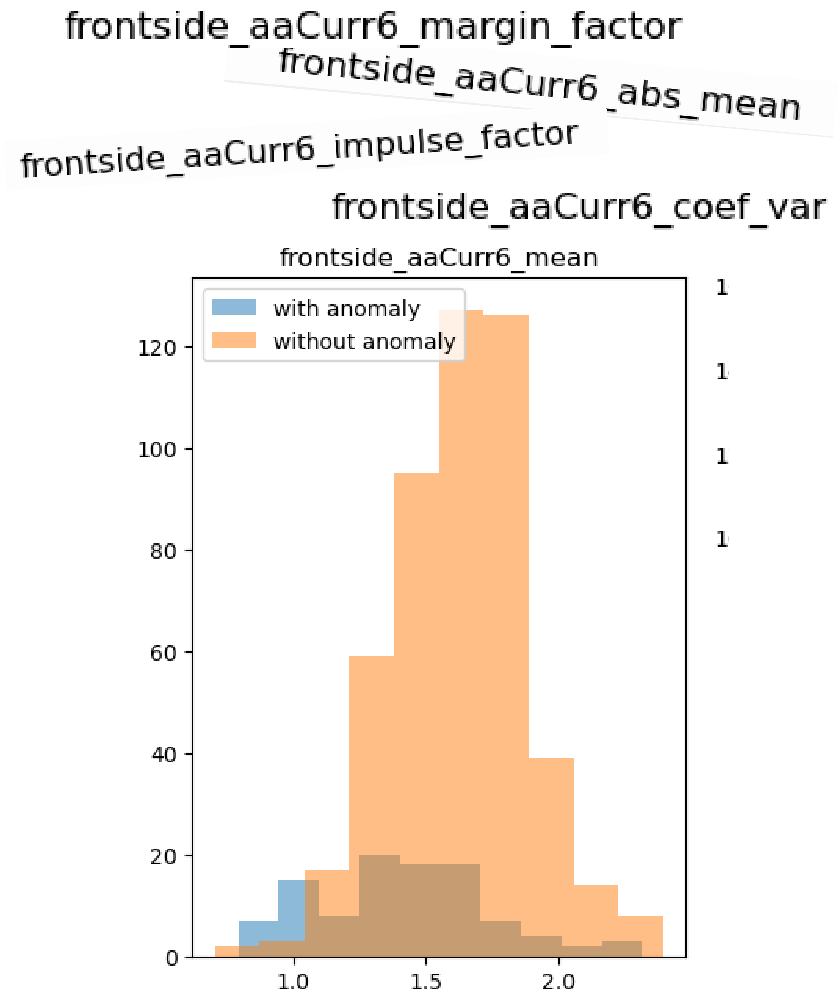
Data Preparation

TECHNISCHE UNIVERSITÄT DARMSTADT

Feature Extraction

Feature Extraction

- Computed various statistical measures for each sensor
- Measures included mean, root mean square, kurtosis, skewness, etc.
- Chosen for their ability to describe time series characteristics.
- Assist in understanding patterns and variations.
- Resulted in 900 features per data point.

