SAMSUNG

FailureMiner: A Joint Key Decision Mining Scheme for Practical SSD Failure Prediction and Analysis

Shuyang Wang[1], Yuqi Zhang[1], Haonan Luo[1], Kangkang Liu[2], Gil Kim[3], JongSung Na[3], Claude Kim[3], Geunrok Oh[3], Kyle Choi[3], Ni Xue[1], Xing He[1]

[1]*Samsung R&D Institute China Xi'an, Samsung Electronics*
[2]*Tencent*
[3]*Samsung Electronics*

2026 SRCX

Introduction

➢ **Impact of SSD failure**

In enterprise storage systems and large data centers

Data loss

Maintenance cost

Service unavailability

?

What should we do?

Reset count	28
Power on hour	376
Temperature	41
...	

SSD daily monitoring data

\+

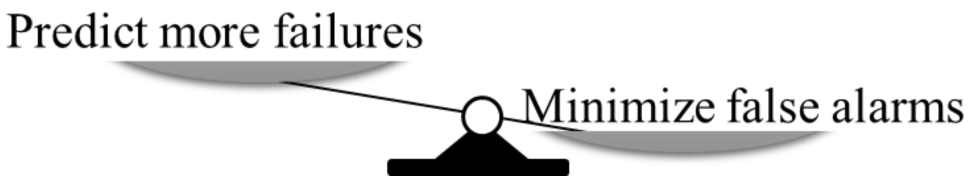
Machine learning

➡ **Predict failures and analyze their root causes**

Introduction

➤ Design goals for SSD failure prediction in production

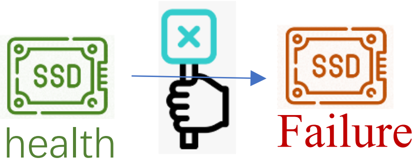
- High-accuracy



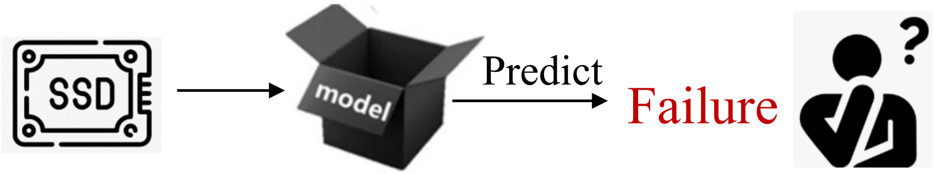
Hundreds of false alarms

↓

Impose substantial operational overhead



- Interpretable



Black box model

Why did it fail?

Lack confidence and necessary information to take actions

- User-friendly

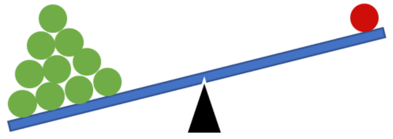
Prediction model → Streamlined and efficient

Prediction model → Cost-effective in deployment

Introduction

➢ Existing work

(1) Data preprocessing

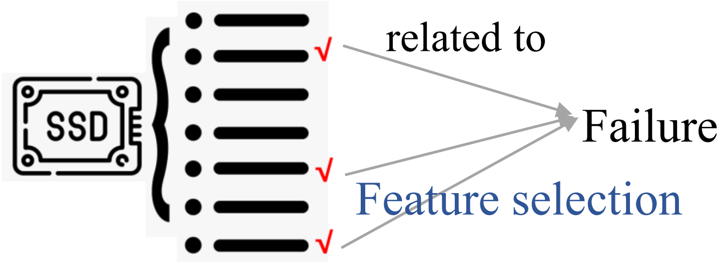


Data imbalance in training data → Low accuracy 

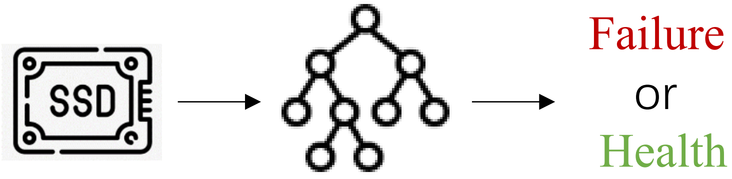
- Healthy samples: Failed samples ⟶ hundreds : 1
  - ↳ Downsampling

