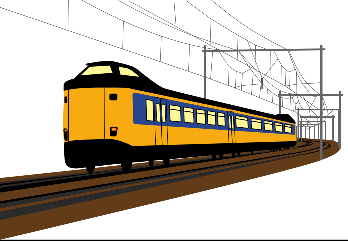
Learning to learn HVAC failures

Layering ML experiments in the absence of ground truth

RSSRail2022
Reliability, Safety and Security of Railway Systems



Learning to learn HVAC failures

Layering ML experiments in the absence of ground truth

Carlos E. Budde[1] **Duncan Jansen**[2]

Inka Locht[3] **Mariëlle Stoelinga**[2,4]

[1] University of Trento, Italy

[2] University of Twente, the Netherlands

[3] Dutch Railways (NS), the Netherlands

[4] Radboud University, the Netherlands

NS

UNIVERSITY OF TWENTE.

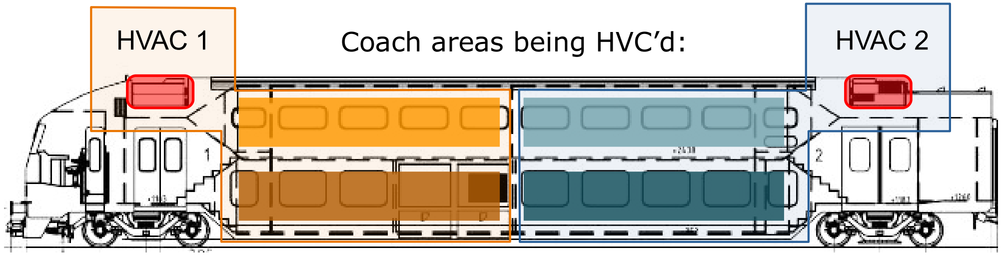
UNIVERSITAS ATHESINA STUDIORUM

RSSRail2022

Reliability, Safety and Security of Railway Systems

Problem: “sudden” failures in HVACs

Heating,
Ventilation, and
Air-
Conditioning units



Each HVAC handles up & down of half coach

Summer months => “Service requests” for **HVAC failures**

Dormant failures:
no use of function → no detection

Motivation: early detection of failures

Self-diagnose of HVACs

Digital codes (ON/OFF) for specific events, e.g. “high pressure in compressor”

ML supervised learning to **correlate the codes to malfunctioning HVACs**

Motivation: early detection of failures

Self-diagnose of HVACs

Digital codes (ON/OFF) for specific events, e.g. “high pressure in compressor”

ML supervised learning to **correlate the codes to malfunctioning HVACs**

But there’s **no automatic detection of HVAC failures**

Malfunctions only found by reported functionality loss

“is it me or it’s getting hot in here?”

Motivation: early detection of failures

Self-diagnose of HVACs

Digital codes (ON/OFF) for specific events, e.g. “high pressure in compressor”

ML supervised learning to **correlate the codes to malfunctioning HVACs**

But there’s **no automatic detection of HVAC failures**

Malfunctions only found by reported functionality loss

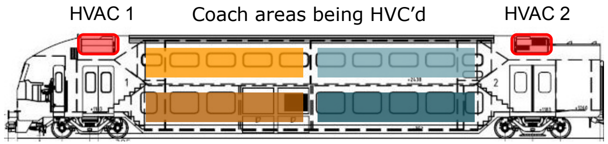
“is it me or it’s getting hot in here?”

Use **temperature in coach** to recognise *cooling* failures

Build an artificial ground truth

Objectives and plan of action

1. Use temperature readings to detect HVAC failures



Compare values of temperature sensors in coaches, to find areas that are too hot

Apply to historic data → build artificial ground truth (**AGT**) for full dataset

2. Use digital codes to detect and foretell HVAC failures

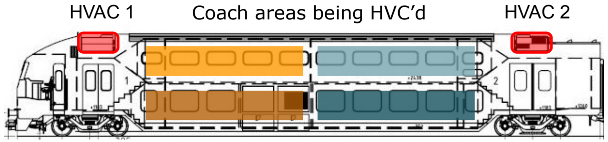
Use AGT (on full dataset) to find useful codes, correlated to failures

Extend to codes that occurred before failures, looking for predictive capabilities of codes

Objectives and plan of action

1. Use temperature readings to detect HVAC failures

Needs human inspection: tedious, error prone. But humans are good ***to interpret plots***



Compare values of temperature sensors in coaches, to find areas that are too hot

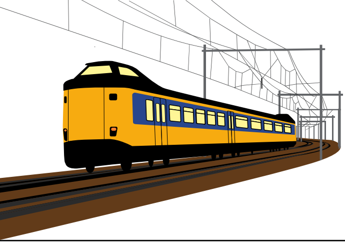
Apply to historic data → build artificial ground truth (**AGT**) for full dataset

2. Use digital codes to detect and foretell HVAC failures

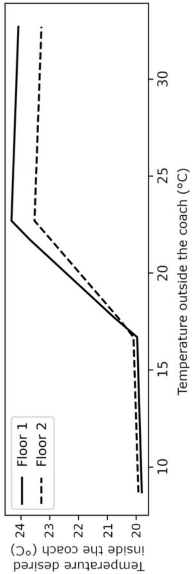
Interpreting digital data (a mesh of codes like "resistor fault" that turns on and then off) is harder

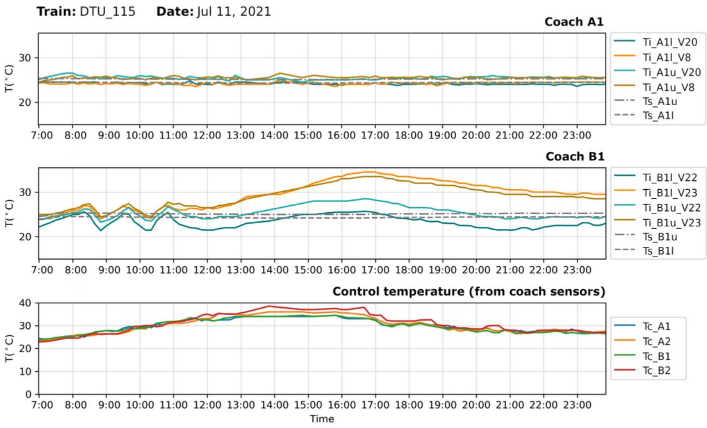
Use AGT (on full dataset) to find useful codes, correlated to failures

Extend to codes that occurred before failures, looking for predictive capabilities of codes