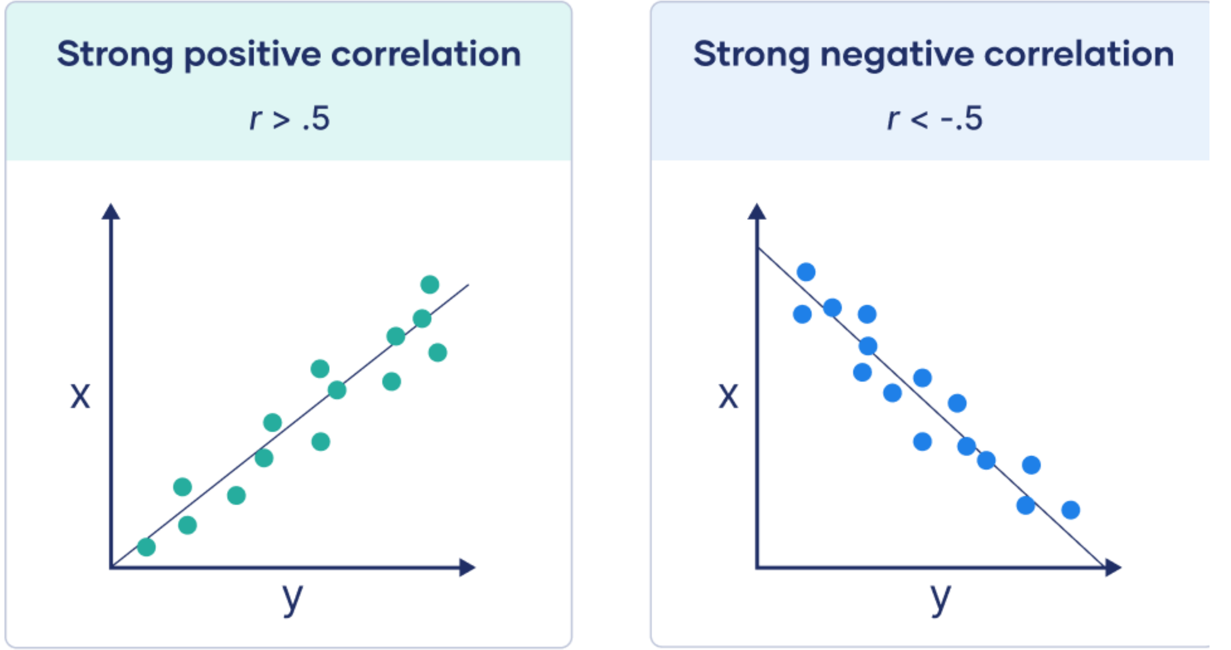


Data Preparation

Feature Extraction → Feature Selection

Feature Selection

- Feature selection reduces data dimensionality.
- A correlation-based method was used.
- Pearson correlation coefficient measured relationship strength.
- Features with correlation < 0.1 were filtered out.
- Reduced features from 900 to 161, significantly.