(2) Model building

A lot of monitoring attributes (noisy) 



Random Forest (RF)



binary classification models

(3) Model interpretability-based analysis

Failure-related global attributes	SHAP for a single failure	SDE to capture important decisions (i.e., bad NAND block > 1)
✖ No insight for a single failure	SHapley Additive exPlanations (SHAP*)	Similar decision extraction (SDE) in MVTRF*

SHAP: A unified approach to interpreting model predictions. NIPS'17
MVTRF: Multi-view feature-based SSD failure prediction: What, when, and why FAST23

Attempts and lessons

➢ Attempt 1: data balancing

- Downsample healthy samples (Fewer health)
- More true alarms (TA), but unacceptable high false alarms (FA)

Table 1. TA and FA of RF with downsampling are x-times higher than those without downsampling.

The ratio of healthy to failed samples after downsampling	Random downsampling TA	Random downsampling FA	Representative downsampling TA	Representative downsampling FA
100 :1	2.0x	5.4x	1.8x	4.2x
10 :1	3.1x	33.5x	3.2x	25.9x
1 :1	4.3x	210.3x	4.1x	203.3x

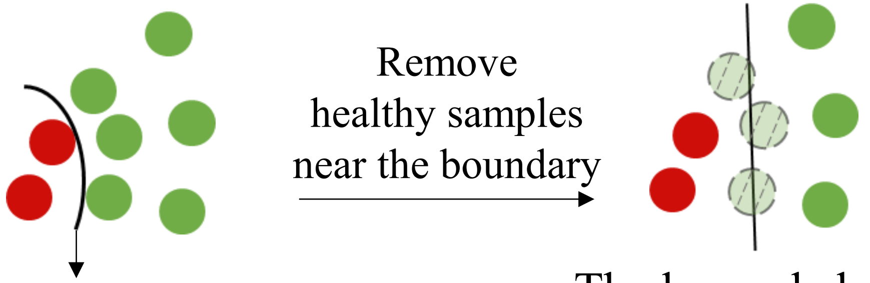
Healthy samples decreases

Correctly predict more failed SSDs (TA)

Misclassify more healthy SSDs as failed ones (FA)

We found

- These methods overlook key healthy samples near classification boundaries.
- Blurred classification boundaries lead to more misclassifications.



Classification boundary

The learned classification boundary is inaccurate

➢ Lessons:

- Samples near the classification boundary are critical for ensuring the accuracy and robustness of classification models

Randomly downsample: SSD failures in the field: symptoms, causes, and prediction models. SC '19
Representative downsampling: Predicting disk replacement towards reliable data centers. KDD '16

Attempts and lessons

➢ Attempt 2: model building

- Feature selection (WEFR*) reduces noisy data to lower FA, but may also remove useful data and lead to fewer TA

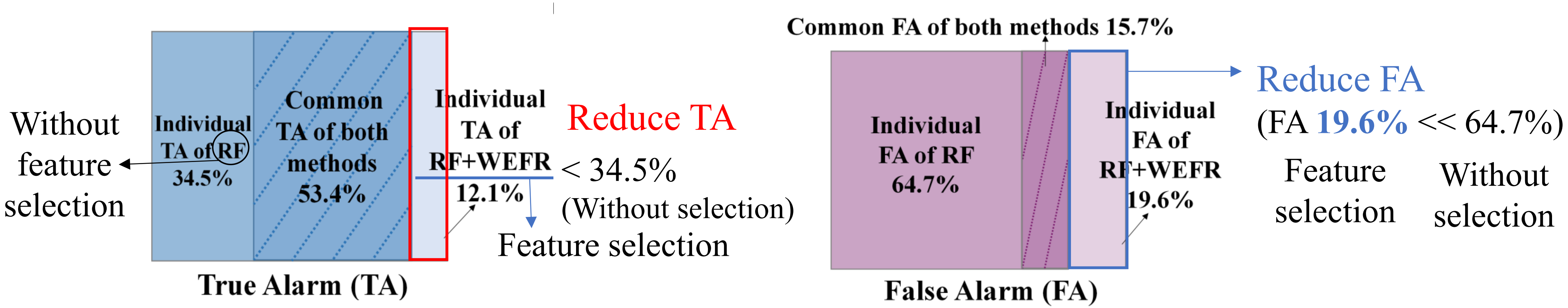


Figure: Proportion of true/false alarms from RF or RF+WEFR to the union of true/false alarms from both methods

