Data Engineering · Team 6 Project Proposal

Unsupervised Anomaly Detection and Analysis of Stable Operating Conditions in a Spot-Welding Process

Team 6

2026. 10. 20. · Using the KAMP Welding Machine AI Dataset

We chose the KAMP Welding Machine AI Dataset: real process data, applicable to the course.

Data Source — KAMP



Manufacturing AI platform led by the Ministry of SMEs and Startups

KAMP Welding Machine AI Dataset (provided via competition)

Why This Dataset

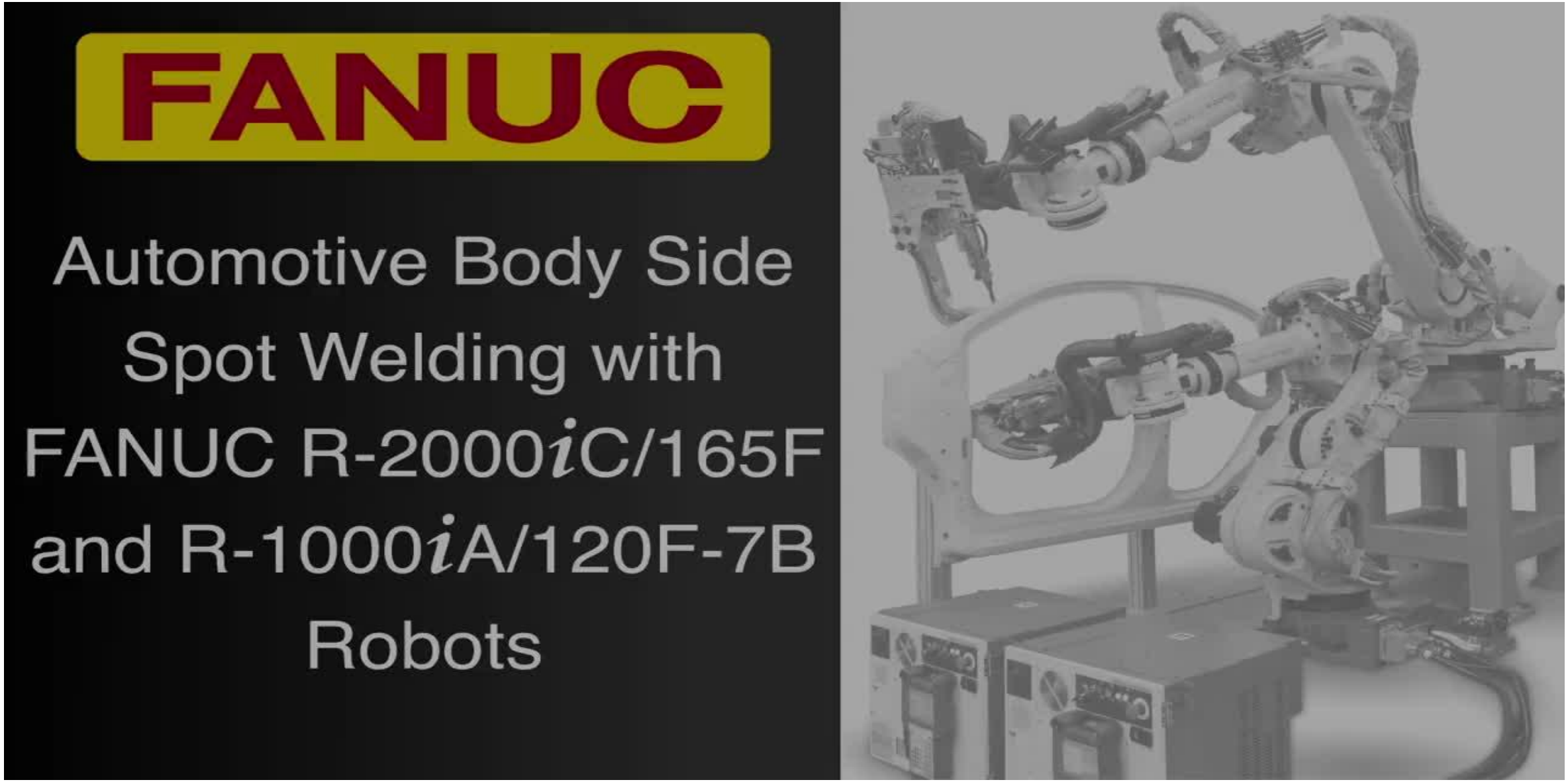
1 Credible public data
Published with a guidebook on a government (MSS)-led platform
→ Clear source and terms of use