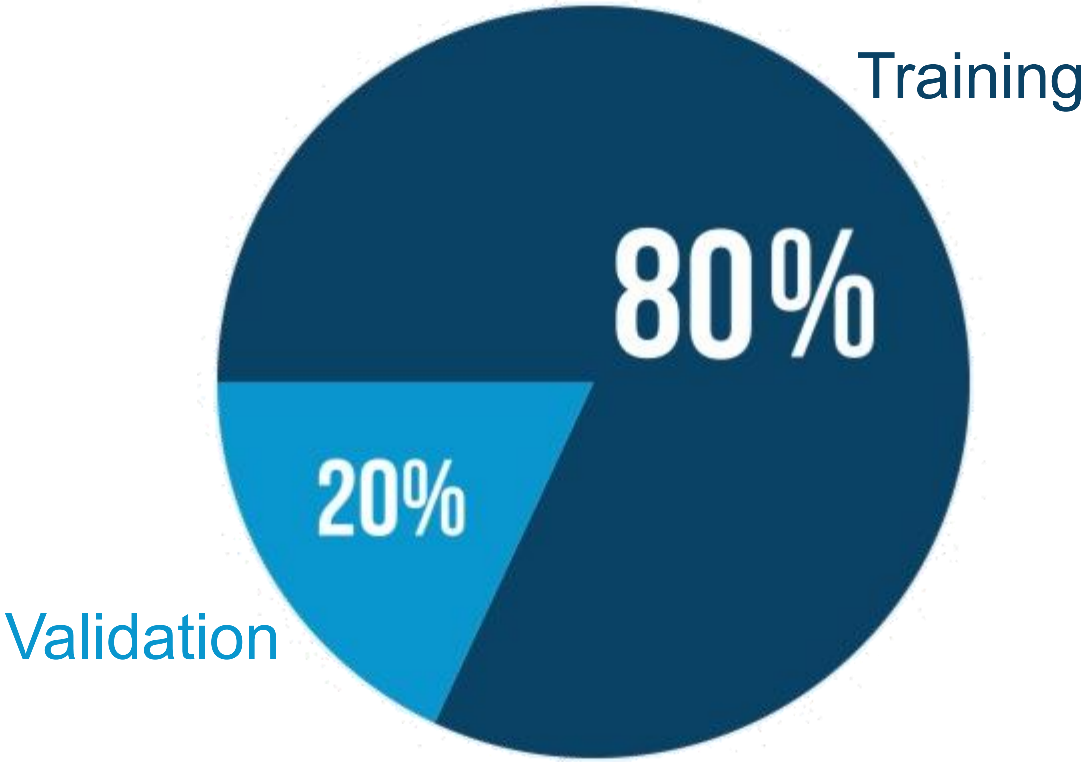
Data Preparation

TECHNISCHE UNIVERSITÄT DARMSTADT

Feature Extraction → Feature Selection → Data Split

Data Split

- Data split into training and validation sets.
- Training set for model training.
- Validation set for model performance evaluation.
- Ratio: 80% training, 20% validation.
- Stratified to preserve class distribution.



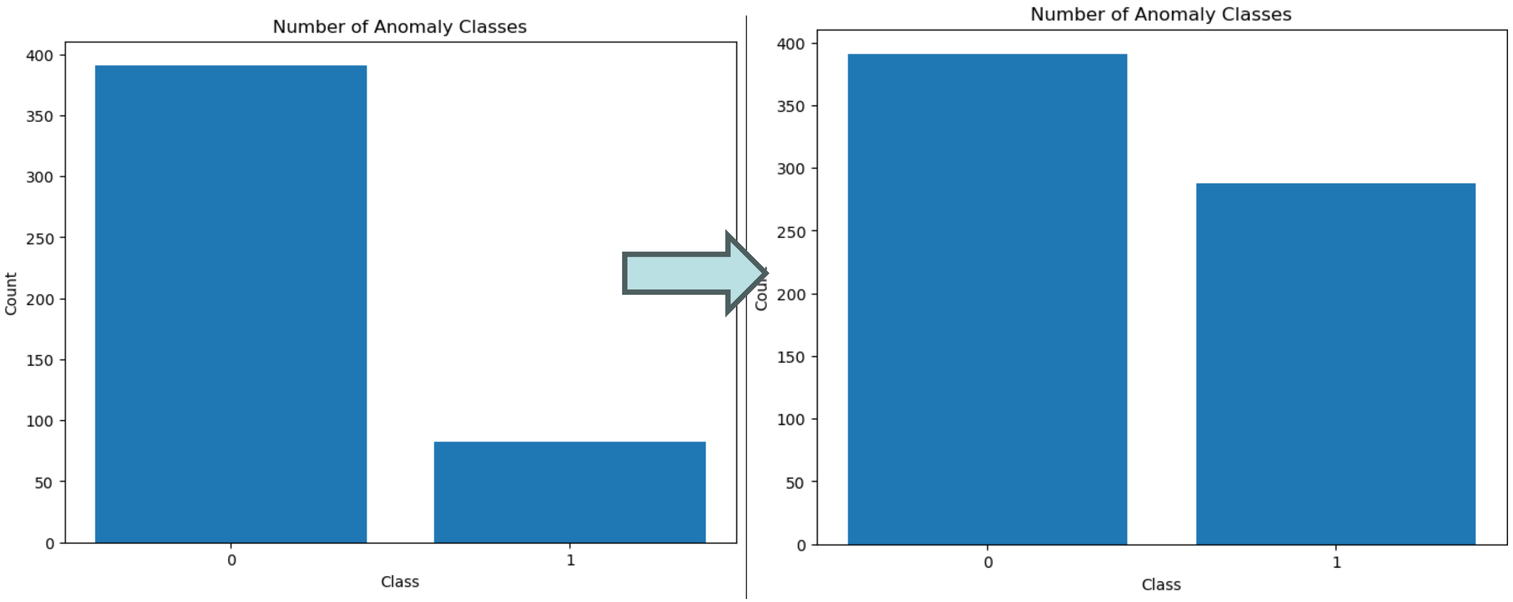
Data Preparation

Feature Extraction → Feature Selection → Data Split → Class Imbalance

Class Imbalance

- Data c class imbalance.
- Addressed using SMOTE technique.
- SMOTE generates new minority class samples.
- Parameters: sampling_strategy.