➢ Lessons:

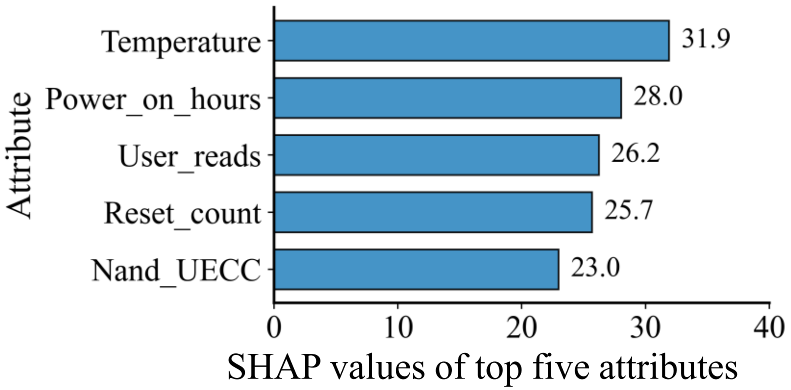
- Some attributes may not be directly associated with failures, but they can play an auxiliary role in failure prediction when combined with other attributes
- Feature selection at the attribute level is coarse-grained and may inadvertently remove auxiliary attributes

WEFR: General feature selection for failure prediction in large-scale SSD deployment. DSN

Attempts and lessons

➢ Attempt 3: **model interpretability-based analysis**

- SHAP values show the importance of attributes about failure, but **attribute-level analysis is coarse-grained**.
- SDE extracts frequent decisions to identify SSD's abnormal attributes and values, but **these decisions may not be important, and decisions' interrelationships remain unexplored**.



Higher SHAP values = More important attributes

SDE focuses on individual decisions

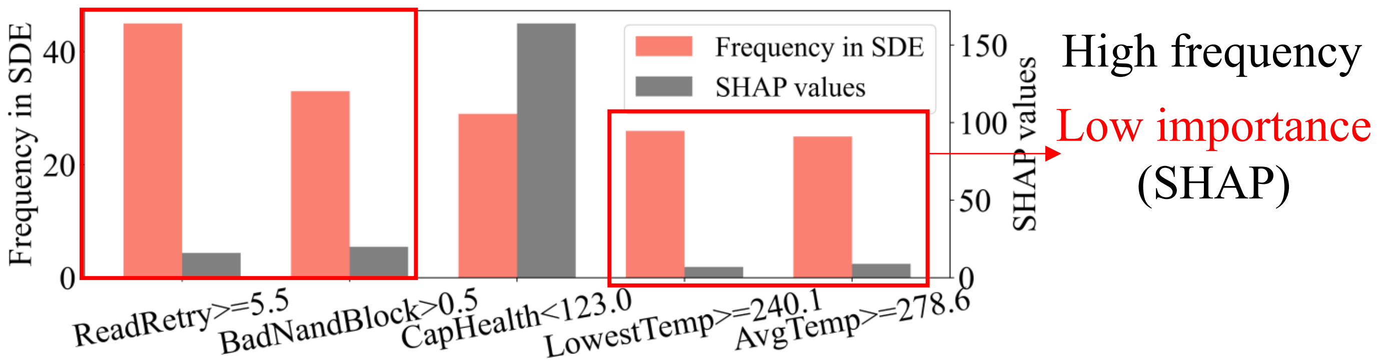


Figure: SHAP values of top five decisions extracted by SDE in identifying a failed SSD.

➢ Lessons:

- Extracting important decisions to locate failure causes is a fine-grained manner
- Some failure patterns are reflected in abnormal changes across multiple attributes. We should focus on the joint contribution of multiple decisions rather than individual ones.