2 Real shop-floor data
PLC–MES logs from a real robot welding line, not a simulation
→ Real recording paths and record units can be examined

3 Industrially important problem
~98% of welds in a passenger car are spot welds
→ Quality (nugget) is invisible, so data-driven screening is needed

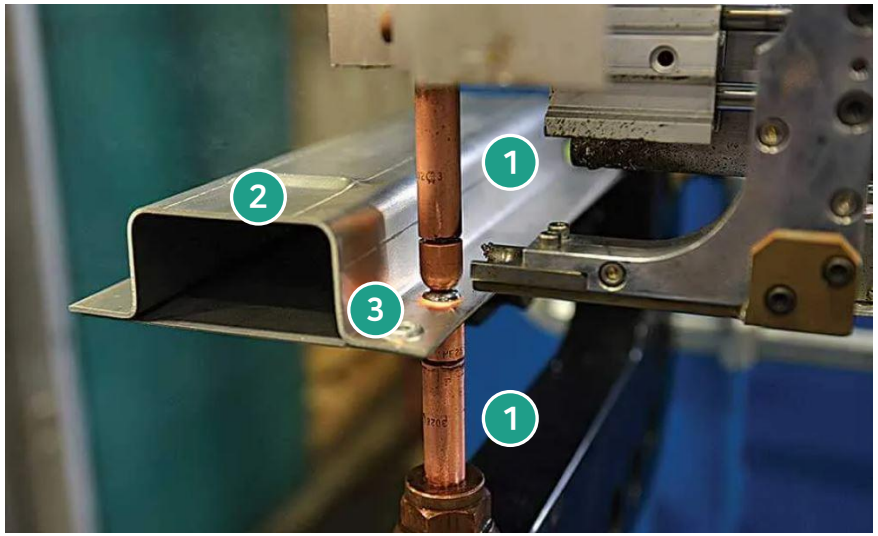
4 Rich preprocessing challenges
86.8% duplicates · missing labels · mismatched record units
→ Course cleaning & preprocessing techniques apply directly

What is “Spot welding?”



Spot welding joins two sheets pressed by electrodes via a melted-and-solidified nugget*

* where three types of defects can occur.



Components

1. **Electrode (copper-alloy tip)**
   Presses sheets and carries current
2. **Steel sheets**
   Two overlapped sheets (lap joint)
3. **Nugget**
   Joint melted and solidified between sheets

Welding Process

Squeeze ▶ **Weld** ▶ Hold

Why the Nugget Matters

- Nugget size determines weld strength
- 3,400–3,800 welds per car are spot welds
- the nugget forms inside the sheets(invisible)
- Defects easily slip past visual inspection

Defects: indentation · insufficient weld · crack

Ref. Good Weld	Type 1 Indentation	Type 2 Insufficient Weld	Type 3 Crack
			
Enough nugget + shallow mark Criterion: nugget dia. ≥ ~4.2 mm	Electrode pressed too deep, dented Thinner sheet, lower strength 14 cases in 8 days	Too little heat, small/no nugget Looks joined, easy to miss 13 cases in 8 days	Cracks while nugget solidifies Fracture risk under vibration 12 cases in 8 days (3/31 missing)

Following the course pipeline, we focus most on exploration and preprocessing before this proposal.

Final Objective

Clean unlabeled welding logs and detect anomalous welds without supervision, to provide inspection targets and stable operating windows (control ranges)

01 Collection	02 Exploration	03 Preprocessing	04 Model	05 Application
Data was provided Trace collection path	Distributions · trends · link to defects	Diagnose data issues Clean · set rules	ML Model for unsupervised learning	Applicable simulation model for industry
Given & Limited	Proposal	Proposal	Final	Final

What We Did Before This Proposal

- **Collection:** traced how the raw and result sheets were made and why their record units differ
- **Exploration:** compared normal vs. abnormal days and linked the repeated force-anomaly segment to defects
- **Preprocessing:** diagnosed 9 data issues, built the 2,184-row modeling data, and set scaling · split rules

