



Making Disk Failure Prediction SMARTer!

***Sidi Lu*[1]**, Bing Luo[1], Tirthak Patel[2], Yongtao Yao[1], Devesh Tiwari[2], Weisong Shi[1]

[1]Wayne State University, [2]Northeastern University



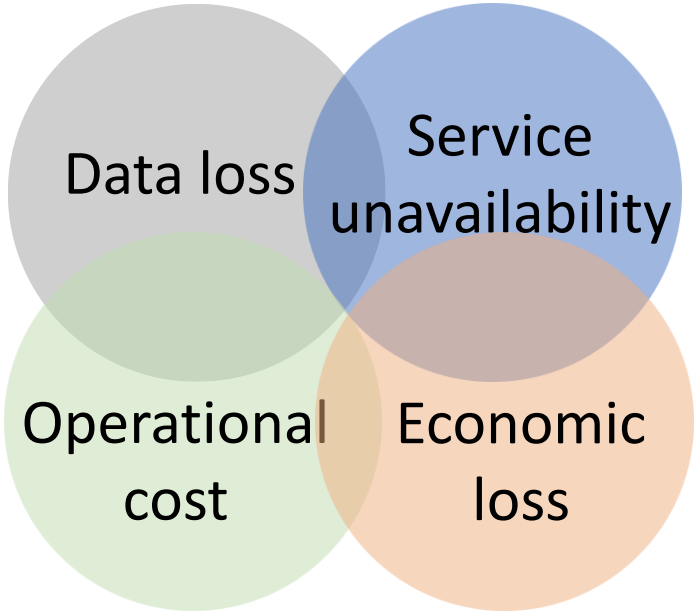


Northeastern



Introduction

- **Hard disk drives (HDD)**
  - Key driving factor of data centers
  - frequently replaced hardware components
  - main reason behind server failures





Tencent was Claimed Ten Million for Data Loss Due to Cloud Hard Drive Glitch

Source: https://medium.com/genaro-network/tencent-was-claimed-ten-million-for-data-loss-due-to-cloud-hard-drive-glitch-344a26449fe2



Tencent Cloud Storage:
An incident resulted in the loss of around 10 million RMB of data

A startup company named "Front Edge":
Basically lost the entire database that had been accumulating since its establishment

Introduction

- **Significance**
  - Storage community benefits from the disk reliability field-studies
- **Gap**
  - Disk reliability field-studies are infrequent and limited in sample size

Introduction

- **To bridge this gap**
  - One of the largest disk failure analysis studies
    - **380,000** HDDs
    - **10,000** server racks
    - **64** data center sites
    - Over **2** months
- **Hosted by a large enterprise data center operator**
- **Goal**
  - predict disk failure accurately with long prediction horizon

Key Concepts

- For the first time to demonstrate **disk failure prediction** can be highly accurate by combining
  - Disk performance data (e.g., capacity, throughput-related attributes)
  - Disk location data (disk/server/rack/room/site)
  - Disk monitoring data (Self-Monitoring, Analysis, and Reporting Technology - SMART)

- Traditional work
  - Focused on SMART data only
    - e.g., correctable errors, disk spin-up time
    - indicative of eventual failure

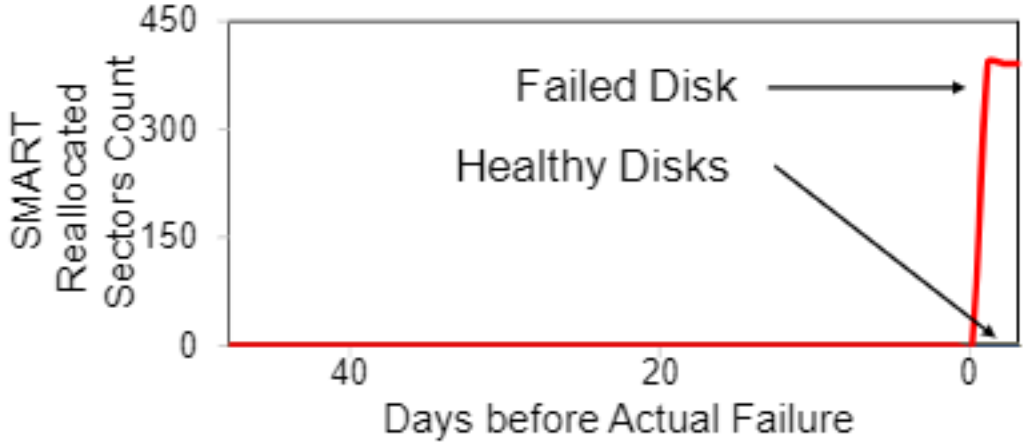
Conventional knowledge holds true

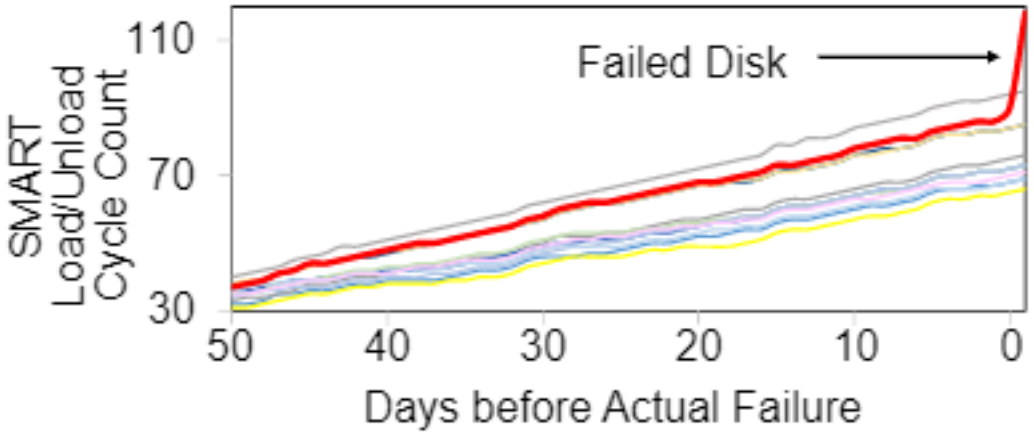
↓

Why we still consider other data?

Why not only SMART data?

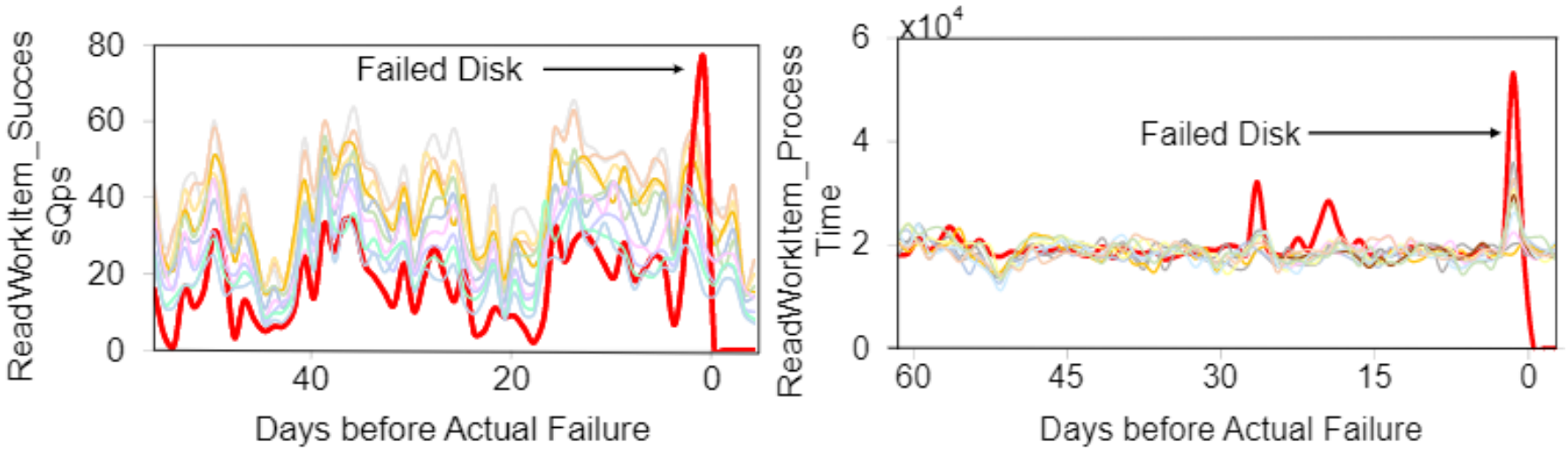
- SMART attributes do **not** always have the **strong predictive capability** at long prediction horizon windows for all disks
  - Value often do not change frequently enough before the actual failure
  - Change is often noticeable only few hours before the actual failure





Why add performance data?