Boundary-preserving downsampling

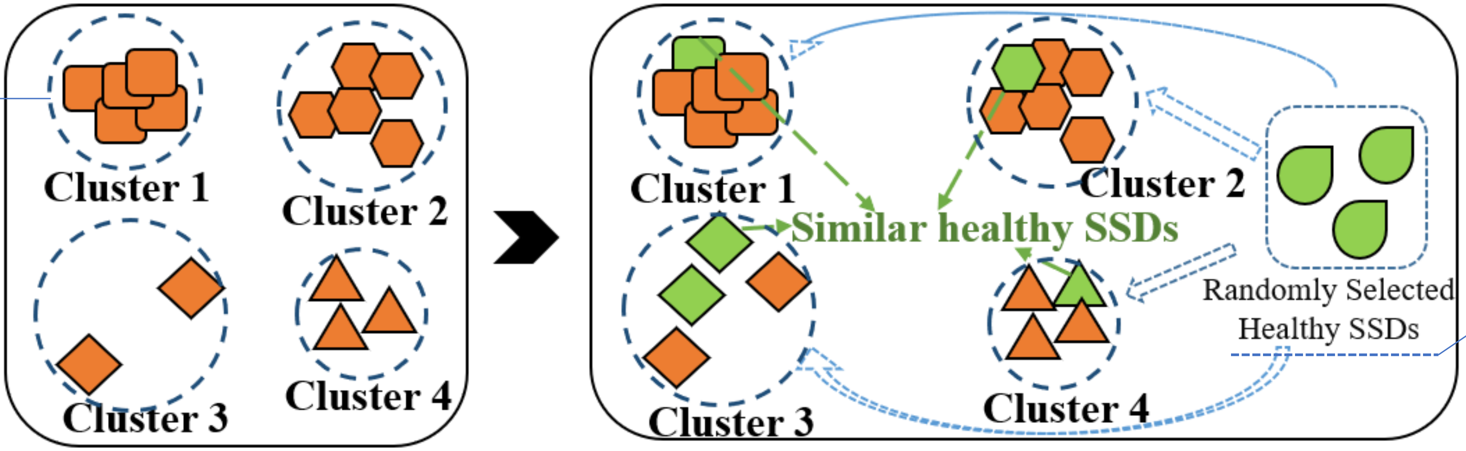
➢ **Lessons:** Samples near the classification boundary are critical

➢ **Ideas:**

✓ Select only near-boundary healthy samples in each cluster to balance data while preserving critical data

- Various failure patterns. ⟶ The data of failed SSDs are partitioned into N clusters using K-means
- Classification boundary ⟶ Data samples near the classification boundary exhibit similar patterns. We use the cluster boundary of each cluster to find nearby healthy SSD with similar patterns.

Cluster boundary: Maximum Euclidean distance from center to any failed SSD in cluster



b) Failed SSD clustering c) Healthy SSD dividing

To avoid overfitting caused by insufficient data

Those healthy SSD data (closer to the cluster center than cluster boundary) will be selected into the cluster.

Joint contribution-based key decision set extraction

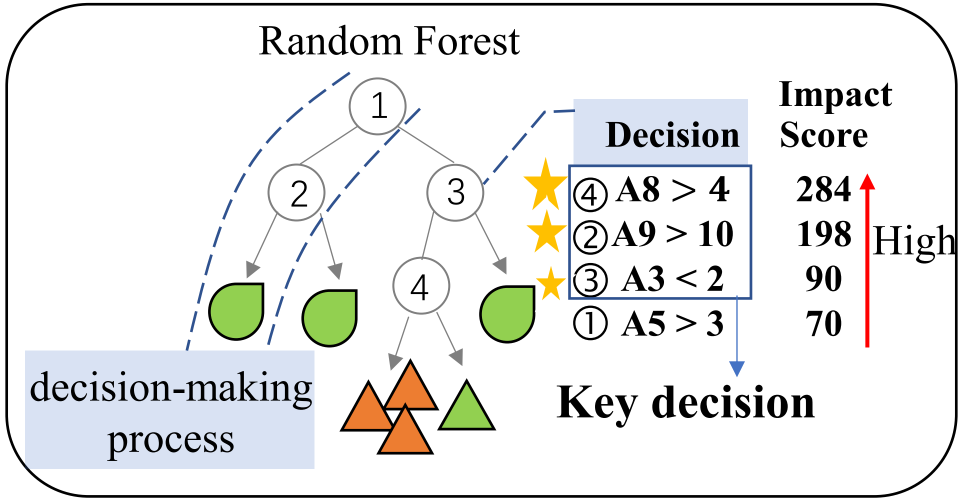
- **Lessons:** Decision-level are finer-grained than attribute-level for both noise removal and model interpretation
- **Idea:**