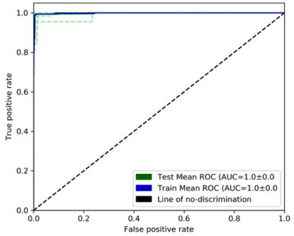


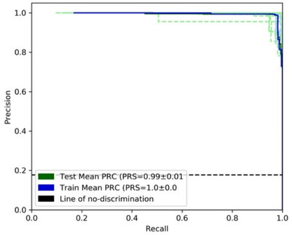
ML layer 1: temperature data to identify failures

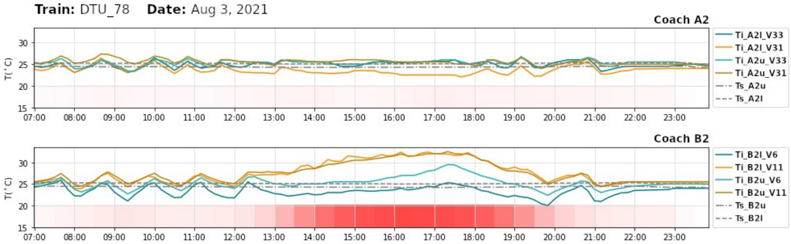




train	coach	HVAC	date	period	symptom
DTU_115	A1	V20	11.07.2021	07:00-23:50	healthy
DTU_115	A1	V8	11.07.2021	07:00-23:50	healthy
DTU_115	B1	V22	11.07.2021	15:00-23:50	healthy
DTU_115	B1	V23	11.07.2021	15:00-23:50	hot

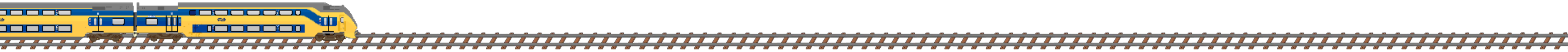






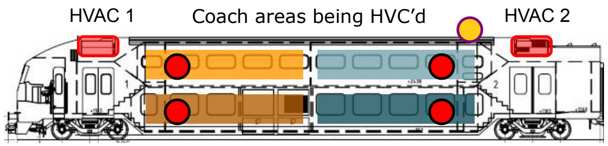
RSSRail2022
Reliability, Safety and Security of Railway Systems

L1—temp.: input & data preprossesing



INPUT

Continuous data: **temperatures read in time** inside and outside the coaches



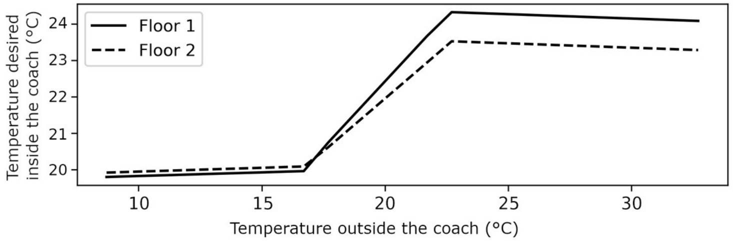
Per coach:

4 internal sensors

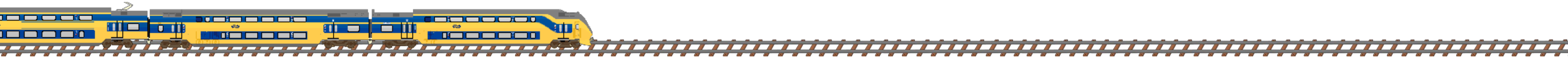
1 external sensor

PREPROCESSING

1. Temperatures < -20 °C or > 60 °C → replace by **`NaN`**
2. Missing data:
   a) Less than 90 min → use linear interpolation
   b) More than 90 min → replace by **`NaN`**
3. Compute desired temperature inside the coaches

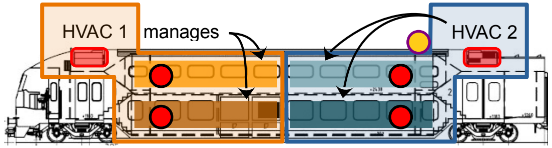


L1—temp.: human data labelling “bootstrap”



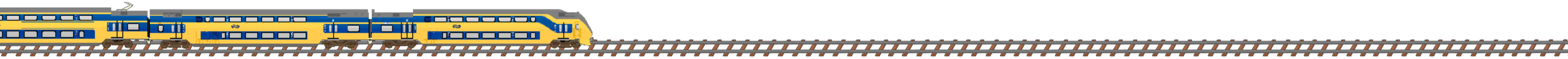
Training data for supervised learning? Must build it by hand!

(there’s no ground truth: true failures are not in data)



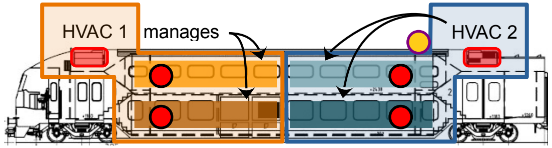
1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

L1—temp.: human data labelling "bootstrap"

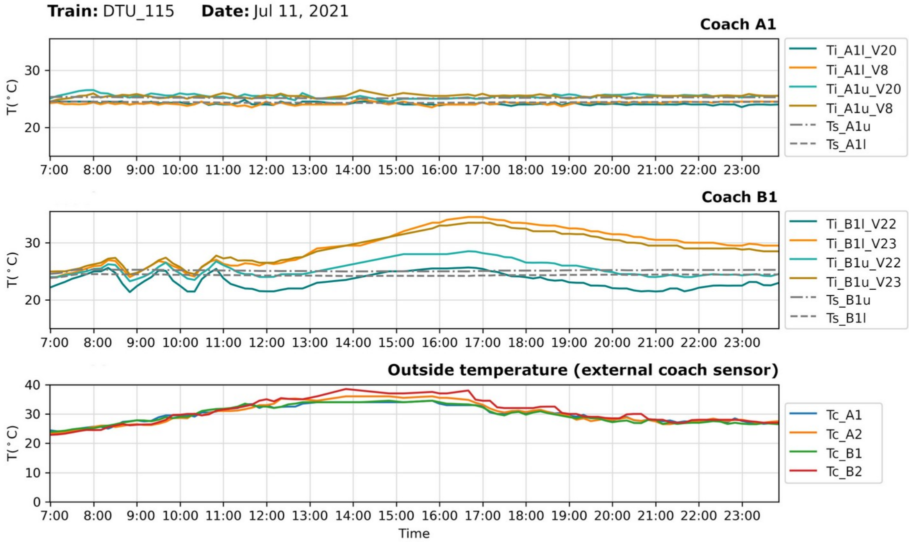


Training data for supervised learning? Must build it by hand!