Oversampled data:



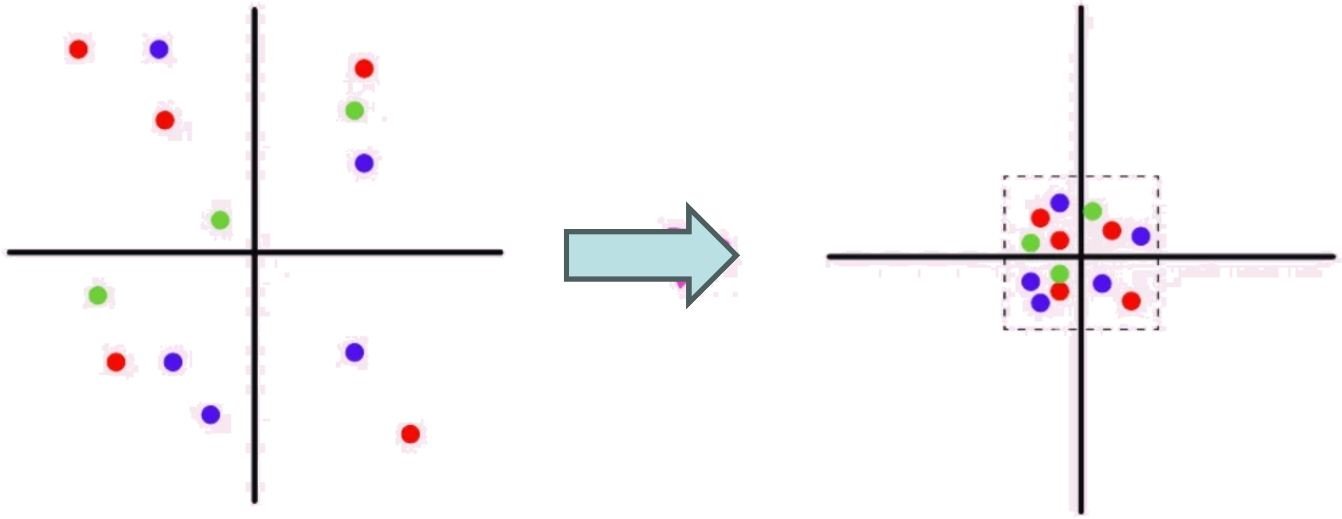
Data Preparation

Feature Extraction → Feature Selection → Data Split → Class Imbalance → Feature Scaling

Feature Scaling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Data scaled using standard scaler.
- Transforms features to have mean zero, std deviation one.
- Improves performance of sensitive models.



