

CS5344 Project Formal Problem Formulation

1 General Setting

Description. We first present the common anomaly-detection setup shared by Tasks 1–3, covering the dataset formulation, training/validation splits, learning objective, and evaluation protocol. We then describe the task-specific instantiations for each track.

Dataset. Let $\mathbf{D} = [x_{ij}] \in \mathcal{X}^n$ denote an $n \times m$ *mix-typed* table, where n is the number of samples (row number) and m is number of features (column number). The feature space \mathcal{X} is the product of typed column domains \mathcal{X}_j : $\mathcal{X} = \prod_{j=1}^m \mathcal{X}_j$. Each column j of \mathbf{D} corresponds to a feature with domain \mathcal{X}_j . Typical spaces include:

$$\mathcal{X}_j \in \{\mathbb{R} \text{ (numeric)}, \mathcal{C}_j \text{ (finite categorical set)}, \Sigma_j^* \text{ (strings)}, \dots\}.$$

Equivalently, each row i is a feature vector depicting the sample information:

$$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,m}) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_m.$$

Let $\mathbf{y} = (y_1, \dots, y_n)$ denote the label vector corresponding to the rows of \mathbf{D} (i.e., y_i is the label of \mathbf{x}_i). $y_i \in \{0, 1\}$, where $y_i = 1$ indicates abnormal and $y_i = 0$ indicates normal. We refer to (\mathbf{D}, \mathbf{y}) as the labeled dataset.

Training/validation splits. We split the labeled dataset (\mathbf{D}, \mathbf{y}) into training set $(\mathbf{D}_{\text{train}}, \mathbf{y}_{\text{train}})$ and validation set $(\mathbf{D}_{\text{valid}}, \mathbf{y}_{\text{valid}})$ by rows. Let $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{valid}} \subset \{1, \dots, n\}$ be disjoint index sets that partition the rows. The training split contains only normal samples, i.e., $y_i = 0$ for all $i \in \mathcal{I}_{\text{train}}$, while validation split is mixed:

$$\begin{aligned} \mathbf{D}_{\text{train}} &:= (\mathbf{x}_i)_{i \in \mathcal{I}_{\text{train}}}, & \mathbf{D}_{\text{valid}} &:= (\mathbf{x}_i)_{i \in \mathcal{I}_{\text{valid}}}, \\ \mathbf{y}_{\text{train}} &:= (y_i)_{i \in \mathcal{I}_{\text{train}}} = \mathbf{0}, & \mathbf{y}_{\text{valid}} &:= (y_i)_{i \in \mathcal{I}_{\text{valid}}}. \end{aligned}$$

Objective. The objective of this task is to learn an anomaly scoring function

$$f: \mathcal{X} \rightarrow [0, 1], \quad \hat{y}_i = f(\mathbf{x}_i) \text{ for } i = 1, 2, \dots, n.$$

so that \hat{y}_i closely approximates y_i , yielding low scores for normal samples and high scores for abnormal samples. Particularly, f should be trained only on $\mathbf{D}_{\text{train}}$ to model the distribution of normal samples.

Evaluation. To evaluate the results of f , we select threshold-free metrics to compare y_i and \hat{y}_i for $i \in \mathcal{I}_{\text{valid}}$. To be specific, we choose Average Precision as the primary evaluation metric, and AUC-ROC as secondary evaluation metric.

2 Track 1: Manufacturing

Description. We study trip-level anomaly detection for vehicle operations in this task. Each trip comprises (i) a static context vector describing the vehicle and environment, and (ii) an irregularly sampled, variable-length sequence of multi-sensor readings collected during the trip. The trip-level binary outcome label indicates whether a breakdown occurred by the end of the trip or not [2].

Models may jointly exploit the static context and the full temporal sensor readings to capture normal operating patterns and assign higher anomaly scores to trips that deviate from those patterns. Key challenges include cross-trip heterogeneity, mixed-type feature domains, variable sequence lengths, and the rarity and diversity of abnormal events.