✓ Decision-level selection: finer-grained than attribute selection—removes noisy decisions, even on important attributes

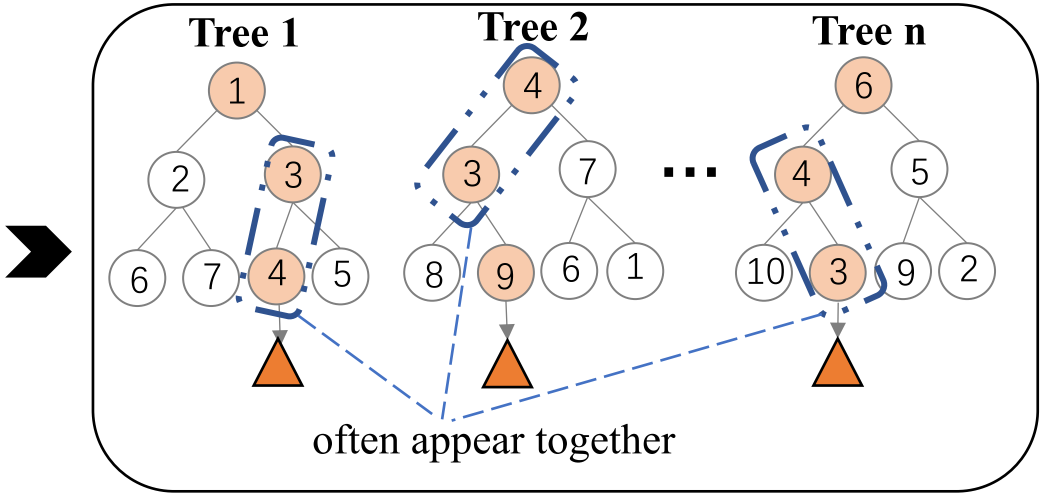
- Decision-level → We extract the impact score (decision-based SHAP value) of each decision to find key decision

✓ Joint decisions: even "minor" attributes can contribute when combined

- Joint contribution of multiple decisions → We extract joint key decisions, which often appear together in the decision-making process and jointly make a major contribution to the final failure identification.