We traced how the provided data was made (inferred). Process vs. defect logs differ in record unit

Process

Welding machine — 1 unit (Spot-01) ▶ **PLC** — Current·voltage·force·time ▶ **MES** — Stored per product ▶ **Raw data** — 11,939 rows · per product

Quality label

Inspectors — Visual inspection ▶ **Inspection log** — Good/defect · type ▶ **Daily aggregation** — Counts by type ▶ **result** — 23 rows · day × type

Source Files & Usage

"Raw Data" Column diagnosis (11,939 rows × 10 cols)

Column	Example	Measured range	Unique	Treatment
idx	2	1 ~ 2,000	—	Used
~~Machine · part no.~~	~~Spot-01~~	~~Constant~~	~~1~~	~~Excluded~~
working time	03-24	03-24 ~ 04-07	9	Used
~~Thickness 1 · 2~~	~~0.7 / 0.7~~	~~Constant~~	~~1~~	~~Excluded~~
weld force	2.33	1.74 ~ 10.54 bar	176	Used
weld current	14.57	14.52 ~ 15.07 kA	43	Used
weld voltage	2.701	2.464 ~ 2.861 V	113	Used
weld time	72	70 ~ 73 ms	8	Used

"Result Sheet" diagnosis (23 rows × 6 cols)

Field	Values	Unique	Treatment
date	03-24 ~ 04-07	8	To working day
defect type	Type 1 · 2 · 3	3	Pivot to columns
defect count	0 ~ 3 per day × type	4	Used (daily label)
~~3/27 record~~	~~No row (1,650 welds)~~	~~—~~	~~Excluded~~
3/31 crack	Blank	—	Kept as missing
welds per day	669 ~ 2,000 (raw)	8	Joined from raw

The provided data has 7 data issues, and only the first 2,184 rows (18.3%) are free of repeated blocks

Uniqueness

1 **Repeated anomaly** — the same 426 rows on 4 working days
→ split by working day to prevent leakage

Accuracy · Consistency

4 **Wrong dates** — 2 rows marked 3/25, actually 3/26–27
→ logged as errors, not used

5 **Floating-point residue** — e.g., 14.570000000000002
→ compare with tolerance

Label Completeness

2 **Record unit mismatch** — process: product, defect: day
→ unsupervised detection without labels