(there's no ground truth: true failures are not in data)

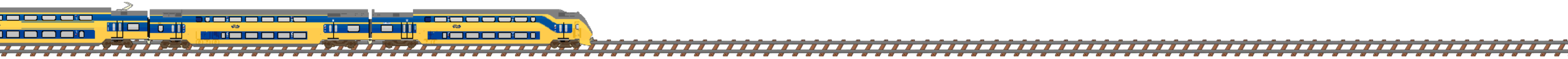


Time-aligned plots for temperatures in coaches, that the human eye is good to interpret



1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

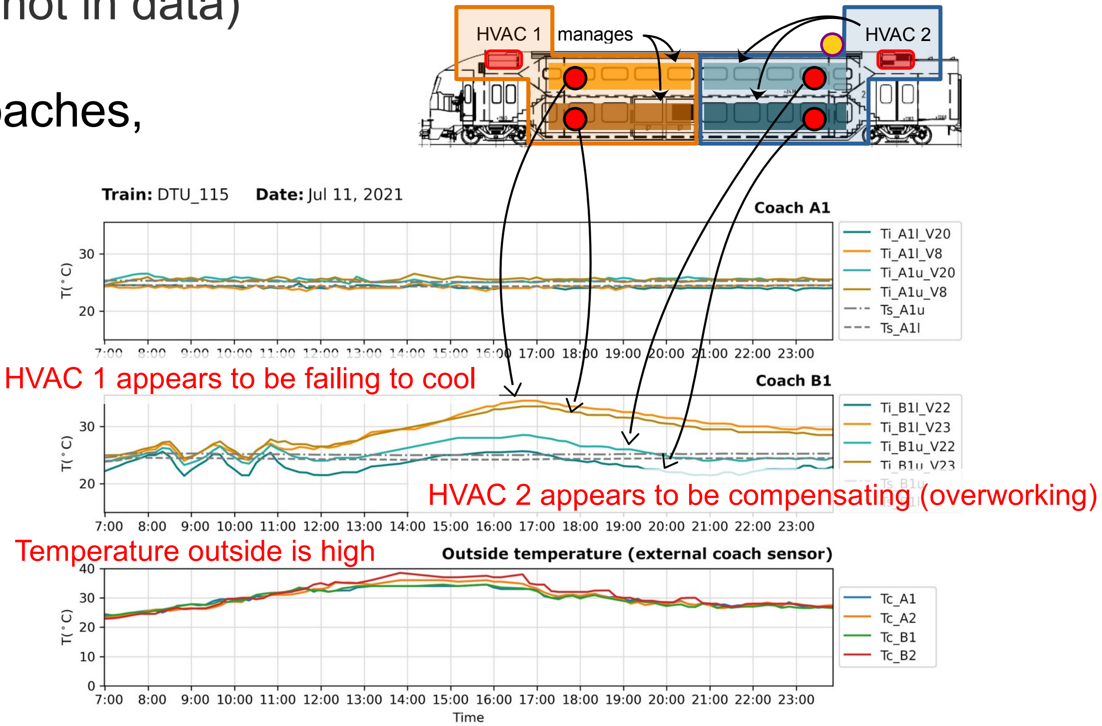
L1—temp.: human data labelling “bootstrap”



Training data for supervised learning? Must build it by hand!

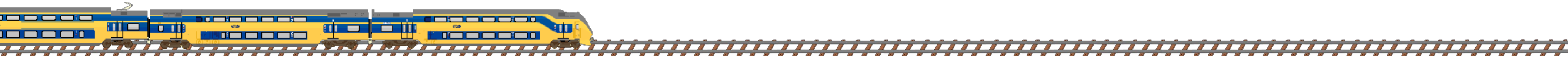
(there’s no ground truth: true failures are not in data)

Time-aligned plots for temperatures in coaches, that the human eye is good to interpret



1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

L1—temp.: human data labelling “bootstrap”



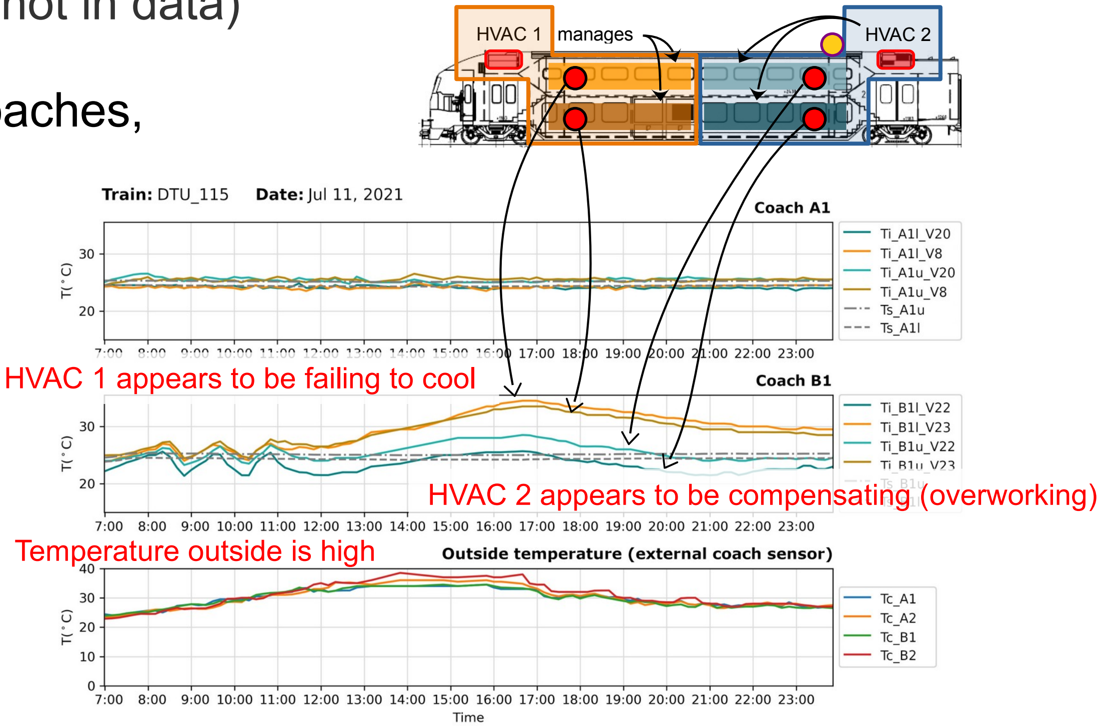
Training data for supervised learning? Must build it by hand!