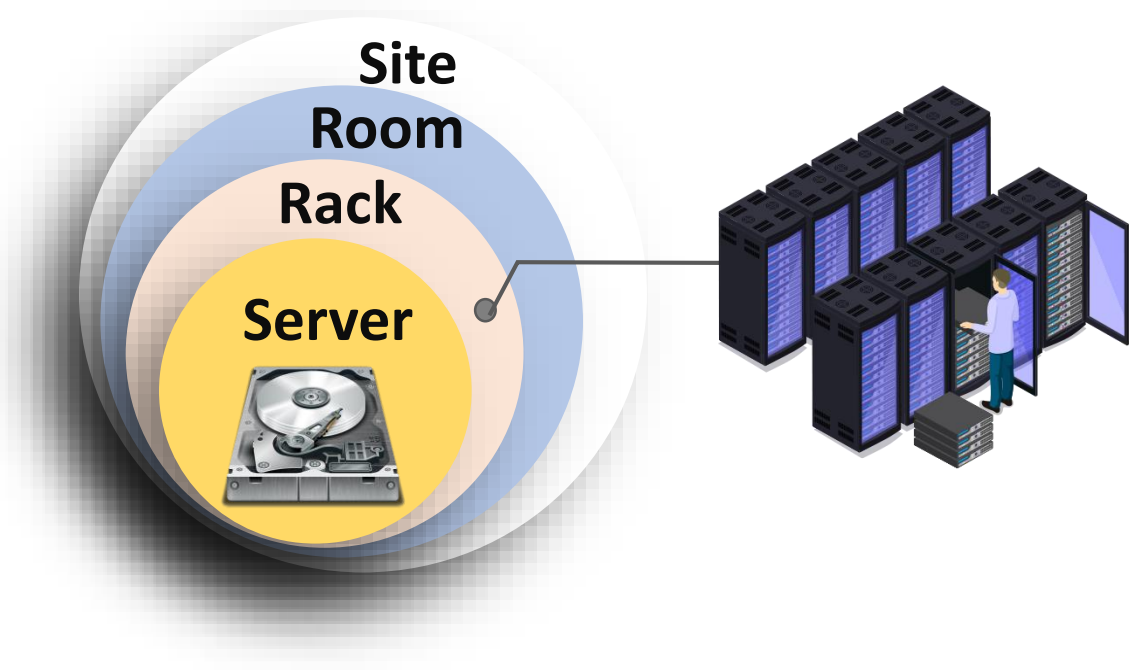
- The value of **performance** metrics (related to capacity, throughput, etc.)
  - Exhibit more variations before the actual drive failure
  - Show distinguishable behavior from healthy disks



Performance increases the coverage in **capturing the workload characteristics** beyond what SMART attributes cover

Why add location data?

- Prediction can be further improved by incorporating the **location** information
  - Disks in close spatial neighborhood
    - Affected by the same environmental factors (such as humidity and temperature)
    - Experience similar vibration level (known to affect the reliability of disks)



Location information increases our coverage of the **operating conditions** of disks

Disk SMART Data

- Select SMART attributes

  (1) Raw value: specific to the disk manufacturer

  (2) Normalized value: mapping corresponding raw value to 1-byte

- Reported at per-day granularity

ID	Attribute Names	ID	Attribute Names
1	Read Error Rate	7	Seek Error Rate
9	Power-On Hours	192	Power-off Retract Count
10	Spin Retry Count	193	Load/Unload Cycle Count
3	Spin-Up Time	194	Temperature
12	Power Cycle Count	197	Current Pending Sector Count
4	Start/Stop Count	198	Uncorrectable Sector Count
5	Reallocated Sectors Count	199	UltraDMA CRC Error Count

Performance Data

- Selected disk-level performance metrics (per-hour granularity)

Disk-level Performance Metrics	
DiskStatus	Background_Checksum_ReadFail_QpsByDisk
IOQueueSize	TempFile_WriteWorkItem_SuccessQps
ReadSuccess_Throughput	TempFile_WriteSuccess_Throughput
ReadWorkItem_QueueTime	NormalFile_WriteWorkItem_SuccessQps
ReadWorkItem_SuccessQps	NormalFile_WriteWorkItem_QueueTime
ReadWorkItem_ProcessTime	NormalFile_WriteSuccess_Throughput

- Selected server-level performance metrics (per-hour granularity)

Categories	Metrics	Categories	Metrics
disk_util	max	udp_stat	udp_outdatagrams
tcp_segs_stat	tcp_outsegs	disk_sys_read_write	read
page_activity	page_in	net_pps_summary	net_pps_receive
disk_summary	total_disk_read	net_summary	receive_speed
disk_throughput	read	page_activity	page_out
disk_util	avg	udp_stat	udp_indatagrams
memory_summary	mem_res	disk_summary	total_disk_write
tcp_currestab	NONE	net_pps_summary	net_pps_transmit
cpu_summary	cpu_kernel	tcp_segs_stat	tcp_insegs

Disk Spatial Location Data

- **Location markers**
  - Each disk has four levels of location markers associated with it: site, room, rack, and server
  - Capture the concept of neighborhood
  - Do not explicitly indicate the actual physical proximity between two disks (physical distance is not captured by our location coordinates)

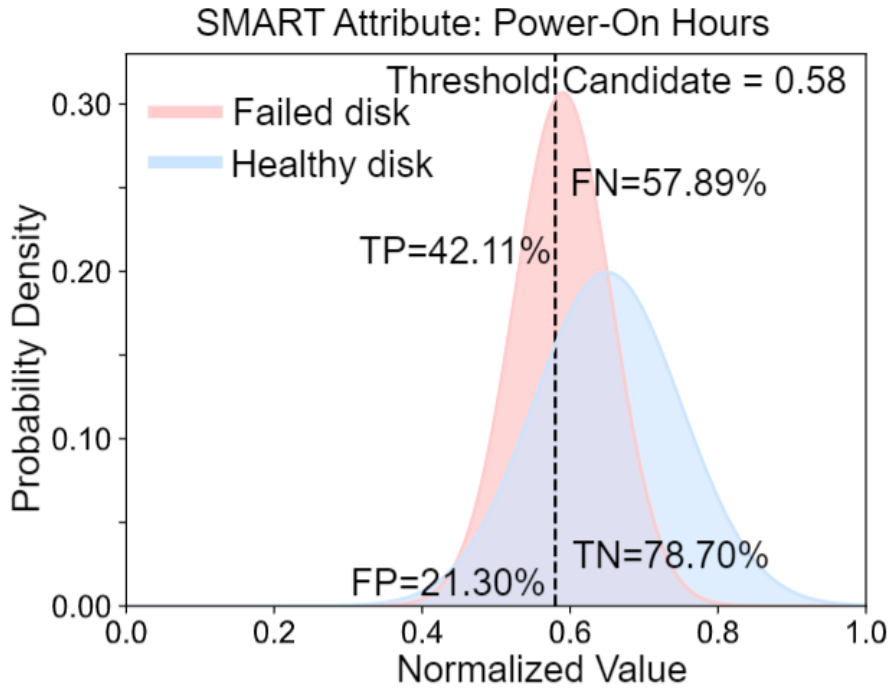
	# of Sites	# of Rooms	# of Racks	# of Servers
Total	64	199	10,440	120,000

Selection of Attributes

- Min-max normalization: $x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$
- J-Index classification:

$$\text{J-Index} = \text{True Positive Rate} + \text{True Negative Rate} - 1$$

$$= \frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}} - 1$$



For a given feature:

(1) Set a series of threshold candidates with the step of 0.01,

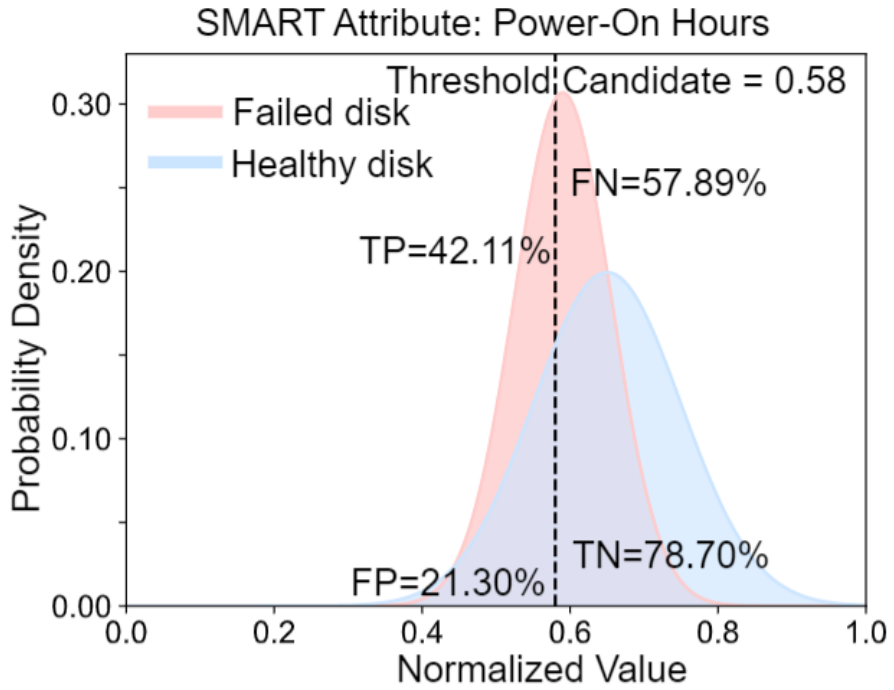
(2) Calculate their corresponding **J-Indexes**

Selection of Attributes

- Min-max normalization: $x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$
- J-Index classification:

$$\text{J-Index} = \text{True Positive Rate} + \text{True Negative Rate} - 1$$

$$= \frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}} - 1$$



- **Higher** J-Index: **more distinguishable**
- The threshold candidate with the **highest J-Index**: the best (final) threshold
- Select features with **highest J-Indexes** as the informative features

Selection of Attributes

FAST'20

Highest J-Indexes for SMART attributes

ID	Threshold	J-Index	% of failed disks		% of healthy disks	
9R	0.58	0.21	**42%**	58%	21%	**79%**
9N	0.54	0.19	52%	**48%**	**72%**	28%
3R	0.72	0.18	80%	**20%**	**98%**	2%
5R	0.49	0.18	**18%**	82%	0%	**100%**
194N	0.50	0.18	**45%**	55%	27%	**73%**
194R	0.50	0.15	**96%**	4%	81%	**19%**
1R	0.38	0.14	**28%**	72%	14%	**86%**
12N	0.04	0.14	65%	**35%**	**79%**	21%
4N	0.01	0.13	64%	**36%**	**77%**	23%
5N	0.01	0.13	87%	**13%**	**100%**	0%
3N	0.59	0.13	77%	**23%**	**90%**	10%