Dataset. Each row of \mathbf{D} corresponds to a trip. For $i = 1, \dots, n$, the i -th trip is represented as

$$\mathbf{x}_i = (\mathbf{s}_i, (t_{i,1}, \mathbf{r}_{i,1}), (t_{i,2}, \mathbf{r}_{i,2}), \dots, (t_{i,T_i}, \mathbf{r}_{i,T_i}), y_i),$$

where

- \mathbf{s}_i is the static (nature) feature vector (e.g., vehicle type),
- $\{t_{i,k}\}_{k=1}^{T_i}$ are timestamps with $t_{i,1} < \dots < t_{i,T_i}$ (irregular sampling allowed),
- $T_i \in \mathbb{N}$ is the number of observations for trip i (varies across trips),
- $\mathbf{r}_{i,k}$ is the multi-sensor reading vector at time $t_{i,k}$ ($k = 1, \dots, T_i$).

Detailed explanation of each column j can be found in Appendix A.

The label corresponding to \mathbf{x}_i is $y_i \in \{0, 1\}$, with $y_i = 1$ indicating a breakdown by the end of the trip and $y_i = 0$ indicating normal operation.

Data Format. We release two logical tables aligned with our formulation:

- **Trips** (static features and labels): one row per trip i keyed by `index` which contains \mathbf{s}_i and y_i .
- **Records** (temporal sensor readings): multiple rows per trip i keyed by `index`, and row k stores a timestamp $t_{i,k}$ and a multi-sensor reading vector $\mathbf{r}_{i,k}$.

We further split these two tables into files based on training/validation split. These files are: `trips_train.csv`, `trips_valid.csv`, `records_train.csv`, `records_valid.csv`.

Mapping to notation. For a given split and trip i (identified by `index`):

$$\begin{aligned} \mathbf{s}_i, y_i &\leftarrow \text{trips_}\{\text{split}\}\text{.csv}[\text{index} = i], \\ \{(t_{i,k}, \mathbf{r}_{i,k})\}_{k=1}^{T_i} &\leftarrow \text{records_}\{\text{split}\}\text{.csv}[\text{index} = i] \text{ (sorted by } t). \end{aligned}$$

3 Track 2: Finance

Description. We study loan-level anomaly detection for repayment behavior. Each loan comprises (i) a static context vector describing the loan and borrower (e.g., borrower attributes), and (ii) a monthly sequence of performance variables (e.g., outstanding unpaid principal balance) observed over the life of the loan. The loan-level binary outcome indicates whether the loan ever fails to meet its monthly obligation or remains current throughout.

Models may jointly leverage the static underwriting information and the full repayment trajectory to characterize typical repayment patterns and assign higher anomaly scores to loans whose trajectories deviate from those patterns. Key challenges include borrower and product heterogeneity, mixed-type feature domains, and strong class imbalance due to the relative rarity of abnormal loans [1].

Dataset. Each row of \mathbf{D} corresponds to a loan. For $i = 1, \dots, n$, the i -th loan is represented as

$$\mathbf{x}_i = (\mathbf{s}_i, (t_{i,1}, \mathbf{r}_{i,1}), (t_{i,2}, \mathbf{r}_{i,2}), \dots, (t_{i,T_i}, \mathbf{r}_{i,T_i}), y_i),$$

where

- \mathbf{s}_i is the static loan information (e.g., borrower attributes),
- $\{t_{i,k}\}_{k=1}^{T_i}$ are months with $t_{i,1} < \dots < t_{i,T_i}$,
- $T_i \in \mathbb{N}$ is the number of months for loan i ,
- $\mathbf{r}_{i,k}$ is the monthly repayment information vector at month $t_{i,k}$ ($k = 1, \dots, T_i$).

Detailed explanation of each column j can be found in Appendix B.