(there’s no ground truth: true failures are not in data)

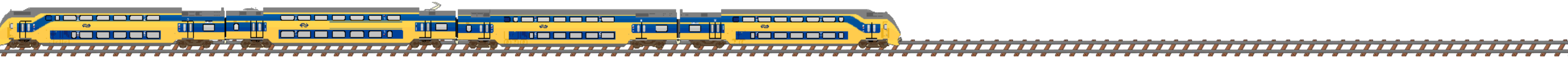
Time-aligned plots for temperatures in coaches, that the human eye is good to interpret

train	coach	HVAC	date	period	symptom
DTU_115	A1	V20	11.07.2021	07:00-23:50	healthy
DTU_115	A1	V8	11.07.2021	07:00-23:50	healthy
DTU_115	B1	V22	11.07.2021	15:00-23:50	healthy
DTU_115	B1	V23	11.07.2021	15:00-23:50	hot

1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT



L1—temp.: build AGT via LR



train	coach	HVAC	date	period	symptom
DTU_115	A1	V20	11.07.2021	07:00-23:50	healthy
DTU_115	A1	V8	11.07.2021	07:00-23:50	healthy
DTU_115	B1	V22	11.07.2021	15:00-23:50	healthy
DTU_115	B1	V23	11.07.2021	15:00-23:50	hot

Our human(s) labels covered 1 % of temperature data
Machine-learn the failures for the rest of the data

Logistic Regression, I choose you *(fast + overfitting-resistant + yields prob. of failure + good for linear data)*

True Positive: LR models says “HVAC cooling failure” and human label says “hot” aka failure
True Negative: LR models says “HVAC doing just fine” and human label says “healthy” aka okay
FP & **FN** anal.

1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

“No label” is also possible: neither failure nor okay

Values that don’t indicate HVAC proper- nor mal-function, e.g. if outside temperature is low so no cooling is needed

L1—temp.: build AGT via LR: features

train	coach	HVAC	date	period	symptom
DTU_115	A1	V20	11.07.2021	07:00-23:50	healthy
DTU_115	A1	V8	11.07.2021	07:00-23:50	healthy
DTU_115	B1	V22	11.07.2021	15:00-23:50	healthy
DTU_115	B1	V23	11.07.2021	15:00-23:50	hot

Our human(s) labels covered 1 % of temperature data

Machine-learn the failures for the rest of the data

ML features from INPUT data:

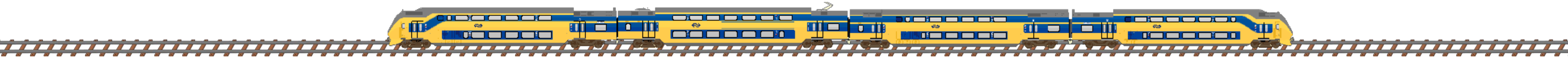
- Set Point (**SP**): difference between desired temperature (per coach, side, up/down) and actual temp.
- Compensation Behaviour (**CB**): temp. difference in opposite sides of the coach (up/sown)
- Comparison to Other Coaches (**COC**): temp. difference from a coach to the median of the train
- Defective Control Sensor (**DCS**): temp. difference in this coache's outside sensor to the train's median

1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

Trained an **LR model** on these features

- NS temperature data for 2 months of ca. 180 trains
- Imbalance of ratio 5:1 in favour of the "healthy" label

L1—temp.: check quality of the human boostrapping



But how reliable is that human-labelling part?

Two independent persons followed same what-to-label instructions

Inter-rater measurement of the resulting coherence:

“Are the human labels too subjective?”

1. Manual label 1–5 % of data
2. Extrapolate to build the full artificial ground truth (AGT)
3. Verify the quality of the AGT

L1—temp.: check quality of the human boostrapping

But how reliable is that human-labelling part?

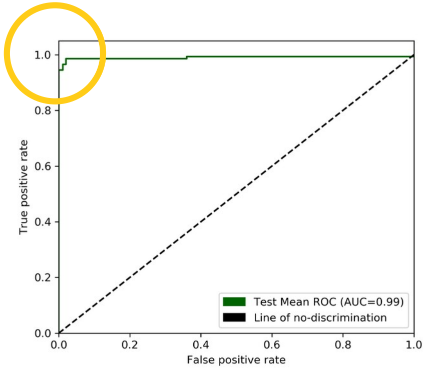
Two independent persons followed same what-to-label instructions

Inter-rater measurement of the resulting coherence:

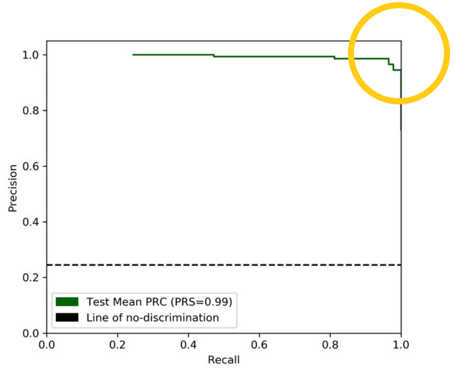
“Are the human labels too subjective?”

Answer*: independent human labels are coherent,* **AGT is useful**

1. **Manual label 1–5 % of data**
2. **Extrapolate to build the full artificial ground truth (AGT)**
3. **Verify the quality of the AGT**



ROC

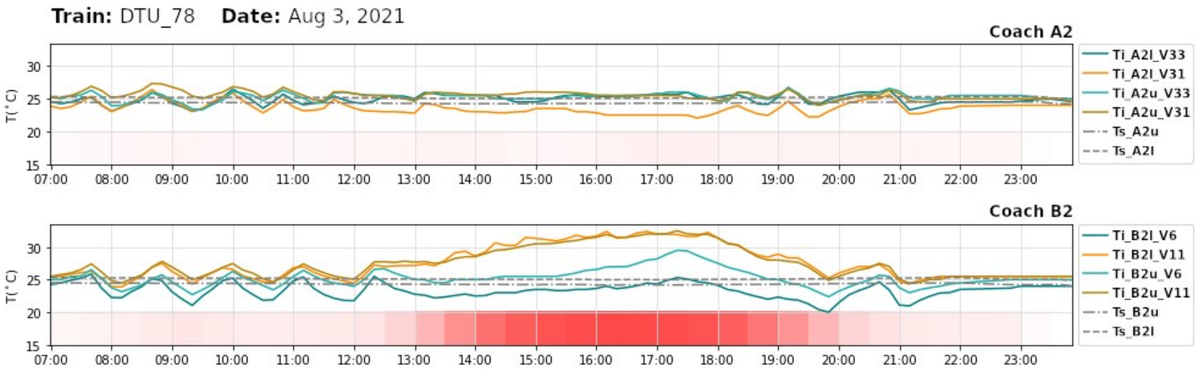


PRC

L1—temp.: LR model tells prob. of cooling failure

Alright, it works, let's use it!