Highest J-Indexes for performance metrics

Threshold	J-Index	% of failed disk		% of healthy disk	
0.13	0.45	**100%**	0%	55%	**45%**
0.11	0.44	2%	**98%**	**46%**	54%
0.10	0.40	8%	**92%**	**48%**	52%
0.16	0.40	8%	**92%**	**49%**	51%
0.15	0.38	12%	**88%**	**50%**	50%
0.12	0.31	25%	**75%**	**56%**	44%
0.10	0.30	27%	**73%**	**56%**	44%

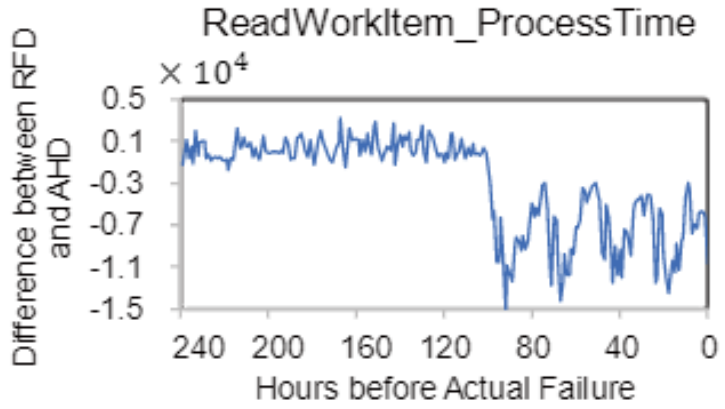
Single **performance metric** has an overall **higher J-Index** than a SMART attribute

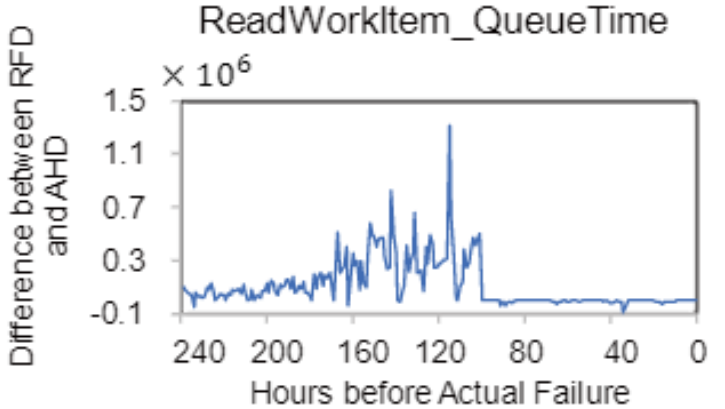
↓

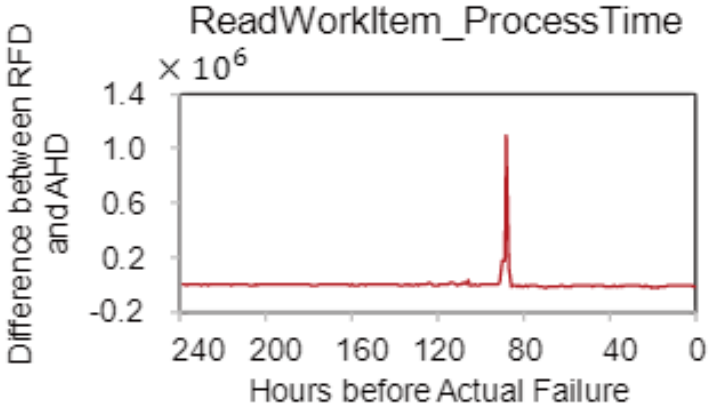
Performance metrics are likely to be **predictive** for disk failures

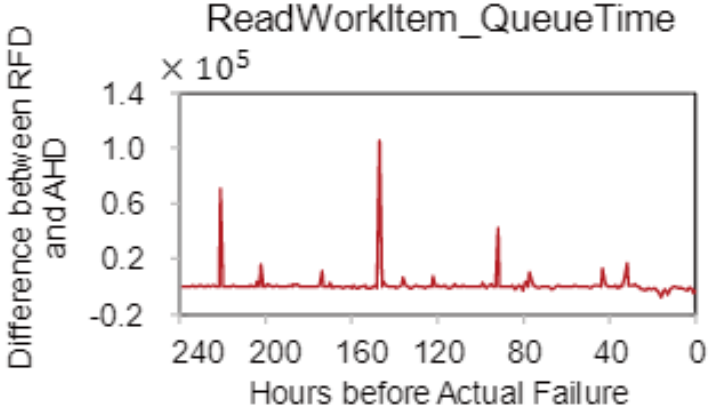
Patterns of Performance Metrics

The difference between the signatures of failed and healthy disks on the same server



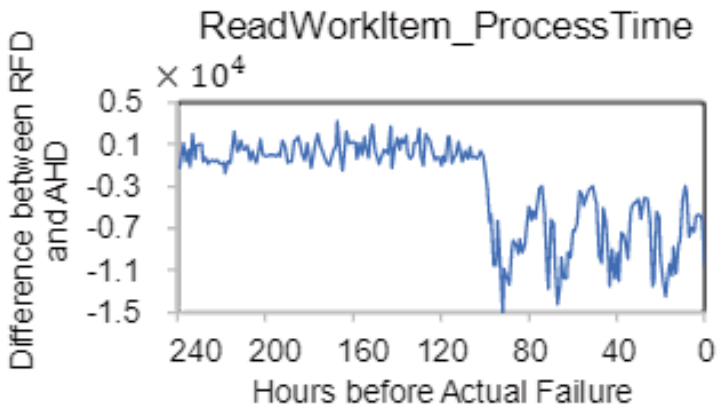


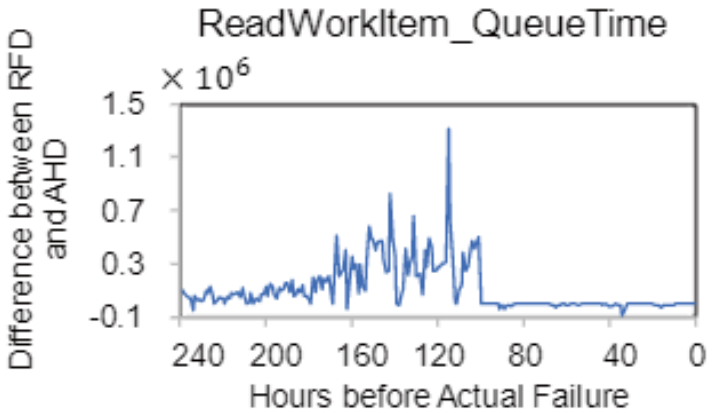




- 240 hours before actual failures
- Raw value of the failed disk (RFD)
- Average value of all healthy disks (AHD)
- RFD - AHD

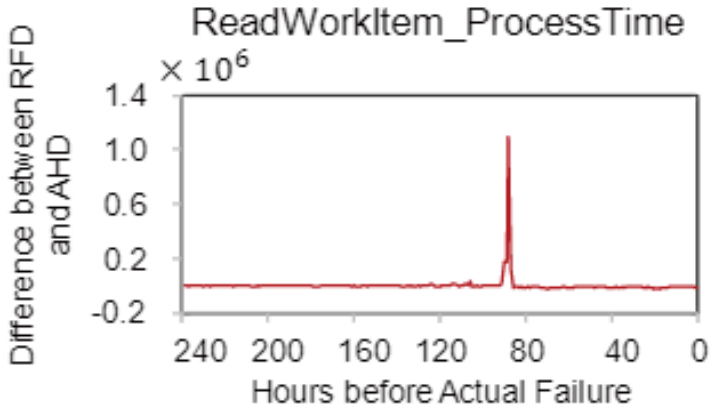
Patterns of Performance Metrics

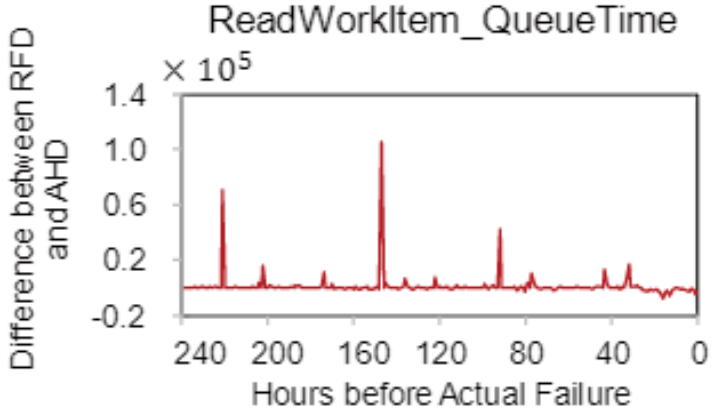




- **Top two graphs:**

  Some failed disks have a similar value to healthy disks at first, but then their behavior becomes unstable





- **Bottom two graphs:**

  Some failed disks report a sharp impulse before they fail

Effective Measurements

- To evaluate the effectiveness of our prediction approaches

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

(Matthews correlation coefficient)

ML Models

Employ five ML methods:

- Bayes classifier (Bayes)
- Random forest (RF)
- Gradient boosted decision trees (GBDT)
- Long short-term memory network (LSTM)
- Convolutional neural network with long short-term memory (CNN-LSTM)