Machine Leraning Models

TECHNISCHE UNIVERSITÄT DARMSTADT

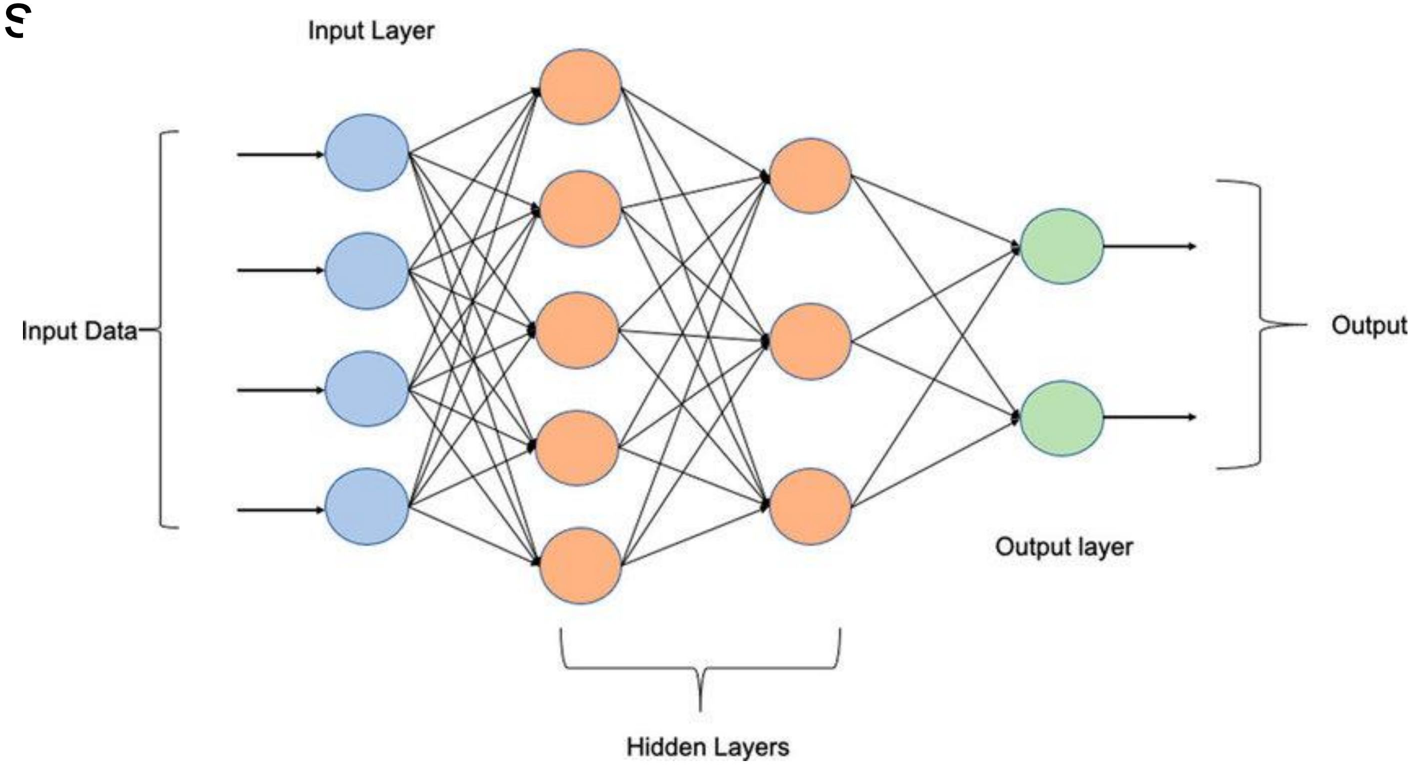
- Multilayer perceptron (MLP)
- Support vector machine (SVM)
- Random forest (RF)

MLP | SVM | RF

Multilayer perceptron (MLP)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Learns complex nonlinear patterns
- Parameters:
  hidden_layer_sizes=(128, 64), activation='relu', alpha=0.01, max_iter=20, random_state=42.
- Chosen based on trial and error.
- Specifies layer sizes, activation function, regularization, and iterations.