LR model in action:



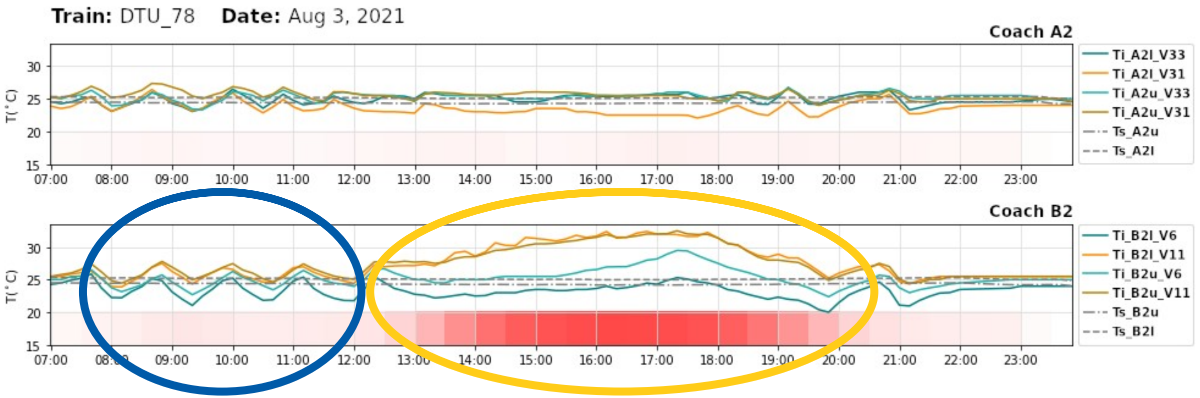
Color intensity directly proportional to probability of cooling malfunction *(red = brace yourselves)*

L1—temp.: LR model tells prob. of cooling failure

Alright, it works, let's use it!

LR model in action:

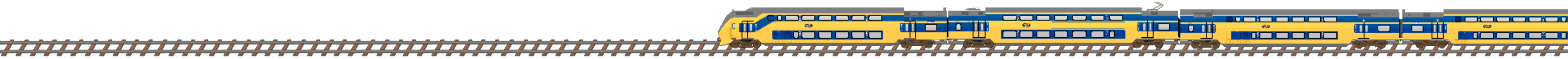
Color intensity directly proportional to probability of cooling malfunction
(red = brace yourselves)



Model correctly **identifies values of likely cooling malfunctions**

Model ignores erratic values, reflecting the human-labelling rules

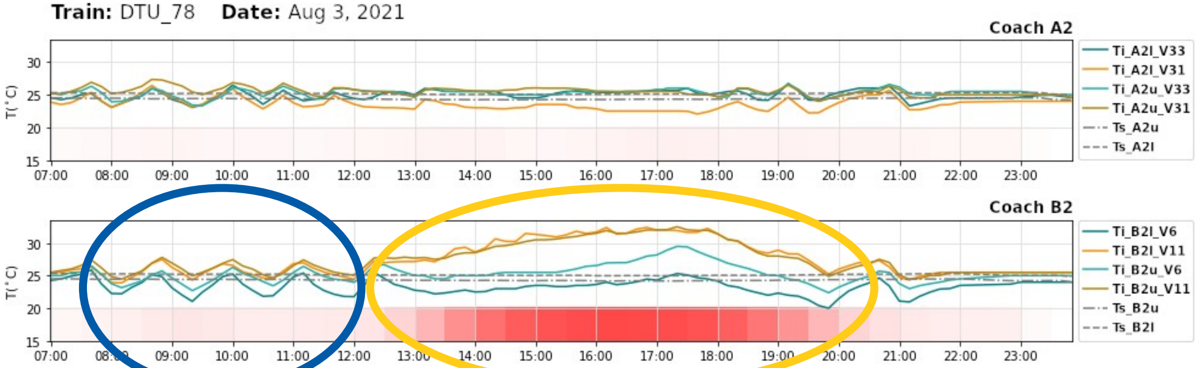
L1—temp.: LR model tells prob. of cooling failure



Alright, it works, let's use it!

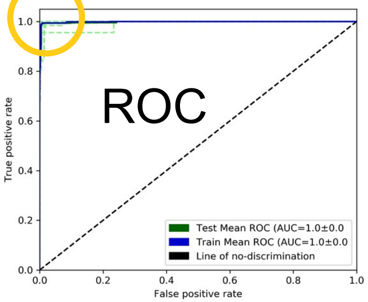
LR model in action:

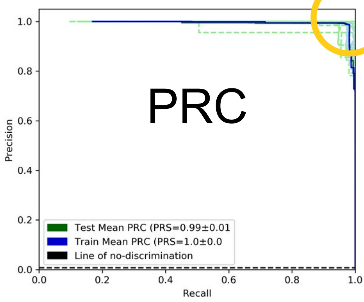
Color intensity directly proportional to probability of cooling malfunction *(red = brace yourselves)*



Model correctly **identifies values of likely cooling malfunctions**

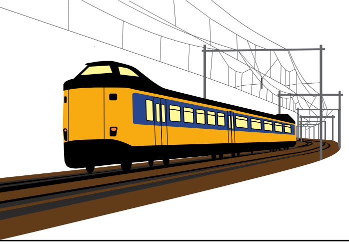
Model ignores erratic values, reflecting the human-labelling rules





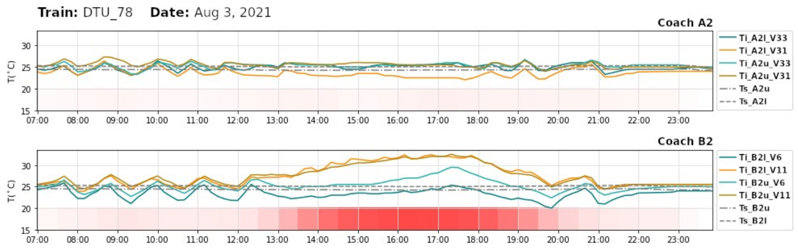
Model quality:

- High predictive capability (ROC, PRC)
- No evidence of overfitting (PRC)



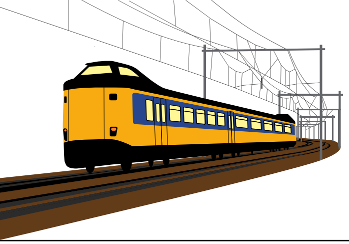
ML layer 1: temperature data to identify failures

<u>Objective 1</u>: “use temperature readings to detect HVAC failures”



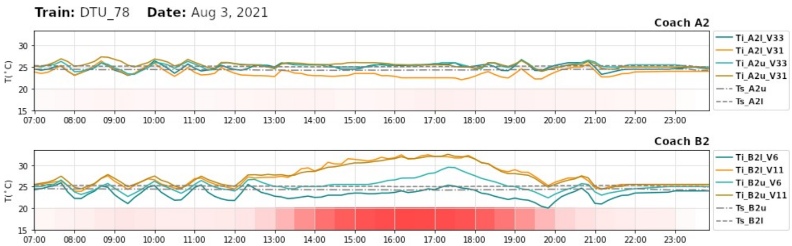
Process and results:

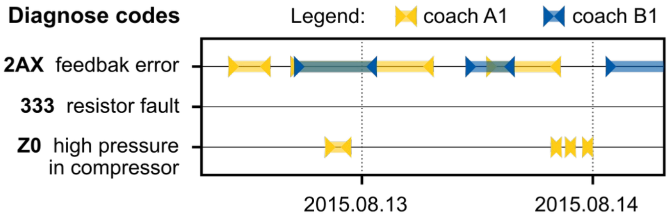
- Generated Logistic Regression model on temperature data input
- Bootstrapped by human labels in 1 % of the data
- Model tells probability of an HVAC cooling failure at present, with high accuracy
- Extrapolated to historic data, serves as AGT for further studies (i.e. detects failures)

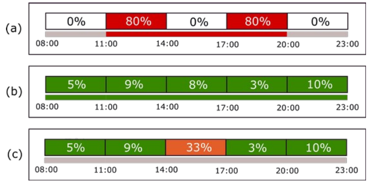


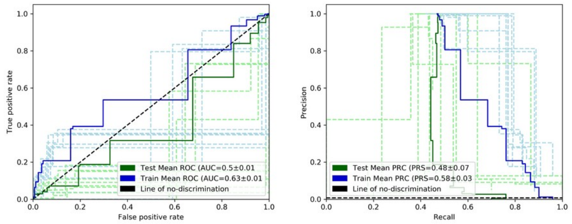
ML layer 2:
digital data to identify/foretell failures

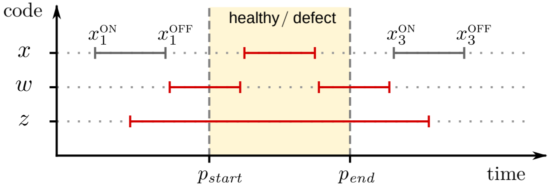
<u>Objective 2</u>: “use digital codes to detect and foretell HVAC failures”











RSSRail2022
Reliability, Safety and Security of Railway Systems

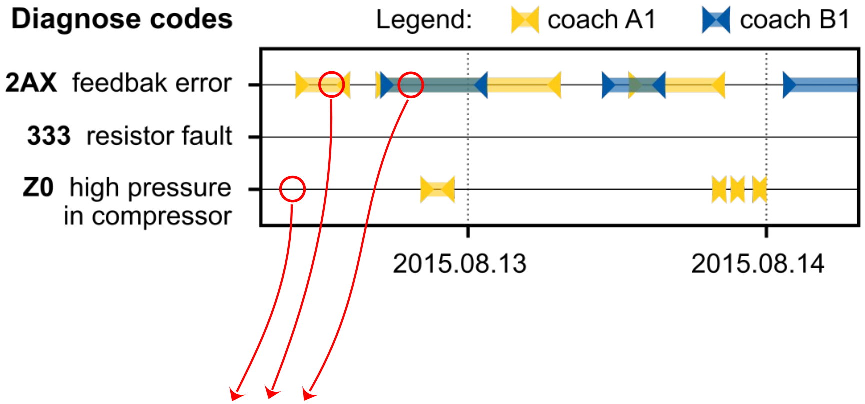
L2—codes: input & data preprossesing

INPUT

Digital data: **diagnostic codes that turn ON/OFF in time**
i.e. also as a time series, but Boolean

PREPROCESSING

1. Make code IDs unique *per failure type*
   a) E.g. if code includes coach ID, group values
2. Missing deactivation (“OFF”) time of code?
   a) Either discard, or else…
   b) Insert default deact. time (midnight)



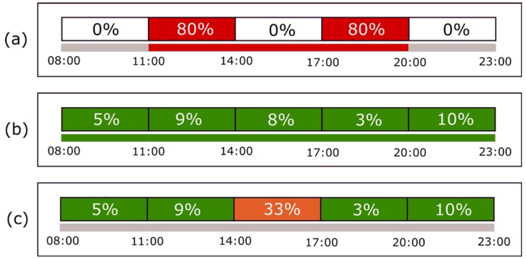
At any point in time
each individual code is either ON or OFF

L2—codes: AGT-based data labelling

Training data for supervised learning? **Use AGT built in previous step**

Tech details:

- Match temperature data to codes data by **discretising the day in windows**
- Use AGT (prob. value of HVAC failure) to label each window
  - As before, 3 labels: failure, okay, dunno.

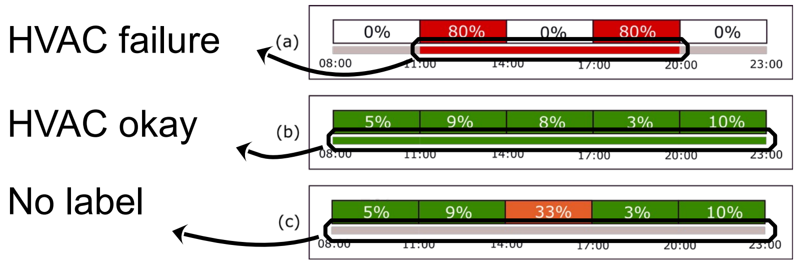


L2—codes: AGT-based data labelling

Training data for supervised learning? **Use AGT built in previous step**

Tech details:

- Match temperature data to codes data by **discretising the day in windows**
- Use AGT (prob. value of HVAC failure) to label each window
  - As before, 3 labels: failure, okay, dunno.

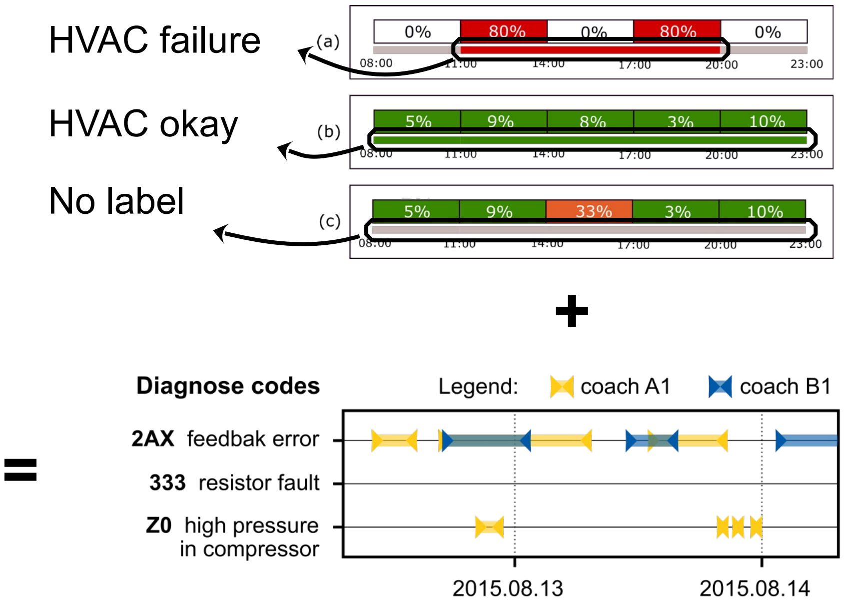


L2—codes: AGT-based data labelling

Training data for supervised learning? **Use AGT built in previous step**

Tech details:

- Match temperature data to codes data by **discretising the day in windows**
- Use AGT (prob. value of HVAC failure) to label each window
  - As before, 3 labels: failure, okay, dunno.