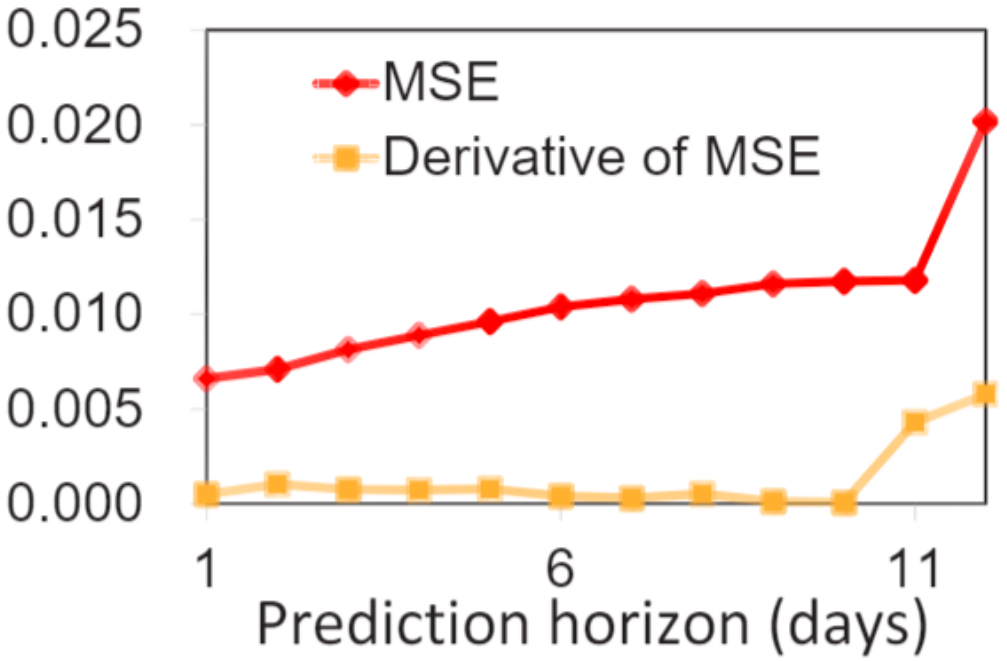
Feature Group Sets

- Construct six groups using different feature combinations
  - to evaluate the effectives of SMART (S), performance (P) and location (L) data

	SMART	Performance			Location
SPL Group	28 attributes	12 metrics	18 metric categories	18 metrics	1 marker
SL Group	28 attributes	NONE			1 marker
SP Group	28 attributes	12 metrics	18 metric categories	18 metrics	NONE
PL Group	NONE	12 metrics	18 metric categories	18 metrics	1 marker
S Group	28 attributes	NONE			NONE
P Group	NONE	12 metrics	18 metric categories	18 metrics	NONE

Prediction Horizon Selection

- Predict if a given disk will fail within the next 10 days
  - **Long enough** for IT operators to conduct early countermeasures
- Sensitivity study (Mean Squared Error)
  - Derivative of MSE reaches the minimum on the 10th day
  - MSE increases after 10 days

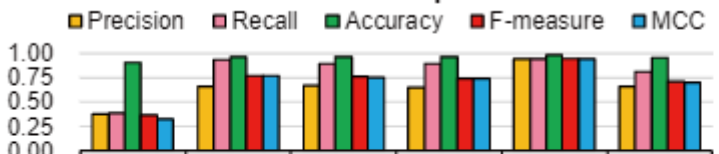


Experimental Results

FAST'20

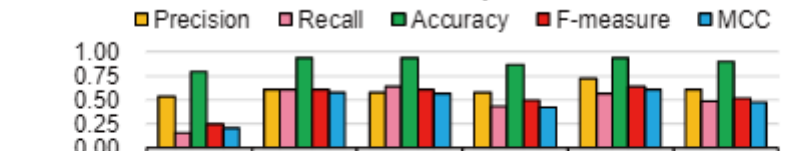
Results of 6 groups by employing 5 machine learning methods:

SPL Group



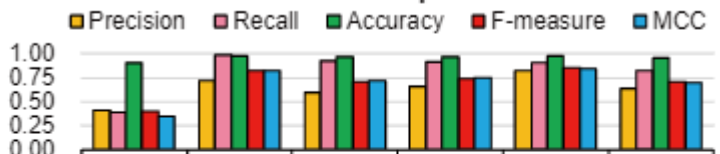
	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.38	0.66	0.67	0.65	0.95	0.66
Recall	0.39	0.94	0.89	0.89	0.95	0.81
Accuracy	0.90	0.97	0.97	0.97	0.99	0.96
F-measure	0.36	0.77	0.76	0.74	0.95	0.72
MCC	0.32	0.77	0.76	0.74	0.95	0.71

SL Group



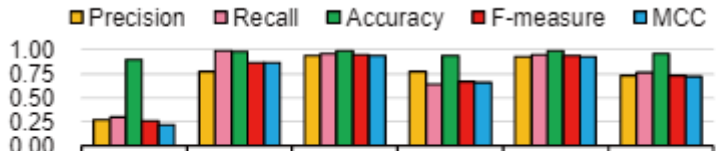
	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.53	0.61	0.57	0.58	0.72	0.60
Recall	0.16	0.61	0.64	0.43	0.57	0.48
Accuracy	0.79	0.94	0.93	0.87	0.94	0.89
F-measure	0.25	0.61	0.61	0.49	0.63	0.52
MCC	0.20	0.58	0.57	0.42	0.60	0.48

SP Group



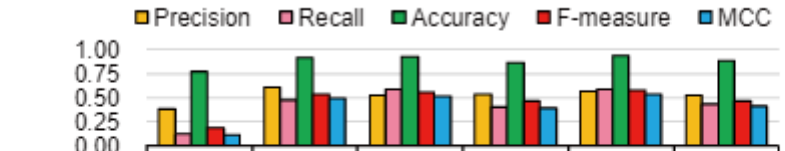
	BayesBayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.41	0.72	0.60	0.66	0.82	0.64
Recall	0.39	0.99	0.92	0.91	0.90	0.82
Accuracy	0.90	0.98	0.96	0.97	0.98	0.96
F-measure	0.40	0.82	0.71	0.74	0.85	0.71
MCC	0.35	0.83	0.72	0.75	0.85	0.70

PL Group



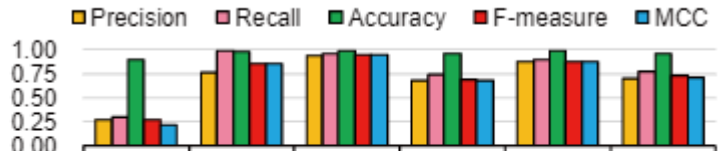
	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.26	0.77	0.93	0.77	0.93	0.73
Recall	0.29	0.99	0.96	0.63	0.94	0.76
Accuracy	0.89	0.98	0.99	0.93	0.99	0.96
F-measure	0.26	0.86	0.94	0.67	0.93	0.74
MCC	0.22	0.86	0.94	0.66	0.93	0.72

S Group



	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.38	0.61	0.52	0.54	0.57	0.52
Recall	0.12	0.48	0.58	0.40	0.58	0.43
Accuracy	0.77	0.92	0.92	0.86	0.93	0.88
F-measure	0.19	0.54	0.55	0.46	0.58	0.46
MCC	0.11	0.50	0.51	0.39	0.54	0.41

P Group

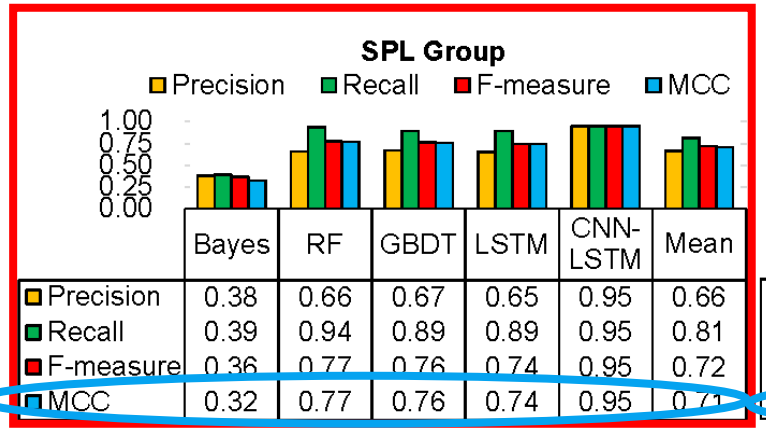


	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.26	0.76	0.94	0.68	0.87	0.70
Recall	0.30	0.99	0.96	0.74	0.89	0.78
Accuracy	0.90	0.98	0.99	0.95	0.98	0.96
F-measure	0.26	0.86	0.95	0.69	0.87	0.73
MCC	0.22	0.86	0.94	0.67	0.87	0.71