Key decision Mining

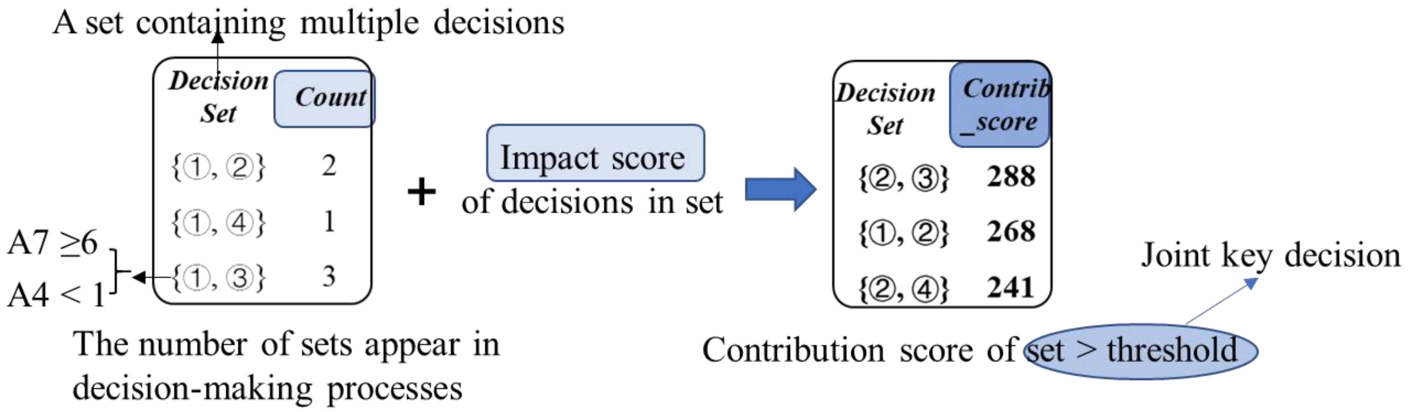


joint key decision extraction

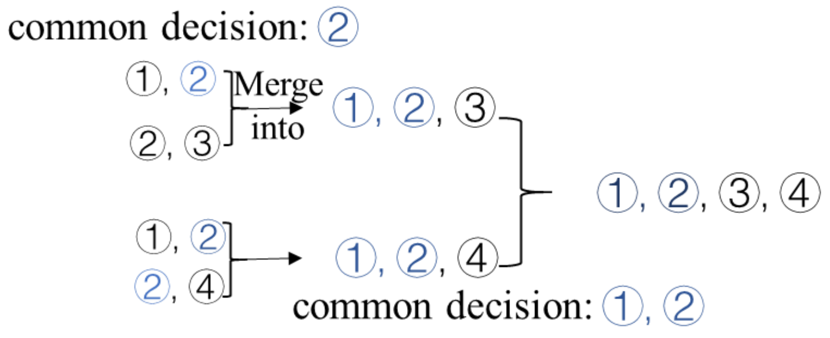
Joint contribution-based key decision set extraction

➢ Extracting joint key decisions

- We calculate the contribution score by combining decision importance and co-occurrence frequency of decision sets to extract joint key decisions that are truly meaningful for failure identification.



- We expand frequent decision sets from small ones to large ones based on their common decisions, aiming to find more interrelated decisions to form a lager decision set



Merging common decisions to form more joint decisions

Evaluation

➢ Dataset setup

- **Tencent: 70 million Telemetry logs (85 attributes) from about 350,000 Samsung SSDs**
- Alibaba: 10 million SMART logs (21 attributes) from 20,000 SSDs

Time span (month) of training set and test set.

Dataset	Training time	Test time
Tencent	1st - 13th	14th - 23rd
Alibaba	1st - 12th	13th - 24th

➢ Four comparison models

- RF : original random forest model
- WEFR : random forest model combined with feature selection
- MVTRF: random forest model combined with temporal features to capture abnormal changes in SSD monitoring attributes
- CNN−LSTM*: Neural network model (Convolutional Neural Network + Long Short-Term Memory)

➢ Three metrics

- Precision (P): the proportion of true alarms to all true alarms and false alarms
- Recall (R): the proportion of true alarms to all actual failed SSDs
- F0.5-score (F0.5): Harmonic average of P and R $\frac{(1+0.5^2)*P*R}{0.5^2*P+R}$

Weighted higher, reflecting the greater impact of false alarms

* CNN−LSTM : Making disk failure predictions SMARTer! FAST’20

Evaluation

➢ Evaluation on failure prediction

On Tencent dataset

Three failure patterns ← (first three rows)

Joint key decisions	Test P	Test R	Test F0.5
(1-1. $Delta_{15}$NandUECC ⩾ 10) or (1-2. $Delta_{15}$NandUECC ⩾ 1 & 1-3. $Delta_{3}$UserRead < 36)	81.9%	23.6%	0.55
2-1. DramCECC ⩾ 13 & 2-2. $Delta_{7}$DramCECC-Add ⩾ 2	72.7%	3.2%	0.14
3-1. CapHealth <68 or (3-2. CapHealth <130 & 3-3. $Delta_{7}$CapHealth < −15)	100.0%	3.2%	0.14
FailureMiner (Ours)	**82.2%**	**29.6%**	**0.61**
RF [6]	55.4%	20.4%	0.41
CNN-LSTM [22]	50.0%	12.8%	0.32
WEFR [12]	67.9%	15.2%	0.40
MVTRF [13]	68.1%	19.6%	0.46