The label corresponding to \mathbf{x}_i is $y_i \in \{0, 1\}$, where $y_i = 1$ denotes an abnormal loan that misses at least one scheduled monthly payment, and $y_i = 0$ denotes a normal loan.

Data Format. We release one logical table aligned with our formulation:

- **Loans** (static features and labels): one row per loan i keyed by `index` which contains \mathbf{x}_i and y_i .

We further split the table into files based on training/validation split. These files are: `loans_train.csv`, `loans_valid.csv`.

4 Track 3: Cybersecurity

Description. We study process-level anomaly detection for server operations. Each example corresponds to a single process observed on a host and comprises an information vector (e.g., timepoint, arguments, return value). The binary outcome indicates whether the process is malicious or benign.

Although each process is represented by one timepoint (i.e., no per-process variable-length sequence), processes on the same host collectively form a time-ordered stream. Models may optionally leverage this inter-process temporal relationship to improve detection. Key challenges include high-cardinality categorical fields, mixed-type feature domains, class imbalance, and temporal correlation among related processes [3].

Dataset. Each row of \mathbf{D} corresponds to a single process. For $i = 1, \dots, n$, the i -th process is represented as

$$\mathbf{x}_i = (t_i, \mathbf{p}_i),$$

where

t_i is the timepoint at which the process observation is recorded,

\mathbf{p}_i is the process information vector (e.g., process name, arguments, return value).

Detailed explanation of each column j can be found in Appendix C.

The label corresponding to \mathbf{x}_i is $y_i \in \{0, 1\}$, where $y_i = 1$ denotes a malicious (abnormal) process and $y_i = 0$ denotes a benign (normal) process.

Data Format. We release one logical table aligned with our formulation:

- **Processes** : one row per process i keyed by `index` which contains \mathbf{x}_i and y_i .

We further split the table into files based on training/validation split. These files are: `processes_train.csv`, `processes_valid.csv`.

5 Track 4 (Special Track): Healthcare – Synthetic Data Generation

Description. This special track departs from the anomaly-detection formulation of Tasks 1–3 and instead focuses on *synthetic longitudinal data generation* for healthcare. Participants are asked to design models that can generate plausible future clinical trajectories given partial patient histories. Unlike the other tracks, this task does not feature a Kaggle leaderboard: final evaluation will be conducted by the instructors and teaching assistants based on the quality of the problem formulation, methodological soundness, and empirical study.

Dataset. We use the Parkinson’s Progression Markers Initiative (PPMI) dataset (Appendix D), which contains longitudinal clinical and biomarker data from more than 1,500 participants, including both Parkinson’s disease patients and controls. Each patient i contributes a sequence of m_i visits at irregularly spaced timestamps, with mixed-type feature vectors covering demographics, clinical assessments, biomarkers, and imaging variables [4].

Task Formulation. For each patient i , let $\{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,m_i}\}$ denote the sequence of visit records, aligned with timestamps $\{t_{i,1}, \dots, t_{i,m_i}\}$. In this task, models are given the first k_i visits $\{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k_i}\}$ and the timestamps of the remaining $m_i - k_i$ visits. The objective is to generate synthetic records $\{\hat{\mathbf{x}}_{i,k_i+1}, \dots, \hat{\mathbf{x}}_{i,m_i}\}$ that resemble the true withheld visits.

Evaluation. Unlike Tasks 1–3, evaluation in this track is not based on anomaly scores but on generative quality. Multiple complementary criteria may be considered:

- **Statistical fidelity:** similarity between synthetic and ground-truth records (e.g., Gower’s distance, KL divergence, MMD).
- **Predictive utility:** performance of downstream models trained on synthetic data and tested on real data.

- **Interpretability and controllability:** ability to explain or steer generated trajectories by clinically meaningful covariates.
- **Baselines:** comparison against strong baselines such as per-patient extrapolation, feature imputation, VAEs, and GANs.