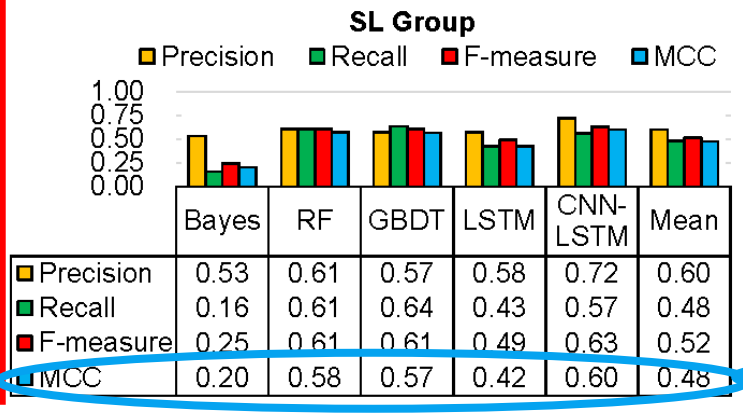
Observation #1

1. **SPL group performs the best across all ML models**

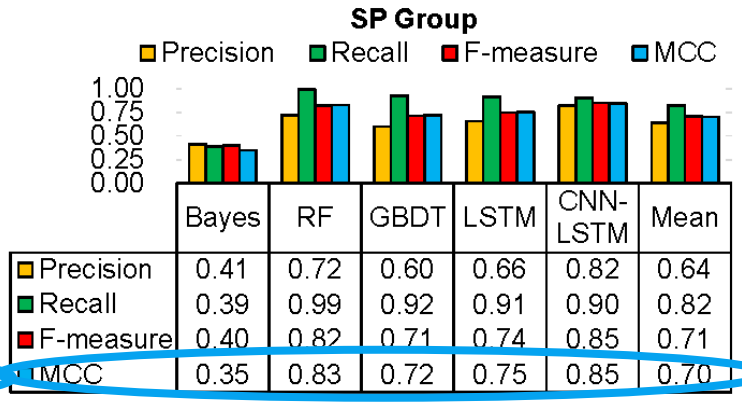
   - **performance and location features improve the effectiveness of prediction**



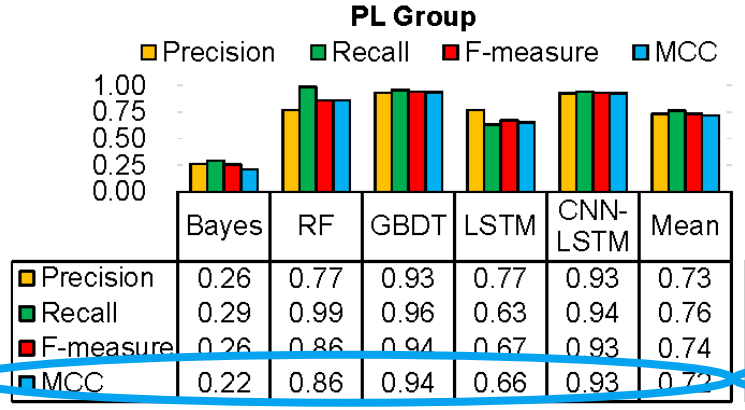
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.66	0.67	0.65	0.95	0.66
Recall	0.39	0.94	0.89	0.89	0.95	0.81
F-measure	0.36	0.77	0.76	0.74	0.95	0.72
MCC	0.32	0.77	0.76	0.74	0.95	0.71



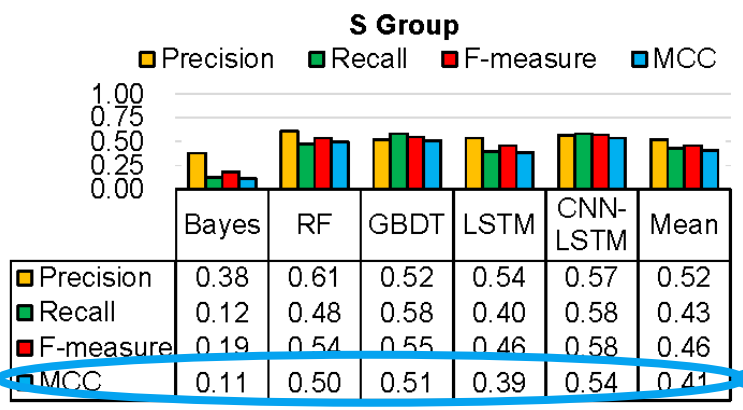
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.53	0.61	0.57	0.58	0.72	0.60
Recall	0.16	0.61	0.64	0.43	0.57	0.48
F-measure	0.25	0.61	0.61	0.49	0.63	0.52
MCC	0.20	0.58	0.57	0.42	0.60	0.48



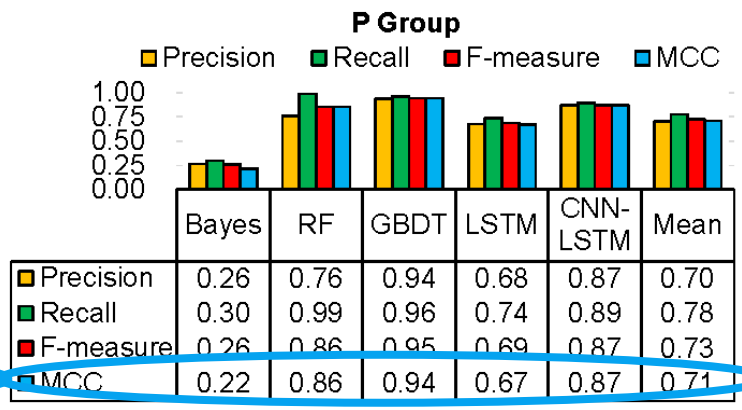
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.41	0.72	0.60	0.66	0.82	0.64
Recall	0.39	0.99	0.92	0.91	0.90	0.82
F-measure	0.40	0.82	0.71	0.74	0.85	0.71
MCC	0.35	0.83	0.72	0.75	0.85	0.70



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.77	0.93	0.77	0.93	0.73
Recall	0.29	0.99	0.96	0.63	0.94	0.76
F-measure	0.26	0.86	0.94	0.67	0.93	0.74
MCC	0.22	0.86	0.94	0.66	0.93	0.73



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.61	0.52	0.54	0.57	0.52
Recall	0.12	0.48	0.58	0.40	0.58	0.43
F-measure	0.19	0.54	0.55	0.46	0.58	0.46
MCC	0.11	0.50	0.51	0.39	0.54	0.41



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.76	0.94	0.68	0.87	0.70
Recall	0.30	0.99	0.96	0.74	0.89	0.78
F-measure	0.26	0.86	0.95	0.69	0.87	0.73
MCC	0.22	0.86	0.94	0.67	0.87	0.71

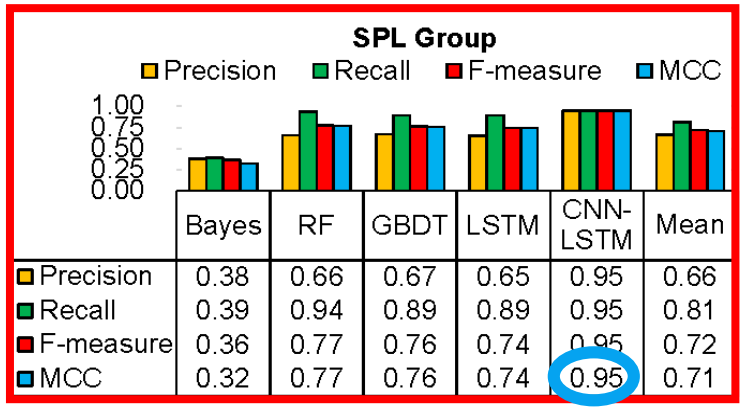
Observation #2

FAST'20

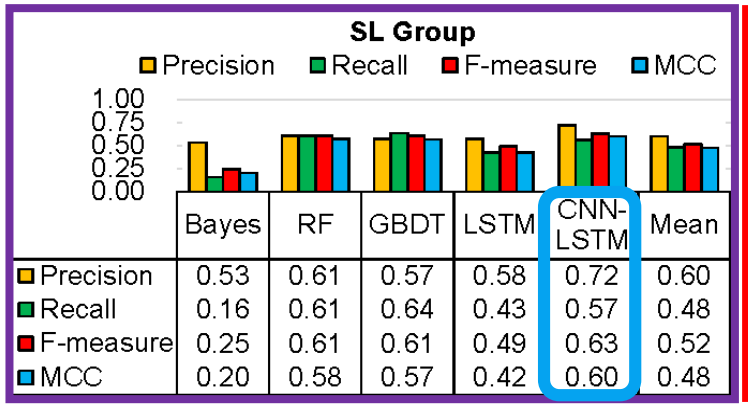
2. **The improvement of adding location info is limited and pronounced only in the presence of performance features**

- Without location markers: 10% reduction for CNN-LSTM in terms of MCC

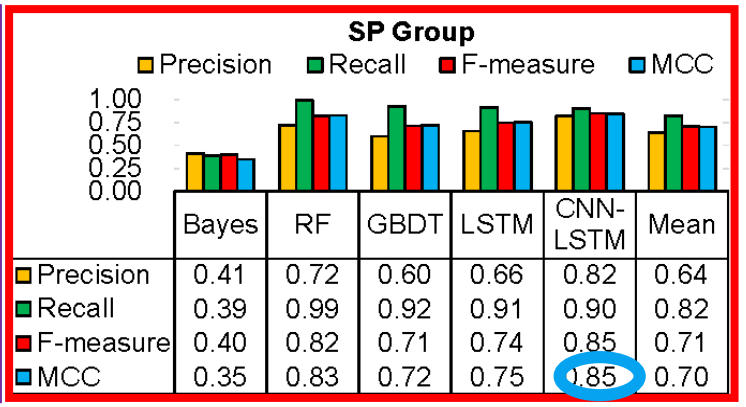
- Location info may help ML models amplify the hidden patterns in performance metrics