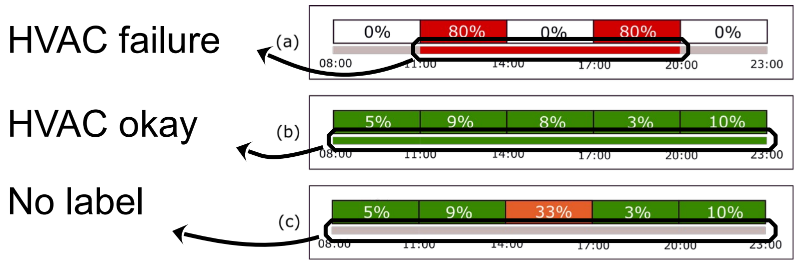


L2—codes: AGT-based data labelling

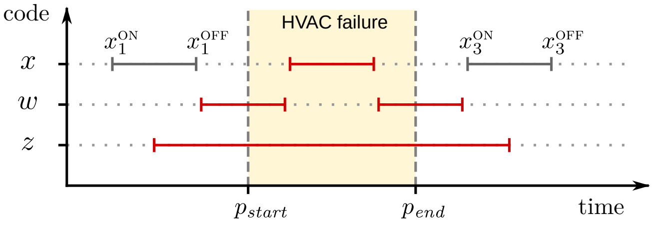
Training data for supervised learning? **Use AGT built in previous step**

Tech details:

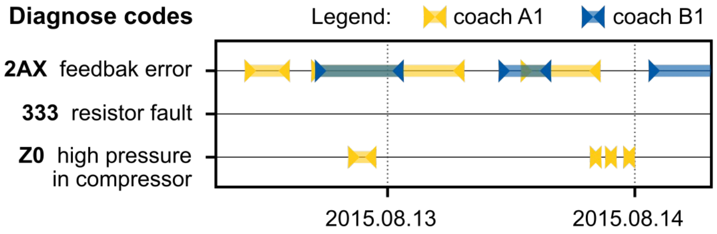
- Match temperature data to codes data by **discretising the day in windows**
- Use AGT (prob. value of HVAC failure) to label each window
  - As before, 3 labels: failure, okay, dunno.



+



=



In red, codes present during HVAC failure

L2—codes: train ML model(s)

Logistic Regression + also **Decision Trees** → *Can tell how each code contributes to total prob. of HVAC failure*

ML features from INPUT data:

(time window of this)

- Code During Period (**CDP**): is this code active during this failure/okay period?
- Code: Number of Days (**CND**): from the end of each failure/okay window, how many days passed since this code was last seen?
- Code: Number of Occurrences (**CNO**): from the end of each failure/okay window, how many distinct times was this code turned on?
- Code: Cumulative Time (**CCT**): as CNO above, but count the accumulated time of this code

Up to 30 days back = periodic mainten. of trains

L2—codes: train ML model(s)

Logistic Regression + also **Decision Trees**

Can tell how each code contributes to total prob. of HVAC failure

Address detection objective

ML features from INPUT data:

(time window of this)

- Code During Period (**CDP**): is this code active during this failure/okay period?
- Code: Number of Days (**CND**): from the end of each failure/okay window, how many days passed since this code was last seen?
- Code: Number of Occurrences (**CNO**): from the end of each failure/okay window, how many distinct times was this code turned on?
- Code: Cumulative Time (**CCT**): as CNO above, but count the accumulated time of this code

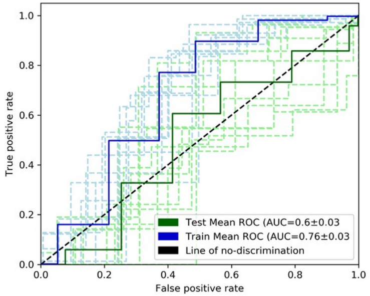
Up to 30 days back = periodic mainten. of trains

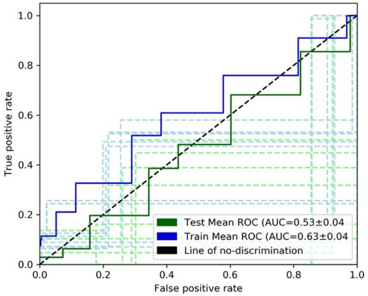
Address prediction objective: codes occurring before failure observed

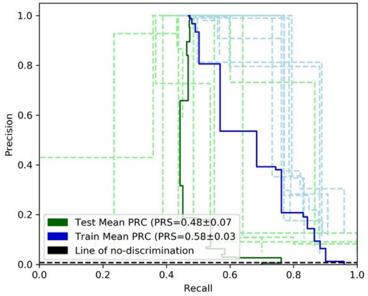
L2—codes: ML experimental results

Model: **Logistic Regression** | **Decision Tree**

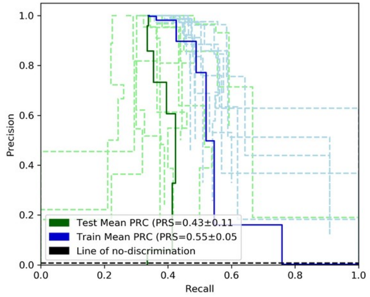
ROC

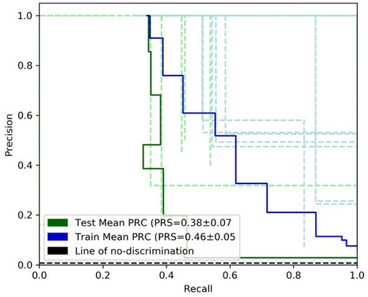


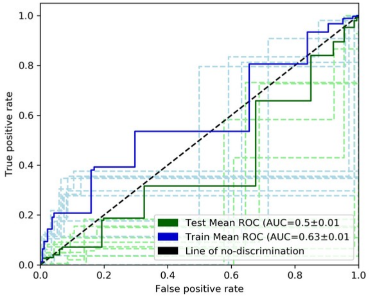




PRC





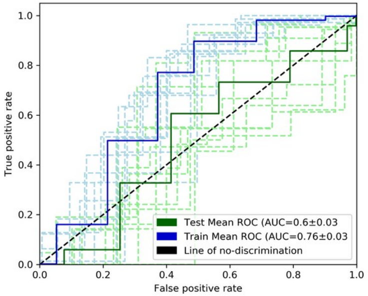


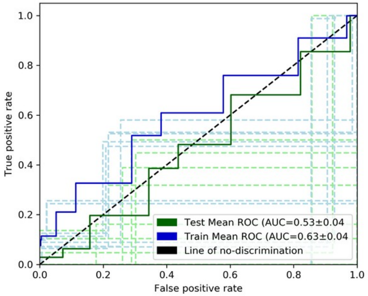
Objective: Detection | Detection | Prediction