On Alibaba dataset

Key decision	Test P	Test R	Test F0.5
4-1. r198 ⩾ 0.5 & 4-2. r174 < 32.5	65.2%	19.1%	0.44
5-1. r1 ⩾ 0.5 & 5-2. r187⩾ 0.5	78.6%	10.9%	0.35
FailureMiner (Ours)	**68.4%**	**26.4%**	**0.52**
RF [6]	32.7%	16.2%	0.27
CNN-LSTM [22]	32.3%	10.2%	0.23
WEFR [12]	56.8%	17.8%	0.40
MVTRF [13]	59.6%	21.5%	0.44

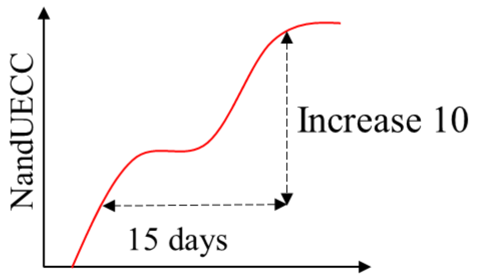
Two datasets have different monitoring attributes, SSD models and enterprises

- Effectiveness: random forest (117,404 decisions) → 3 joint key decisions (streamlined) — deployment-friendly and well-suited for online prediction
- Failure prediction capability: improving Precision (P) by 38.6% and Recall (R) by 80.5% on average (Tencent dataset)
- Generalizability: also maintaining high P and R on the Alibaba SMART dataset

Failure analysis

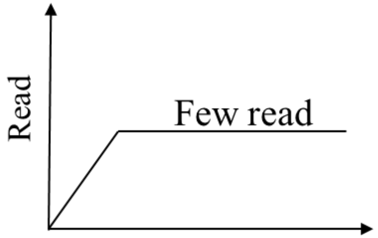
➢ Failure patterns (UECC-related, DRAM-related, capacitor-related)

- UECC-related



$Delta_{15}$NandUECC ≥ 10

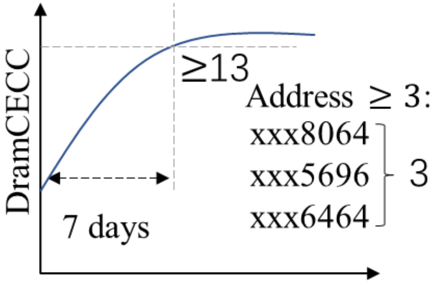
Pattern 1: Rapid increase of UECC errors on NAND



$Delta_{15}$NandUECC ≥ 1
$Delta_{15}$UserRead < 36

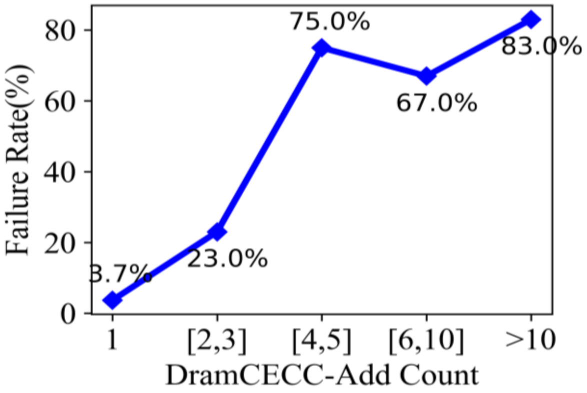
Pattern 2: A few UECC errors, few Read/Write operations

- DRAM-related



DramCECC ≥ 13
$Delta_7$DramCECC –Add ≥ 2

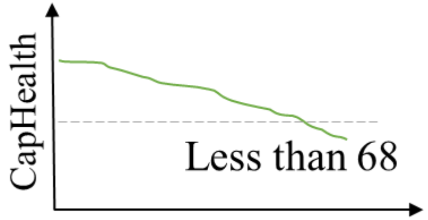
Pattern: Some CECC errors on DRAM, with multiple distinct error addresses