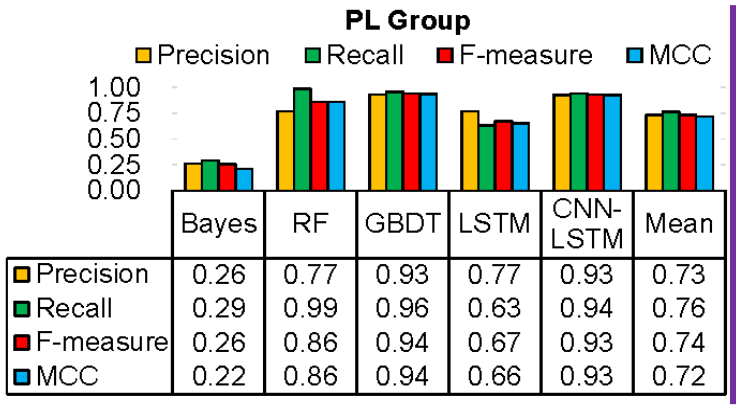
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.66	0.67	0.65	0.95	0.66
Recall	0.39	0.94	0.89	0.89	0.95	0.81
F-measure	0.36	0.77	0.76	0.74	0.95	0.72
MCC	0.32	0.77	0.76	0.74	0.95	0.71



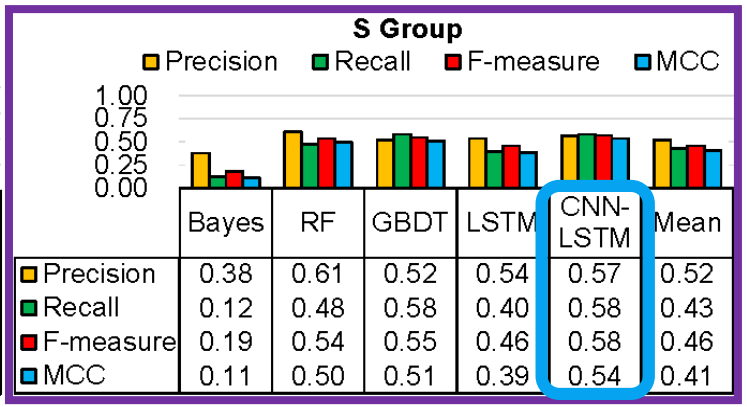
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.53	0.61	0.57	0.58	0.72	0.60
Recall	0.16	0.61	0.64	0.43	0.57	0.48
F-measure	0.25	0.61	0.61	0.49	0.63	0.52
MCC	0.20	0.58	0.57	0.42	0.60	0.48



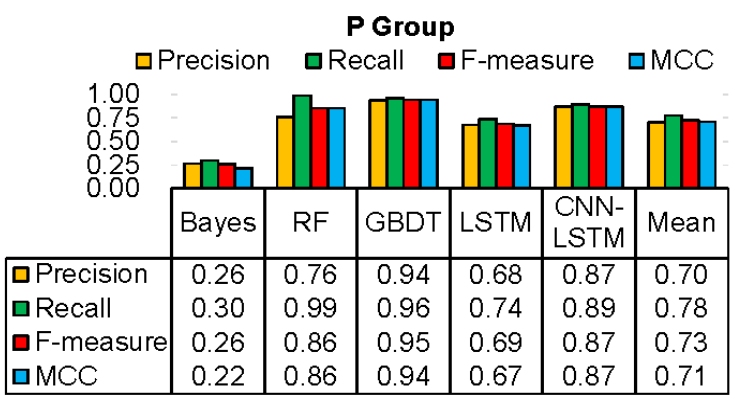
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.41	0.72	0.60	0.66	0.82	0.64
Recall	0.39	0.99	0.92	0.91	0.90	0.82
F-measure	0.40	0.82	0.71	0.74	0.85	0.71
MCC	0.35	0.83	0.72	0.75	0.85	0.70



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.77	0.93	0.77	0.93	0.73
Recall	0.29	0.99	0.96	0.63	0.94	0.76
F-measure	0.26	0.86	0.94	0.67	0.93	0.74
MCC	0.22	0.86	0.94	0.66	0.93	0.72



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.61	0.52	0.54	0.57	0.52
Recall	0.12	0.48	0.58	0.40	0.58	0.43
F-measure	0.19	0.54	0.55	0.46	0.58	0.46
MCC	0.11	0.50	0.51	0.39	0.54	0.41



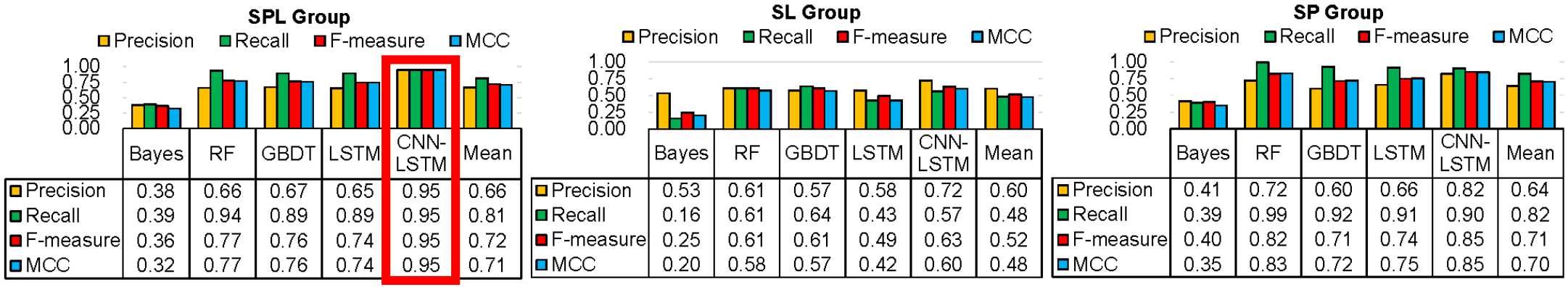
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.76	0.94	0.68	0.87	0.70
Recall	0.30	0.99	0.96	0.74	0.89	0.78
F-measure	0.26	0.86	0.95	0.69	0.87	0.73
MCC	0.22	0.86	0.94	0.67	0.87	0.71

Observation #3

3. **CNN-LSTM performs close to the best in all situations**

   - Achieve MCC score of 0.95 for SPL group

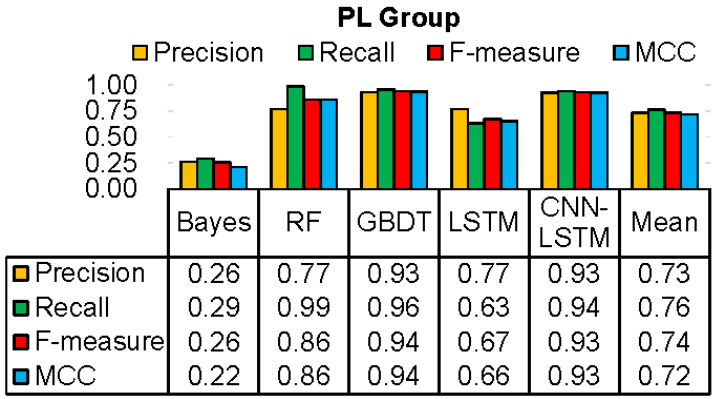
   - LSTM could be further improved by taking better features as the input, which could be provided by CNN through dimensionality reduction



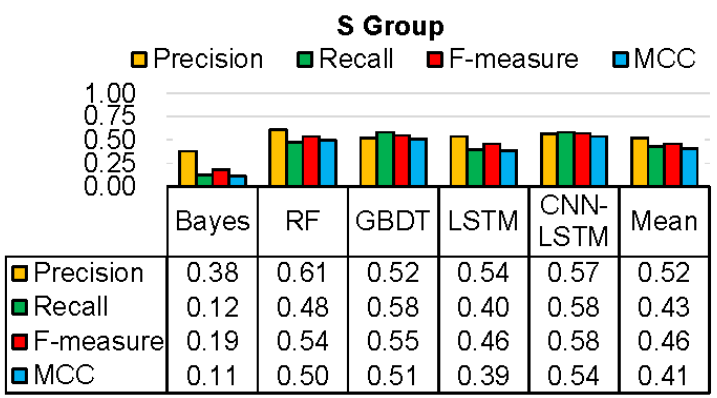
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.66	0.67	0.65	0.95	0.66
Recall	0.39	0.94	0.89	0.89	0.95	0.81
F-measure	0.36	0.77	0.76	0.74	0.95	0.72
MCC	0.32	0.77	0.76	0.74	0.95	0.71

	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.53	0.61	0.57	0.58	0.72	0.60
Recall	0.16	0.61	0.64	0.43	0.57	0.48
F-measure	0.25	0.61	0.61	0.49	0.63	0.52
MCC	0.20	0.58	0.57	0.42	0.60	0.48

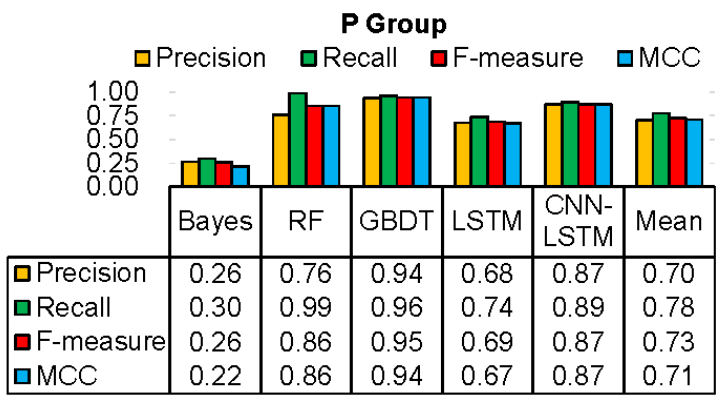
	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.41	0.72	0.60	0.66	0.82	0.64
Recall	0.39	0.99	0.92	0.91	0.90	0.82
F-measure	0.40	0.82	0.71	0.74	0.85	0.71
MCC	0.35	0.83	0.72	0.75	0.85	0.70