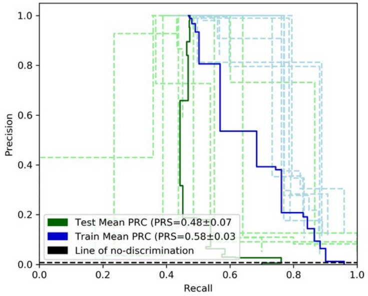
L2—codes: ML experimental results

Model: **Logistic Regression** | **Decision Tree**

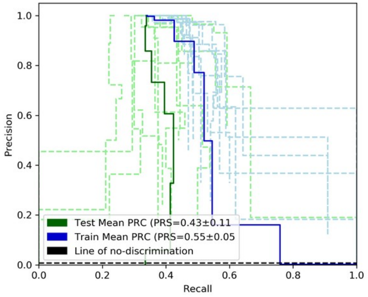
ROC

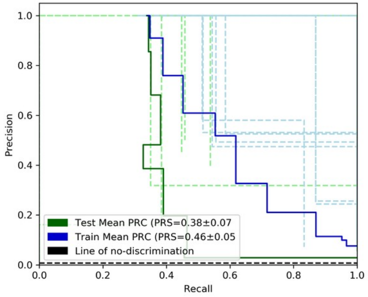


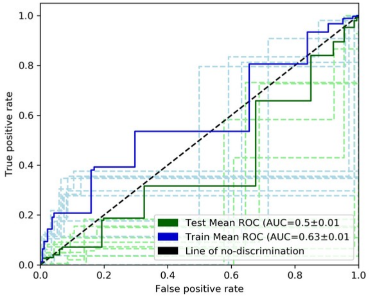




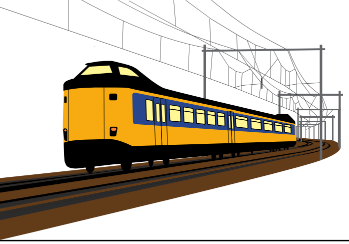
PRC





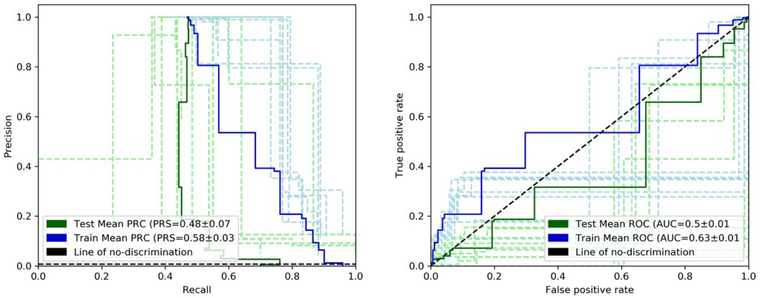


Objective: Detection | Detection | Prediction



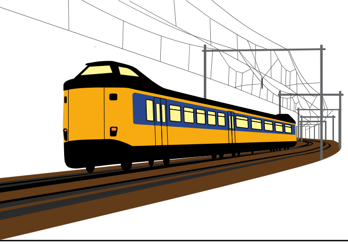
ML layer 2: digital data to identify/foretell failures

<u>Objective 2</u>: “use digital codes to detect and foretell HVAC failures”



Process and results:

- Generated LR and DT models from codes data
- Trained on AGT from previous step
- Neither model shows good capabilities:
  - To detect current ongoing HVAC faliures
  - To foretell HVAC failures to come



General conclusions

T = temperature data readings inside and outside trains

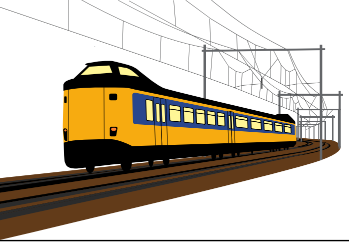
D = diagnose codes from HVACs corresponding to those readings

- LR model (**T** , human_labels(1% of **T**)) = ⋯

  ⋯ = good detection of HVAC failures + can derive AGT for full dataset

- LR model (**D**, AGT) = DT model (**D**, AGT) = bad detection of HVAC failures

- DT model (**D**, AGT) = bad prediction of HVAC failures



General conclusions

T = temperature data readings inside and outside trains

D = diagnose codes from HVACs corresponding to those readings

Overfitting?
Many types of codes for one type of HVAC failure

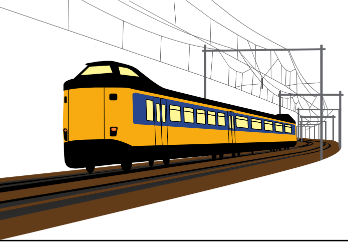
- LR model (**T** , human_labels(1% of **T**)) = ···

  ··· = good detection of HVAC failures + can derive AGT for full dataset

- LR model (**D**, AGT) = DT model (**D**, AGT) = bad detection of HVAC failures

- DT model (**D**, AGT) = bad prediction of HVAC failures

Lack of data?
Only trained on 2 months of codes



General conclusions

¿Prediction!
Try temperature-based
ML to foretell failures

Overfitting?
Many types of
codes for one type
of HVAC failure

T = temperature data readings inside and outside trains

D = diagnose codes from HVACs corresponding to those readings

- LR model (**T** , human_labels(1% of **T**)) = ⋯

  ⋯ = good detection of HVAC failures + can derive AGT for full dataset

- LR model (**D**, AGT) = DT model (**D**, AGT) = bad detection of HVAC failures

- DT model (**D**, AGT) = bad prediction of HVAC failures

Lack of data?
Only trained on 2
months of codes

Learning to learn HVAC failures

Layering ML experiments in the absence of ground truth

RSSRail2022
Reliability, Safety and Security of Railway Systems

C.E. Budde D. Jansen
I. Locht M. Stoelinga