Remarks. Because this task is open-ended and research-oriented, teams must apply in advance for approval to join. Well-executed projects in this track may achieve higher recognition and scores than those in standard tracks; however, underdeveloped submissions may score lower than those in standard tracks.

References

- [1] Freddie Mac. Freddie mac single-family loan-level dataset, 2019. Available at <https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset>.
- [2] Silas Hellenbrand, Dominik Schmid, Simon Albin, Martin Törngren, Zaid Al-Ars, et al. Scania component x dataset for vehicle component failure prediction. *Scientific Data*, 12(1):122, 2025.
- [3] Kate Highnam. Beth honeypot dataset, 2021. Available on Kaggle: <https://www.kaggle.com/katehighnam/beth-dataset>.
- [4] Parkinson’s Progression Markers Initiative. Parkinson’s progression markers initiative (ppmi), 2011. Available at <https://www.ppmi-info.org>.

A Column Names and Descriptions of the SCANIA Component X Dataset

`index` Unique anonymized identifier for each trip in the fleet.

`target` Binary label indicating whether the trip experienced a breakdown:

- 0 = no breakdown
- 1 = breakdown occurred

`Spec_0`, `Spec_1`, `Spec_2`, `Spec_3`, `Spec_4`, `Spec_5`, `Spec_6`, `Spec_7` Eight anonymized categorical specification features describing static vehicle properties (e.g., truck type, configuration). Each takes discrete values labeled as `Cat0`, `Cat1`, ..., up to `Cat28`.

`time_step` Continuous variable measuring the elapsed operational time of Component X for a given vehicle; serves as a timeline index.

- Different vehicles may have different sampling frequencies.
- The dataset provides a curated subset of operational records selected by experts to capture the most relevant operating conditions.

Operational sensor features (all remaining columns) Fourteen anonymized operational variables. Some are recorded as *numeric counters* (single values), while others are *histograms* (multi-bin columns).

- **Numeric counters (single-value variables):**

- 171_0, 666_0, 427_0, 837_0, 309_0, 835_0, 370_0, 100_0

Each of these columns is a continuous numeric measurement (e.g., cumulative counters or direct sensor values).

- **Histogram variables (multi-bin variables):**

- 167_0 ... 167_9 (10 bins)
- 272_0 ... 272_9 (10 bins)
- 291_0 ... 291_10 (11 bins)
- 158_0 ... 158_9 (10 bins)
- 459_0 ... 459_19 (20 bins)
- 397_0 ... 397_35 (36 bins)

Each histogram variable is represented by multiple columns, where the suffix `_binindex` indicates the bin number.

- **Binning concept:**

- Continuous sensor readings are divided into intervals (bins).
- For each observation window, the dataset records how many values fall into each bin.

- Example: If variable `167` represented temperature, then `167_0` could store counts of values below -20°C , `167_1` counts between -20°C and 0°C , etc.
- **Interpretation:**
 - A single row corresponds to one vehicle at one `time_step`.
 - Numeric counters show the sensor’s numeric value directly.
 - Histogram variables describe the distribution of sensor values over that time interval, capturing richer operating patterns than a single reading.

B Column Names and Descriptions of the Freddie Mac Single-Family Loan-Level Dataset

`index` Unique identifier assigned to each loan.

`target` Binary label indicating loan performance outcome (for this project setting):

- 0 = normal loan (no default).
- 1 = abnormal loan (default or anomalous event),

Origination Variables

`CreditScore` Borrower credit score at origination (300–850). Values outside range or missing coded as 9999.

`FirstPaymentDate` First scheduled payment due date (YYYYMM).

`FirstTimeHomebuyerFlag` Y = Yes, N = No, 9 = Not Available.

`MaturityDate` Scheduled maturity date (YYYYMM).

`MSA` Metropolitan Statistical Area code (null if unknown).