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.77	0.93	0.77	0.93	0.73
Recall	0.29	0.99	0.96	0.63	0.94	0.76
F-measure	0.26	0.86	0.94	0.67	0.93	0.74
MCC	0.22	0.86	0.94	0.66	0.93	0.72



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.38	0.61	0.52	0.54	0.57	0.52
Recall	0.12	0.48	0.58	0.40	0.58	0.43
F-measure	0.19	0.54	0.55	0.46	0.58	0.46
MCC	0.11	0.50	0.51	0.39	0.54	0.41



	Bayes	RF	GBDT	LSTM	CNN-LSTM	Mean
Precision	0.26	0.76	0.94	0.68	0.87	0.70
Recall	0.30	0.99	0.96	0.74	0.89	0.78
F-measure	0.26	0.86	0.95	0.69	0.87	0.73
MCC	0.22	0.86	0.94	0.67	0.87	0.71

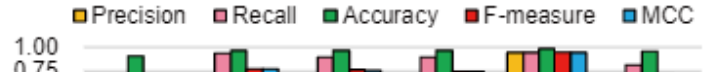
Observation #4

4. **Trade-off between models with respect to different availability of feature sets**

- **In absence of performance and location features**

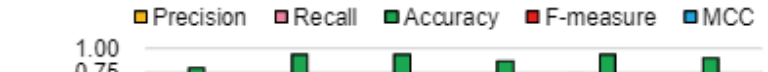
  - Traditional tree-based ML models (RF and GBDT) can provide equally accurate predictions as complicated neural network based model (CNN-LSTM or LSTM)

SPL Group



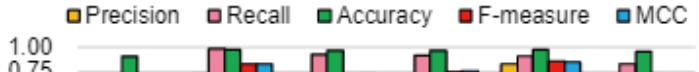
	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.38	0.66	0.67	0.65	0.95	0.66
Recall	0.39	0.94	0.89	0.89	0.95	0.81
Accuracy	0.90	0.97	0.97	0.97	0.99	0.96
F-measure	0.36	0.77	0.76	0.74	0.95	0.72
MCC	0.32	0.77	0.76	0.74	0.95	0.71

SL Group



	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.53	0.61	0.57	0.58	0.72	0.60
Recall	0.16	0.61	0.64	0.43	0.57	0.48
Accuracy	0.79	0.94	0.93	0.87	0.94	0.89
F-measure	0.25	0.61	0.61	0.49	0.63	0.52
MCC	0.20	0.58	0.57	0.42	0.60	0.48

SP Group



	BayesBayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.41	0.72	0.60	0.66	0.82	0.64
Recall	0.39	0.99	0.92	0.91	0.90	0.82
Accuracy	0.90	0.98	0.96	0.97	0.98	0.96
F-measure	0.40	0.82	0.71	0.74	0.85	0.71
MCC	0.35	0.83	0.72	0.75	0.85	0.70

PL Group



	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.26	0.77	0.93	0.77	0.93	0.73
Recall	0.29	0.99	0.96	0.63	0.94	0.76
Accuracy	0.89	0.98	0.99	0.93	0.99	0.96
F-measure	0.26	0.86	0.94	0.67	0.93	0.74
MCC	0.22	0.86	0.94	0.66	0.93	0.72

S Group



	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.38	0.61	0.52	0.54	0.57	0.52
Recall	0.12	0.48	0.58	0.40	0.58	0.43
Accuracy	0.77	0.92	0.92	0.86	0.93	0.88
F-measure	0.19	0.54	0.55	0.46	0.58	0.46
MCC	0.11	0.50	0.51	0.39	0.54	0.41

P Group

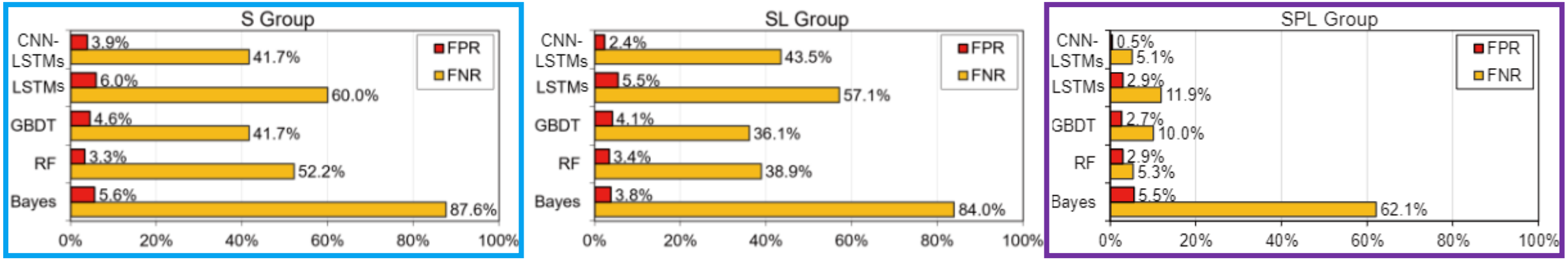


	Bayes	RF	GBDT	LSTMs	CNN-LSTMs	Mean
Precision	0.26	0.76	0.94	0.68	0.87	0.70
Recall	0.30	0.99	0.96	0.74	0.89	0.78
Accuracy	0.90	0.98	0.99	0.95	0.98	0.96
F-measure	0.26	0.86	0.95	0.69	0.87	0.73
MCC	0.22	0.86	0.94	0.67	0.87	0.71

Further Exploration

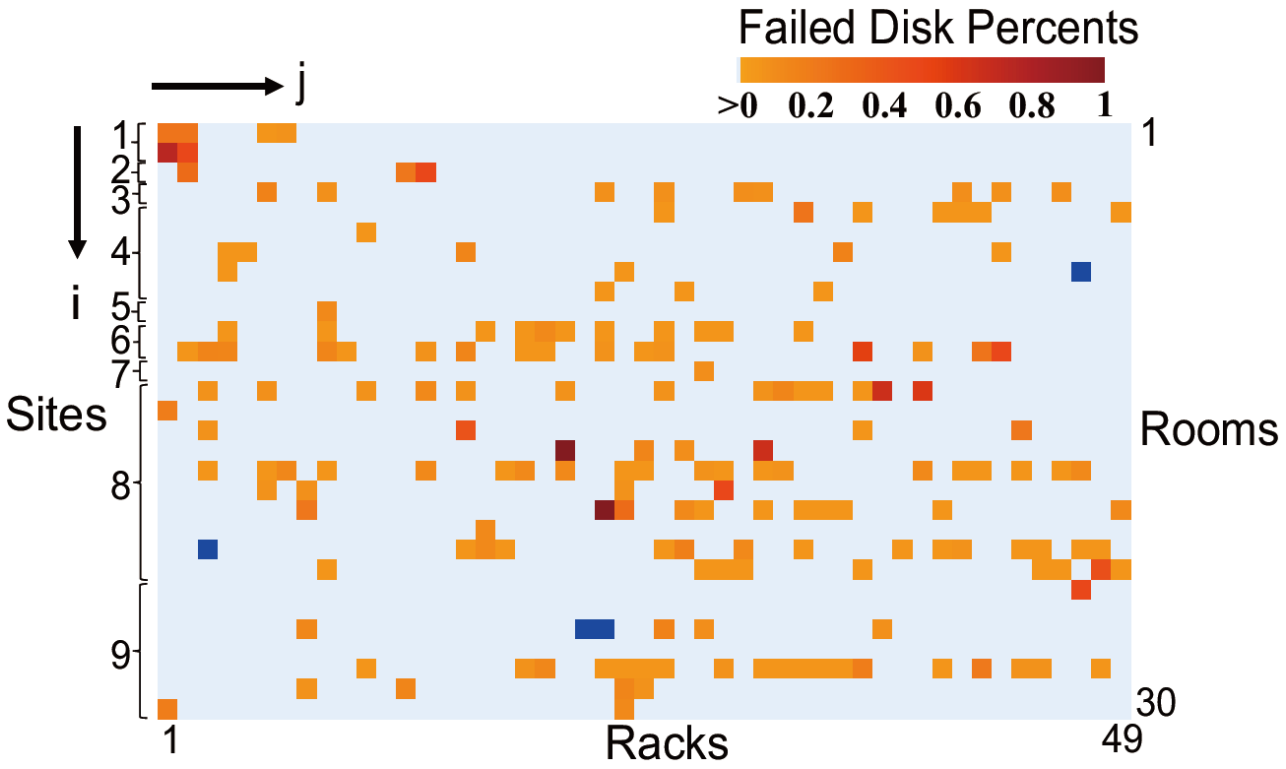
False positive and false negative rates for different ML models and different feature groups

false positive rate ($FPR = FP/(FP+TN)$) false negative rate ($FNR = FN/(TP+FN)$)



- SMART attribute based models:
  - High FNR (failed disks predicted healthy) across all models
- Adding performance and location features
  - Decreases FNR significantly
  - Prediction quality goes up

Where do ML models perform poorly and why?