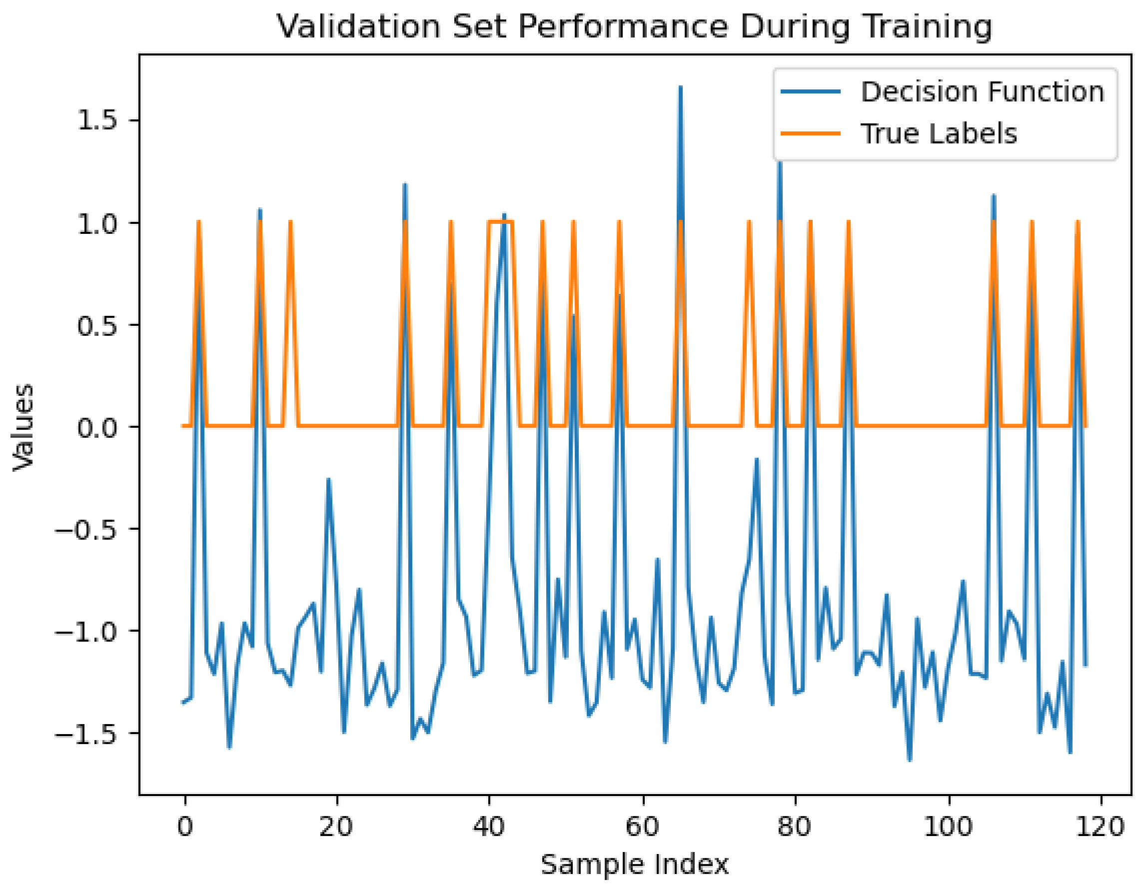


[Multi-layer perceptron (MLP-NN) basic Architecture. | Download Scientific Diagram (researchgate.net)](#)

Support vector machine (SVM)