3 **Missing labels** — no 3/27 record, 3/31 crack blank
→ keep as missing, not 0

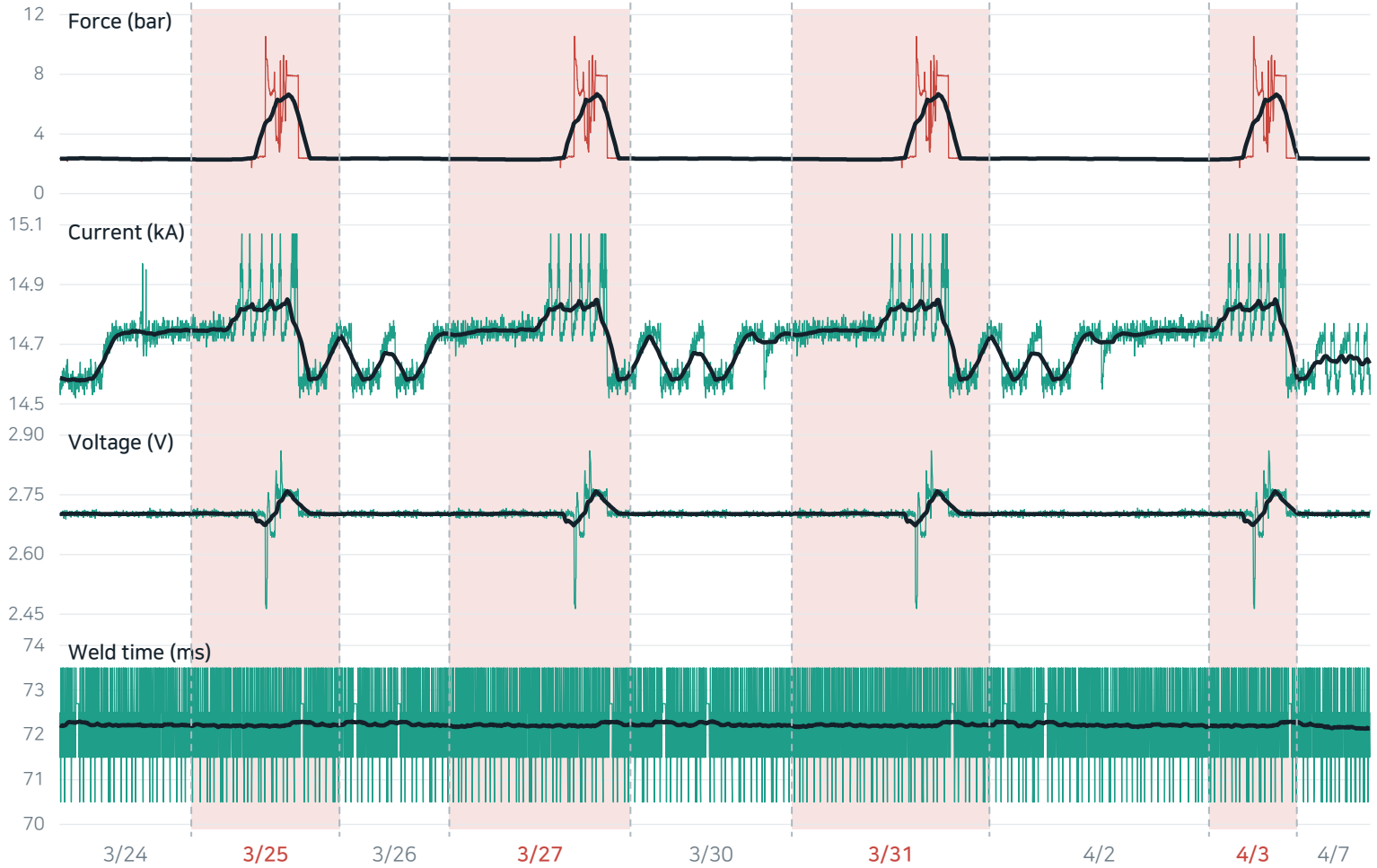
Relevance

6 **Constant columns** — machine · part no. · thickness 1·2
→ excluded from input (drop)

7 **No-signal variable** — weld time has only 8 unique values
→ kept in the table, excluded from model input

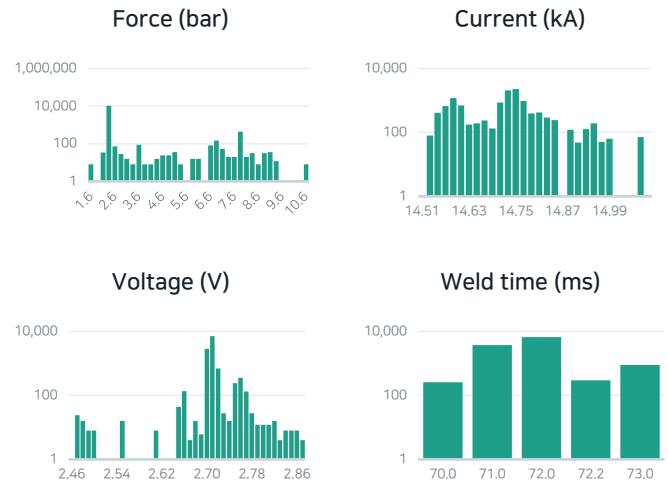
Force outliers cluster in one segment on 4 days (3/25·3/27·3/31·4/3), separating normal and abnormal days

① Force outliers mark 4 abnormal days



Process variables over production order (thin = measured, thick = 200-pt MA, red = abnormal)

Distributions (all 11,939, log y-axis)



Split criterion (force)

Rule Do values above the IQR upper fence (2.46 bar) cluster in one segment of a day?