`MI_Pct` Mortgage insurance percentage. 0 = none, 1–55 = coverage %, 999 = not available.

`NumberOfUnits` Number of dwelling units (1–4).

`OccupancyStatus` P = Primary, I = Investment, S = Second Home, 9 = Not Available.

`OriginalCLTV` Combined Loan-to-Value ratio at origination.

`OriginalDTI` Debt-to-Income ratio (%). Values $> 65\%$ or missing coded as 999.

`OriginalUPB` Original unpaid principal balance (nearest \$1,000).

`OriginalLTV` Loan-to-Value ratio at origination; invalid coded as 999.

`OriginalInterestRate` Note rate at origination.

Channel Origination channel: **R** = Retail, **B** = Broker, **C** = Correspondent, **T** = TPO Not Specified, **9** = Not Available.

PPM_Flag Prepayment penalty: **Y** = Yes, **N** = No.

ProductType **FRM** = Fixed Rate, **ARM** = Adjustable Rate.

PropertyState Two-letter state/territory code.

PropertyType **SF** = Single-Family, **CO** = Condo, **PU** = PUD, **MH** = Manufactured, **CP** = Co-op, **99** = Not Available.

PostalCode Masked ZIP code (first 3 digits + “00”).

LoanPurpose **P** = Purchase, **C** = Refinance Cash Out, **N** = Refinance No Cash Out, **R** = Refinance Not Specified, **9** = Not Available.

OriginalLoanTerm Scheduled term in months.

NumberOfBorrowers Number of borrowers (1–10).

SellerName Entity that sold the loan (“Other Sellers” if below disclosure threshold).

ServicerName Entity servicing the loan (“Other Servicers” if below disclosure threshold).

SuperConformingFlag Indicates whether loan exceeded conforming limits but qualified as “super conforming”.

PreHARP_Flag / ProgramIndicator / ReliefRefinanceIndicator Indicators for HARP and related refinance programs.

PropertyValMethod Appraisal method: **1** = ACE, **2** = Full, **3** = Other (Desktop/AVM), **4** = ACE + PDR.

InterestOnlyFlag **Y** = interest-only payments required, else **N**.

BalloonIndicator **Y** = balloon payment, else **N**.

Performance Panel Variables

For each loan, monthly performance data is provided across multiple periods. The prefix **N_** indicates the month index, where $N = 0, 1, 2, \dots$. Each panel contains the following repeated fields:

N_CurrentActualUPB Current unpaid principal balance (UPB), including both interest-bearing and non-interest-bearing portions.

N_CurrentInterestRate Mortgage interest rate in effect for that period.

N_CurrentNonInterestBearingUPB Non-interest-bearing portion of UPB (e.g., deferred modification amounts).

N_EstimatedLTV Current estimated Loan-to-Value ratio (ELTV) from Freddie Mac’s AVM.
Range: 1–998, with 999 = unknown.

N_InterestBearingUPB Portion of UPB that accrues interest.

N_LoanAge Number of months since the loan’s first payment date (or modification date).

N_MonthlyReportingPeriod Period identifier in YYYYMM format.

N_RemainingMonthsToLegalMaturity Remaining months until scheduled maturity (adjusted if modified).

Notes

- The origination variables provide static background (borrower credit, loan terms, property information).
- The performance panel makes this a longitudinal dataset: each loan is tracked monthly until payoff, maturity, or default.
- For further detail, see the official Freddie Mac user guide.