Analysis: DRAM CECC errors across more addresses → Failure rate ↑

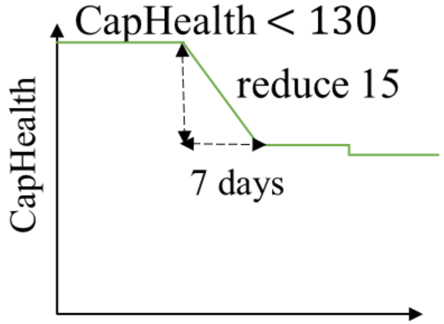
- Capacitor-related

Healthy SSD is around 160



CapHealth < 68

Pattern1: Capacitor health continues to slowly decrease



$Delta_{15}$CapHealth< −15

Pattern 2: Capacitor health is large, but declines rapidly

Failure analysis

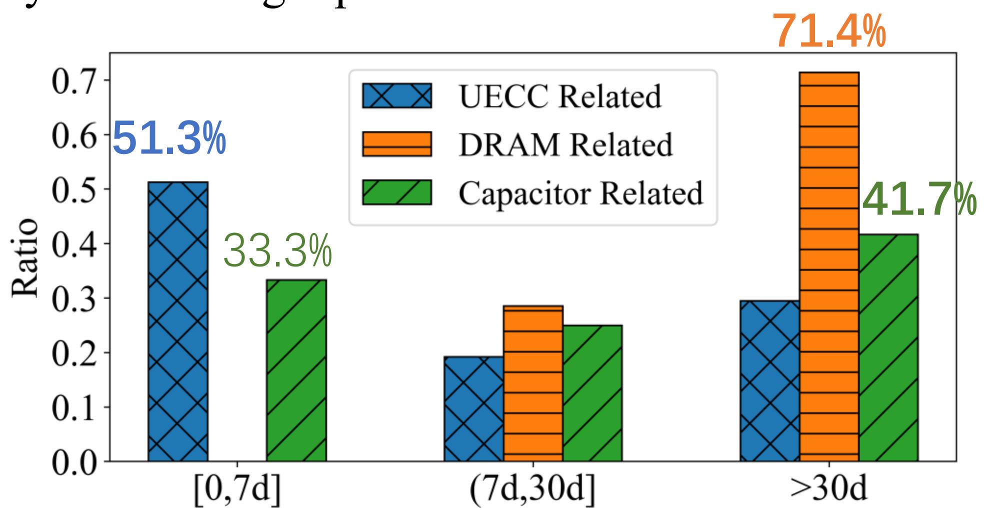
➢ Failure urgency in three patterns

- UECC-related: 51.3% within 0–7 days ⟶ Urgent
  - This may lead to data unavailability or online service interruption
  - 76% of failed SSDs will be processed by operators within two days after being reported
- DRAM-related: 71.5% > 30 days ⟶ Not urgent
  - Failure occurs only when DRAM CECC errors exceed the correction capacity
  - This may lead to SSD loss
- Capacitor-related:

  Capacitor health drop rapidly ⟶ 0–7 days urgent

  Capacitor health declines slowly ⟶ > 30 days Not urgent

  - This may lead to the Power Loss Protection (PLP) failure



Time distribution from joint decision warning to failure report

Failure analysis

➢ Factors affecting SSD health

The failure rate of SSDs is x-times higher than other ordinary SSDs

Factors	Decision	Times	Pattern
PCIe error	BadTLP ≥ 4106 & BadDLLP ≥ 43084	**34x**	Errors occur in data transmission on PCIe
	PHYError ≥ 65535 & $Delta_3$Read < $3.95 * 10^{11}$	**9x**	
Bad block and read retry	BadNandBlock ≥ 10 & ReadRetry ≥ 50	**60x**	Frequent read retries and bad blocks indicate instability or failure in storage cells
End-to-end error	ETEDeteError ≥ 10	**67x**	Errors are detected by the end-to-end error correction protection mechanism (for DRAM, SRAM)

Conclusion

FailureMiner

High-accuracy

- Boundary-preserving downsampling → Solve data imbalance
- Joint key decision extraction → Reduce the noise based on decision level
- Precision > 80%, running in Tencent's data centers over one year

Interpretable

- Joint key decision → Understandable for operators
- Facilitate failure pattern and cause analysis

User-friendly

- No machine learning knowledge required
- Fastest prediction time
- Actionable insights (i.e., failure urgency) for operators based on explainable results