Abnormal 3/25 · 3/27 · 3/31 · 4/3 (364 each)

Normal 3/24 · 3/26 · 3/30 · 4/2 · 4/7 (0)

The same anomaly segment repeats on 4 abnormal days → split by working day & process anomalies

② Same 426 rows repeat on 4 abnormal days

Anomaly segment comparison (all 4 variables)

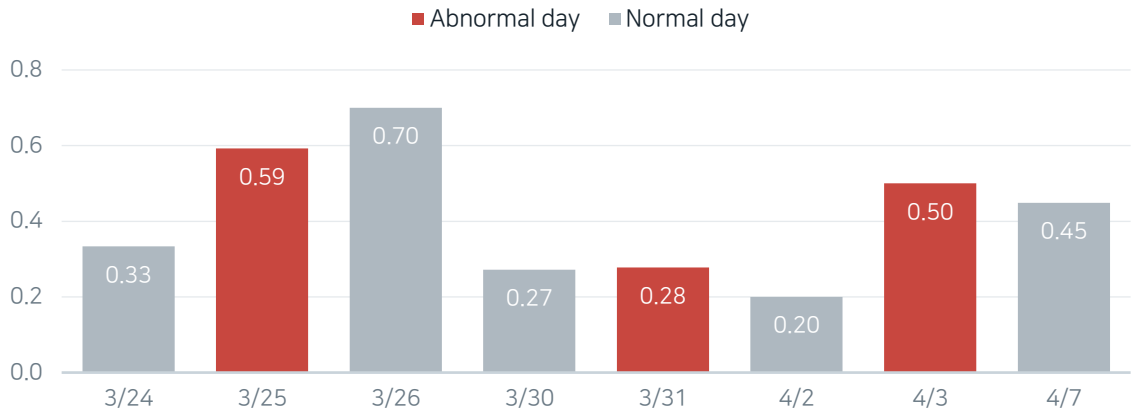
Date	Position (seq.)	Length	Diff. from 3/25
3/25	550 ~ 975	426	Reference
3/27	1,014 ~ 1,439	426	0
3/31	1,004 ~ 1,429	426	0
4/3	280 ~ 705	426	0

Force in the common segment — normal ~2.3 bar → up to 10.5 bar



③ Abnormal day ≠ defect

Daily defect rate (%) — red = abnormal day, 3/27 unlabeled



- Defect rate: abnormal 0.430% vs normal 0.347% → slightly higher
- Counterexample: 3/26 is normal but highest (0.70%)
- Only 8 days (3 abnormal vs 5 normal) → read as a trend

1 Split by Working Day · Validation Reference

Random row splits put identical values in train and test → overestimation
→ Split by day; use the repeated segment as a validation reference

2 Detection Target = Process Anomaly

No confirmed link between the segment and actual defects
→ Claim 'process anomaly detection', not 'defect prediction'

Modeling data is fixed at 2,184 rows · 3 variables; scaling and split rules are set for the final phase

Raw Excel 11,939 rows × 10 cols ▶ **Extract variables** 11,939 rows × 4 cols ▶ **Fix analysis range** Rows 0 ~ 2,183 ▶ **Model input** 2,184 rows × 3 cols

Modeling Data

Item	Column	Use
Input	Force · current · voltage	Model input
Aux. key	Working day · seq.	Split · trace back
Validation	Common segment (426 rows)	Validation reference
Excluded	Weld time · 4 constants	No signal

Why 2,184 Rows?
10-row window repeat rate: rows 0–2,183 0.4% vs after 99.2%
→ Later rows copy earlier blocks and add no new information

Preprocessing Rules

1 **Cleaning** done
Exclude copied blocks (after row 2,184)
Keep missing labels as missing, not 0

2 **Scaling** applied at training
Robust scaling (median · IQR) resists outliers
Fit on training block only → no test leakage

3 **Split** applied at training
The same anomaly segment is copied across days
→ Split train/test by working day, not random rows

Source: KAMP Welding Machine AI Dataset, Raw data sheet

Unsupervised anomaly detection: start from control charts, then extend step by step

Analysis data (2,184 rows · 3 vars) ▶ **Preprocessing** (Scaling · split) ▶ **Baseline** (Univariate chart) ▶ **Model extension** (Multivar. → time) ▶ **Evaluation** (Detection · false alarm) ▶ **Deliverable** (Targets · ranges)

1 Why Unsupervised

- No weld-level defect labels → supervised learning impossible
- Learn the normal process and score deviation as anomaly
- Welds above a threshold become inspection candidates

2 Model Extension

- Baseline: univariate control chart
- Extension ①: multivariate model of force · current · voltage
- Extension ②: temporal model using the trend of past welds
- Add one element at a time to isolate each effect