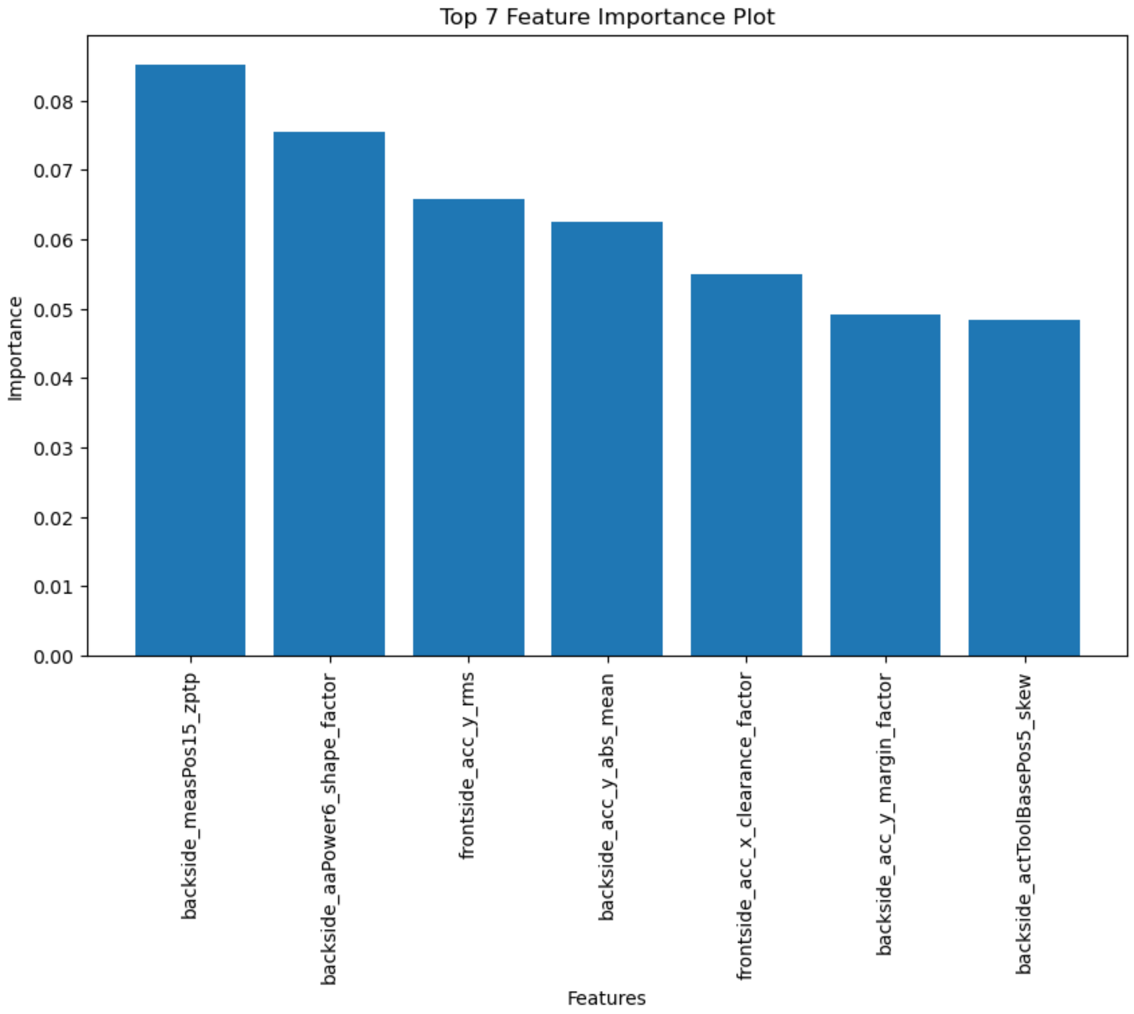
TECHNISCHE UNIVERSITÄT DARMSTADT

- Finds hyperplane to separate data into classes.
- Can handle nonlinear data using kernel function.
- Default parameters: kernel='rbf', C=1.
- 'rbf' kernel commonly used for nonlinear data.
- C parameter set to 1 for moderate penalty balance.



Random forest (RF)

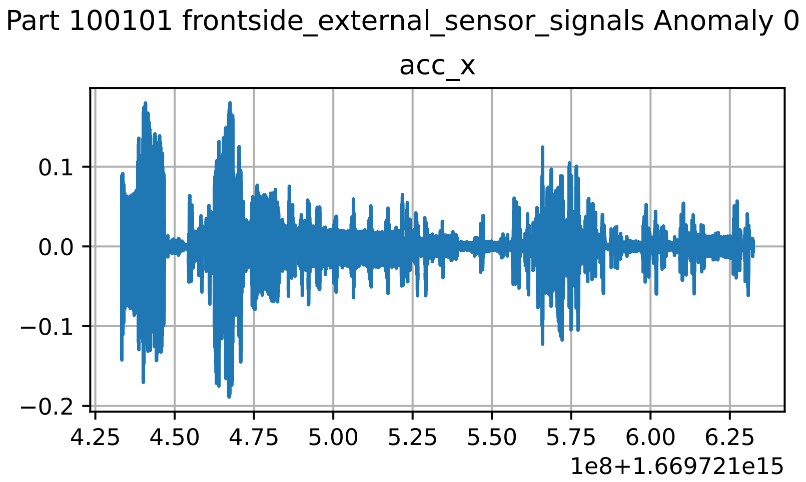
- Ensemble learning combining multiple decision trees.
- Reduces variance and overfitting.
- Chosen based on trial and error.
- n_estimators set to 100 for effective ensemble learning.

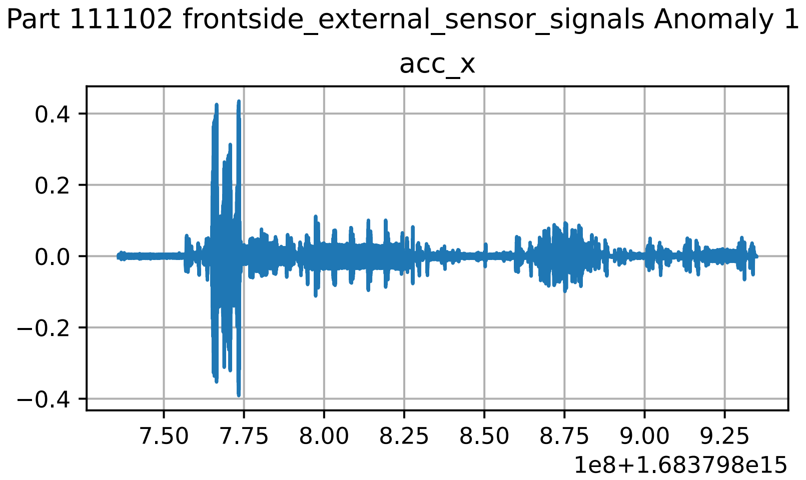


Classification using frontside sensor data

- Frontside acceleration data has visible variations in different anomaly classes