C Column Names and Descriptions of the BETH Honey-pot Logs Dataset

Your file is a *process-event table*. Each row corresponds to one kernel event on a host with decoded arguments. Column meanings:

index (int) Unique identifier assigned to each process.

target (int / bool) Event-level ground truth label provided by the BETH authors: 0 = benign, 1 = malicious.

timestamp (float) Monotonic time (seconds) from the experiment start when the event was captured. Use to order events and build sequences (per host or per process).

processId (int) OS process ID (PID) of the process that triggered the event.

threadId (int) Thread ID (TID); equals PID for single-threaded processes.

parentProcessId (int) Parent PID (PPID) of **processId** at event time. Useful for process-tree reconstruction and lineage features.

userId (int) Numeric user identifier (e.g., 0 = root). Helpful to detect privilege escalation.

mountNamespace (int) Linux mount namespace inode ID; distinguishes containerized contexts and isolates processes sharing the same filesystem view.

processName (string) Short executable name of the emitting process (e.g., **systemd**, **ps**).

hostName (string) Honeypot host identifier (e.g., `ip-10-100-1-28`). Group or split data by host when training/validating. Dataset reports **23** distinct hosts.

eventId (int) Numeric code for the kernel event (stable ID used by the BETH pipeline).

eventName (string) Human-readable syscall or kernel activity name (e.g., `close`, `stat`). BETH’s process subset records kernel-level process calls harvested via eBPF.

stackAddresses (array[int]) Raw kernel/user stack return addresses captured at event time. Helps disambiguate calling contexts; may be empty for some events.

argsNum (int) Count of decoded arguments present in **args**.

returnValue (int) System-call return value (e.g., 0 for success, negative `errno` for failure). Error patterns can be indicative of probing.

args (array[object]) Structured list of event arguments. Each element typically has `{"name", "type", "value"}`, e.g., `{'name': 'pathname', 'type': 'const char*', 'value': '/proc/25'}` for a `stat` call. Use to craft semantic features such as file paths touched, FDs used, flags, etc.

BETH provides fully labeled activity for benchmarking.

Notes

- **Units & ordering:** sort by (`hostName`, `timestamp`) to build per-host timelines; for process-centric analysis, sort by (`hostName`, `processId`, `timestamp`).
- **Context features:** derive parent-child chains from `parentProcessId`, privilege from `userId`, container scope from `mountNamespace`, and behavioral n-grams from (`eventName`, `args`).

D Column Names and Descriptions of the PPMI Parkinson’s Progression Dataset

The Parkinson’s Progression Markers Initiative (PPMI) dataset is a longitudinal clinical study tracking Parkinson’s disease (PD) patients, prodromal individuals, and healthy controls. It provides multi-modal measurements collected over multiple years. Below we summarize its key variable groups.

Identifiers and Demographics

PATNO Unique anonymized patient identifier.

EVENT_ID Visit identifier (e.g., BL = baseline, V01, V02, ...).

AGE Age at baseline (years).

SEX Biological sex (M/F).

APPRDX Diagnosis at enrollment (e.g., PD, healthy control, SWEDD).

Clinical Assessments

MDS_UPDRS_I--IV Unified Parkinson's Disease Rating Scale subscores (non-motor, motor, daily living, and motor complications).

MOCA Montreal Cognitive Assessment score.

GDS Geriatric Depression Scale.

SCOPA_AUT Autonomic dysfunction questionnaire.

Biomarkers

CSF_ABETA, CSF_TAU, CSF_P_TAU, CSF_ALPHA_SYNUC Cerebrospinal fluid (CSF) protein levels.

GENOTYPE Selected genetic variants and SNPs.

Imaging Variables

DATSCAN_PUTAMEN, DATSCAN_CAUDATE Dopamine transporter imaging values.

MRI_VOL MRI-derived brain volumetrics (subset of participants).

Progression and Outcomes

H_Y_STAGE Hoehn and Yahr stage (disease severity).

CLINICAL_EVENTS Milestone flags (e.g., initiation of PD therapy, complications, withdrawal).

Notes

- Each patient has a sequence of visits $\{\mathbf{x}_{i,t}\}_{t=1}^{m_i}$ with mixed-type attributes (continuous, categorical, ordinal).
- Missingness is common and non-random, reflecting real-world clinical data.
- Data access requires registration at <https://www.ppmi-info.org>.