3 Evaluation Design

- Split train/test by working day (blocks leakage from duplicates)
- Reference: common anomaly segment (426 rows, 4 days)
- Metrics: segment detection rate + false alarms on unseen days
- Compare with the control chart under the same conditions

Scenario Overview

DATA LAYER
Build a dataset we can trust

ANALYSIS LAYER
Find conditions that differ from normal

APPLICATION
Deliver to the floor

01
Collect
KAMP welding dataset
Spot-level process values
Daily quality records

02
Validate
Detect duplicated segments
Data quality check
Leak-free train / test split

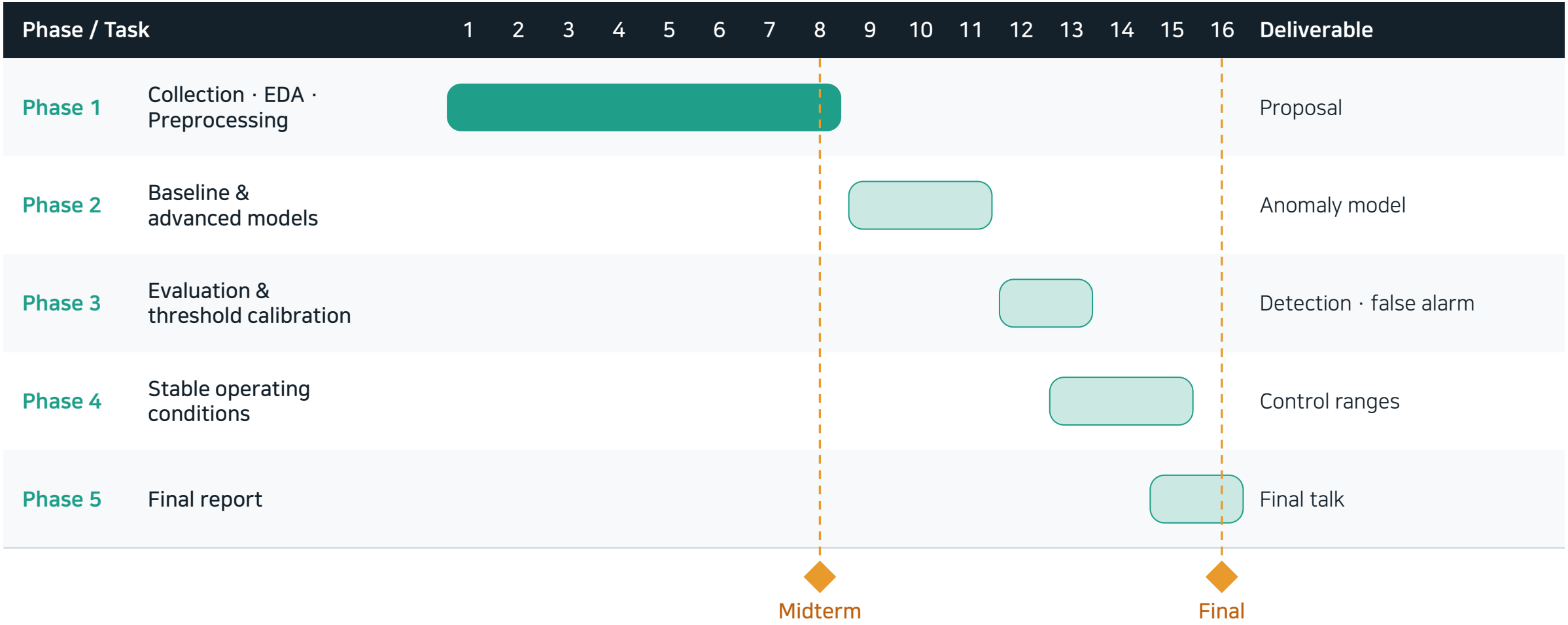
03
Detect
Learn normal process patterns
Variable & temporal relations
Spot-level anomaly score

04
Judge
Model result + process control rules
Output per weld spot
Normal
Suspected defect

05
Monitor
Virtual robot cell link
Product & weld-spot history
Role-based views & inspection requests

Future extension · inspection results → spot-level quality labels → model retraining & criteria recalibration

the final phase covers modeling, evaluation and stable operating windows



TO BE CONTINUE