Mispredicted failures (blue) tend to occur in low failure locations for all models

Where:

Concentration of failures is relatively lower

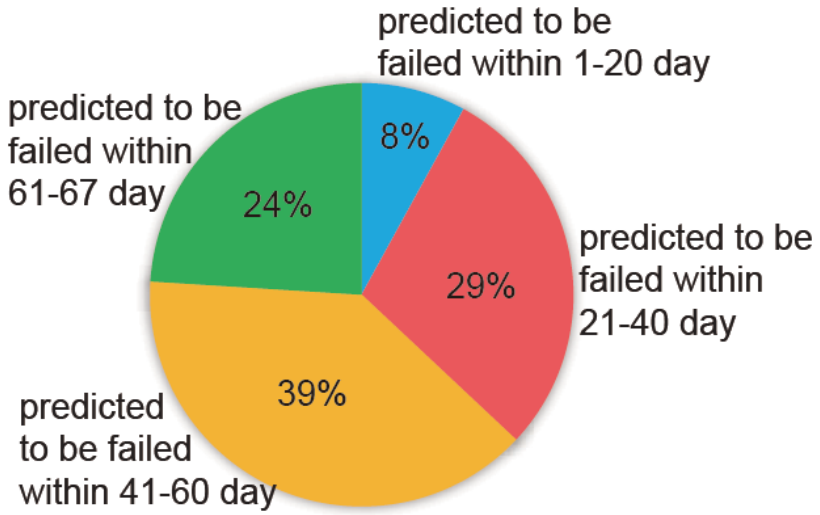
Why:

ML models are not able to collect enough failed disk samples

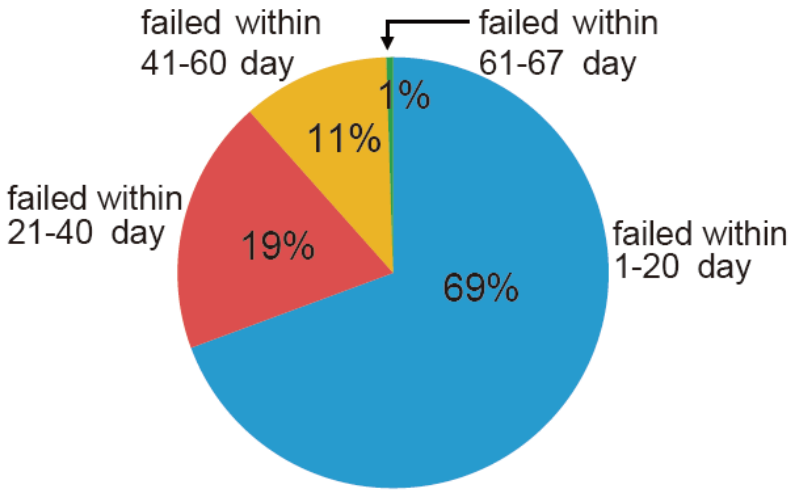
Significance:

Emphasizes the need for adding location markers in disk failure prediction models

When do ML models fail to predict and why?



false positives (healthy disks predicted as failed) categorized in 20-day windows for CNN-LSTM model



false negative (failed disks predicted healthy) categorized in 20-day windows for CNN-LSTM model

When:

The number of false positives is very low initially as it predicts many disks as healthy although they eventually failed in that window

- and, this is why the false negatives are high initially

Why:

ML model does not have enough data and (conservatively) predicts that disks are healthy

Significance:

The need for sufficiently long testing periods before concluding the prediction quality

Is the prediction model portable across data center sites?

FAST'20

- If simply to train on one data center site and port it to another
  - disk failure prediction model may not work

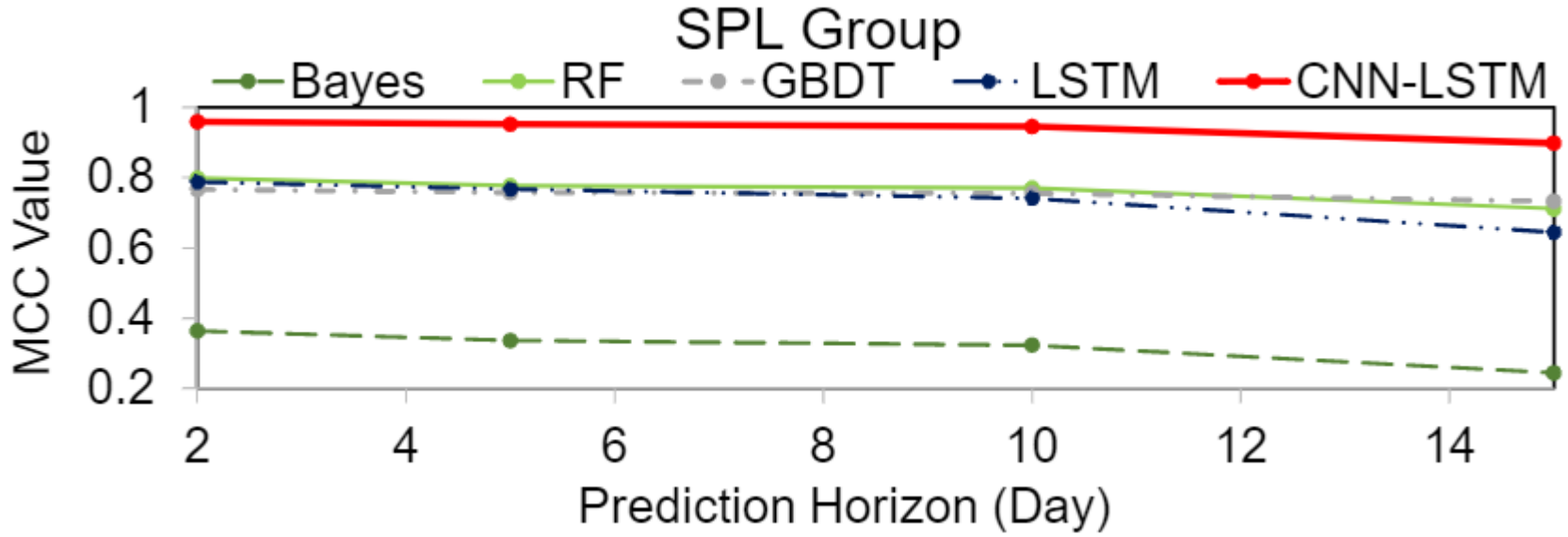
- Training on multiple data sites before testing on a new unseen data site provides reasonable accuracy

	Precision	Recall	F-measure	MCC
Bayes	0.35	0.37	0.36	0.31
RF	0.66	0.94	0.78	0.78
GBDT	0.65	0.89	0.75	0.74
LSTM	0.66	0.88	0.75	0.74
CNN-LSTM	0.93	0.94	0.94	0.93

Tested on two unseen data center sites A and B, while training models on rest of the 62 sites

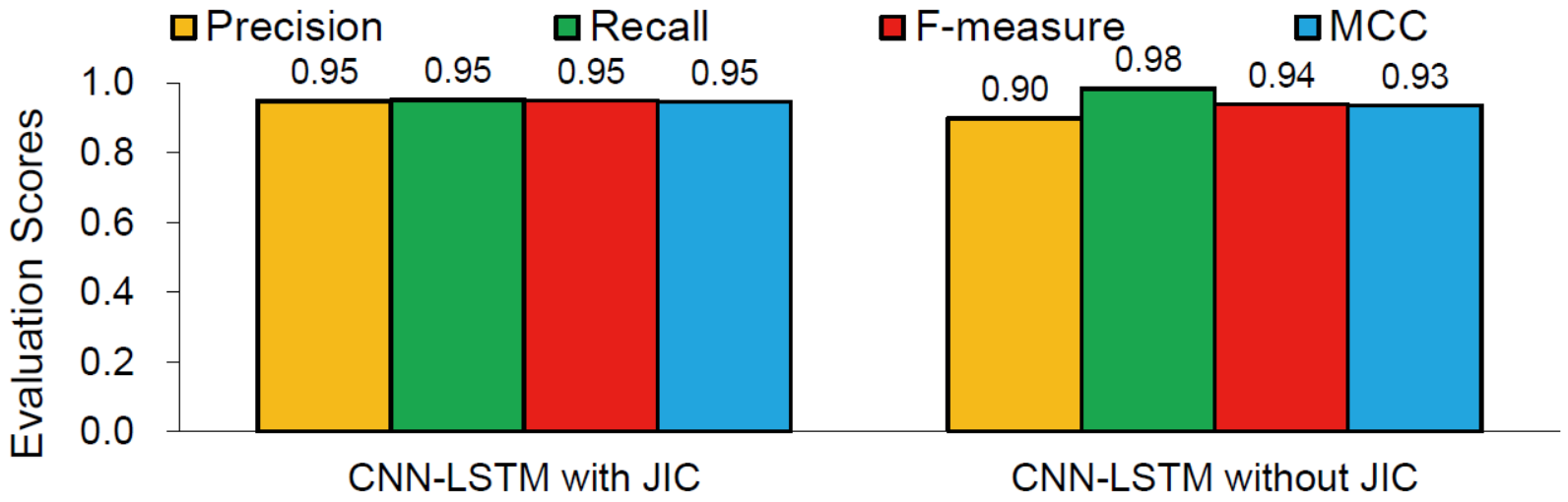
Prefer **CNN-LSTM model** if portability is a requirement

Is the prediction model effective at different prediction horizon (lead time)?



- Prediction quality indeed goes down with increasing prediction horizon window
- Rate of decrease is not steep for any model

Does J-Index classification for feature selection degrade the overall prediction accuracy compared to models trained with all features?



- Provide similar quality results
- Suggestion:
  use J-Index to manage the storage overhead of storing attributes without risking the prediction quality significantly

Conclusion

- One of the largest disk failure prediction studies
  - Covering 380,000 hard drives across 64 sites of a leading e-commerce site
- Performance and location attributes are effective in improving the disk failure prediction quality
- No single machine learning model is a winner across all scenarios, although CNN-LSTM is fairly effective across different situations
- Train models up to 0.95 F-measure and 0.95 MCC score for a 10-day prediction horizon

Q & A

- Our disk failure prediction framework and the dataset used are hosted at http://codegreen.cs.wayne.edu/wizard
- Email: lu.sidi@wayne.edu