- Used for
  - CNN Time Series Classification
  - Random Forest Classifier
  - Decision Tree Classifier

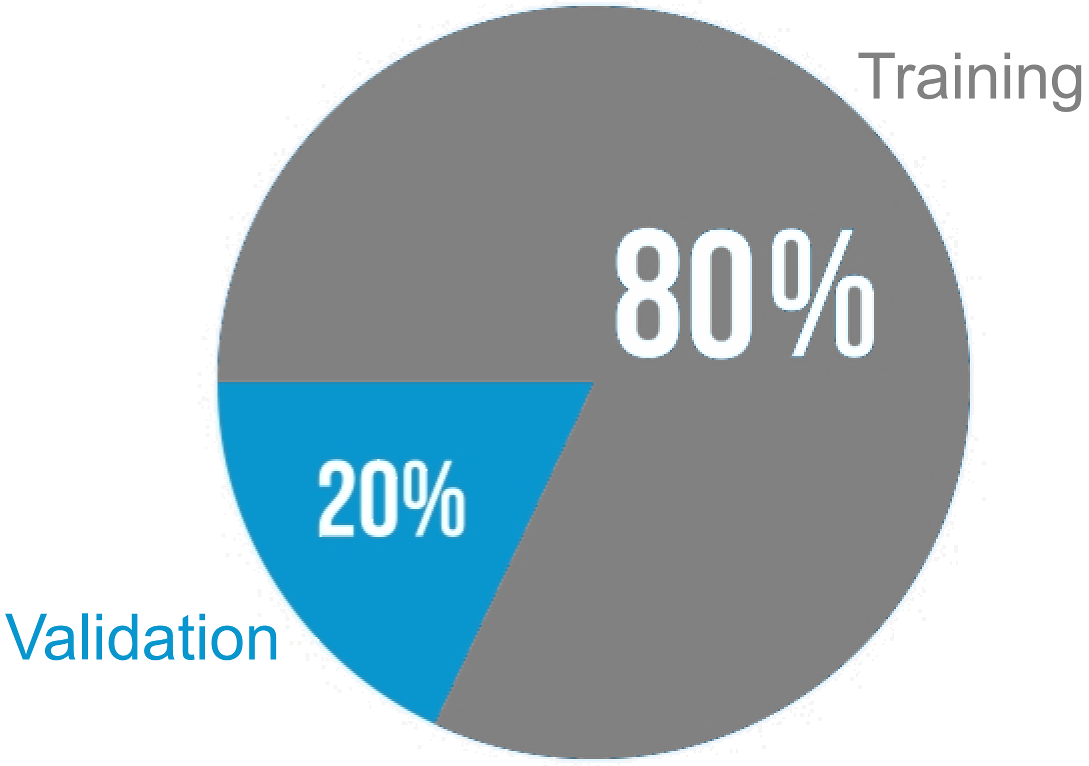




Model Evaluation

- Testing the models on the 20% test split

- Assessment based on the f1-score



Scores

- Highest Score: Random Forest Classifier using correlation-based features at 94.74%
- Promising: Time Series and Random Forest Classification based on acceleration data

	Correlation-based features (all channels)			Frontside acceleration data		
	Multi Layer Perceptron	Support Vector Machine	Random Forest	CNN (Time Series)	Random Forest	Decision Tree
f1-score	79.07%	89.47%	94.74%	93.33%	94.74%	87.18%
False positives	5.04%	0.84%	0%	0%	0.88%	2.63%
False negatives	2.52%	2.52%	1.68%	1.74%	0.88%	1.75%

Final Model

+

High number of features → high chance that useful features can be found

Makes use of all the data channels

Fast actual training

−

Feature correlation might be coincidental

Can‘t detect details in the signal course

Needs heavy data preprocessing

Applicability of the ML Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Models are already quite accurate
- Can help reduce quality control expenses
- Not reliable yet

Strategies for Improvement

- Stronger oversampling of anormal parts
  - Should improve reliability
  - Might induce losses in general accuracy

- Collect further data


BIG
DATA