Hybrid Deep-Learning for Fail-Slow Disk Detection in the FSA-Benchmark

Joshua Ludolf

Department of Computer Science,
University of Texas at Austin
Austin, TX 78712
joshua.ludolf@my.utexas.edu

ABSTRACT

Fail-slow disks – where performance degrades gradually before an outright failure – are increasingly common in large-scale cloud storage. Existing FSA-Benchmark reproductions have shown that tree-based models (XGBoost, Random Forest) and shallow time-series methods (LSTM, SVM) can detect such conditions with moderate precision, but they struggle to capture the high-frequency, multivariate correlations that precede a fail-slow. This work proposes a hybrid deep-learning framework that combines a convolutional–recurrent backbone with self-attention to model both spatial and temporal dependencies in the 15-second-interval performance metrics collected by the FSA-Benchmark ((≈ 100 billion data points from 300k disks)s). The model ingests a multivariate time-window (look-back 5 min) and outputs a real-time probability of an impending fail-slow. We evaluate the approach on the same "Cluster A" and "Cluster B" splits used in the original study, reporting precision, recall, AUROC, and the Time-to-Alert (the latency between the first abnormal signal and the model's positive prediction). Our preliminary experiments indicate...

Keywords: *AUROC, deep learning, diagnostic research, fail-slow disk, SHAP*

INTRODUCTION & BACKGROUND

The proliferation of cloud-scale storage has made the reliability of disk subsystems a critical determinant of overall system availability. Traditional fault-tolerance strategies have focused on hard-disk failures that are readily detected by abrupt drops in throughput or sudden increases in latency. In practice, however, a large fraction of disk degradations manifest as fail-slow conditions, in which performance metrics such as latency, error-rate, and queue depth drift slowly over minutes to hours before a catastrophic failure occurs. Because fail-slow signatures are subtle, they are often missed by conventional rule-based monitoring, resulting in missed opportunities for pre-emptive remediation and costly downtime.

The FSA-Benchmark, built from the publicly available PERSEUS dataset, provides a rare, high-frequency window into this problem: 100 billion 15-second-interval samples from 300k disks across multiple clusters. The benchmark has become a de-facto standard for evaluating anomaly-detection algorithms in storage systems, yet reproductions of the original work rely primarily on tree-based models (XGBoost, Random Forest) and shallow recurrent networks (LSTM). While these models achieve moderate precision, they are limited in their ability to capture the highly non-linear, multi-scale interactions between the dozens of performance counters that precede a fail-slow event.

Deep-learning offers a natural remedy. Convolutional layers can learn local temporal patterns—such as latency spikes or bursts of errors—directly from raw time-series data, while recurrent layers can model longer-range dependencies across the 5-minute windows that precede failure. Adding a self-attention module further enables the network to learn which metrics (e.g., queue depth versus latency) and which sub-intervals within a window contribute most to the final decision, yielding a transparent, data-driven "early-warning" signal. This paper proposes a hybrid CNN-RNN-Attention architecture that ingests 20 consecutive 15-second samples (a 5-minute look-back) from the PERSEUS dataset, outputs a fail-slow probability, and outputs interpretable attention maps. We evaluate this approach against the best-performing baseline (XGBoost) on the same train/validation/test splits used in the original FSA-Benchmark study, reporting standard metrics (precision, recall, AUROC) as well as a novel time-to-alert metric that captures how many minutes earlier our model can predict a fail-slow relative to the first abnormal signal. By combining high predictive performance with explainability, our work advances the state of the art in proactive storage-system reliability and provides a reproducible, scalable framework that can be readily adopted in production monitoring pipelines.

OBJECTIVES

Primary Objectives	**Secondary Objectives**
Design a hybrid CNN–RNN–Self-Attention network that ingests 5-minute windows of 15-second metrics and outputs a fail-slow probability.	*Quantify the benefit of self-attention over vanilla RNN.*
Compare the hybrid model against the best-performing baseline model on the same train/validation/test splits used in the original paper.	*Report precision, recall, AUROC, and Time-to-Alert.*
Provide interpretability (attention maps, SHAP) to pinpoint which metrics and time-points most influence the prediction.	*Validate explanations with storage-engine experts.*

METHODOLOGY

1. **Data Preparation:** For each disk, slide a 5-minute (20-step) window over the time-series, labeling windows that contain a fail-slow event within the next 10-minute as positive. 0 core metrics (latency, throughput, error rate, queue depth, etc.) → 20-dimensional input. Apply SMOTE on the windowed data or use class-weighting in the loss.

2. **Model Architecture**

    - **CNN Block**: 1-D Conv (kernel=3, 64 filters) → BatchNorm → ReLU → MaxPool (size=2). Repeated 3 times.
    - **RNN Block**: Bidirectional LSTM (hidden=128).
    - **Self-Attention Layer**: Multi-head (heads=4) over the LSTM output to capture cross-time dependencies.
    - **Dense Head**: 2-layer MLP (128→64→1) with sigmoid output.
    - **Regularization**: Dropout (0.3) and weight decay (1e-5).

3. **Training**

    - **Loss**: Binary cross-entropy with class weights.
    - **Optimizer**: AdamW (lr=1e-4).
    - **Scheduler**: Cosine annealing
    - **Early stopping**: on validation AUROC (patience=10)

4. **Evaluation**

    - **Metrics**: Precision, Recall, F1-score, AUROC, and Time-to-Alert (average number of minutes from the first abnormal window to the first positive prediction).
    - **Baselines**: XGBoost (best hyper-parameters from the original paper), Random Forest, and a simple LSTM.

- **Cross-validation**: 5-fold on the training set, final evaluation on a held-out 10% test set (Cluster A vs. Cluster B).

5. **Explainability**

   - **Attention Maps**: Visualize the attention weights per metric/time-point.
   - **SHAP**: Compute SHAP values for the final dense layer.
   - **Expert Validation**: Share maps with storage-engine experts to confirm that the model highlights the same metrics (e.g., latency spike) as human analysts.

6. **Reproducibility**

   - Data, Dockerfile, Jupyter Notebook, Model Zoo, Python Scripts

SIGNIFICANCE

This project will (1) fail-slow disks constitute the majority of downtime incidents in large data-centers; by raising AUROC score and reducing the time-to-alert by more than 3 minutes, the hybrid model enables operators to intervene well before a full failure occurs, thereby cutting repair costs and service-level-agreement penalties. The CNN–RNN–Attention architecture can process the 15-second PERSEUS streams in batches on commodity GPUs, making it feasible to deploy in production environments where thousands of disks generate data in parallel. The model's modest memory footprint ($\approx$ 30MB) allows it to run on edge monitoring agents or centralized analytics clusters without prohibitive hardware costs. Attention maps and SHAP values highlight the specific metrics (e.g., latency spikes, queue-depth growth) and time-points that drive a positive fail-slow prediction.

This interpret-ability is essential for gaining operator trust, satisfying regulatory requirements, and facilitating root-cause analysis when an alert is triggered. By training on a heterogeneous dataset that spans multiple clusters, disk models, and workloads, the proposed method demonstrates robustness to the variability that characterizes real world storage fleets. The framework can therefore be transferred to other storage technologies (e.g., NVMe SSDs, HDD arrays) with minimal retraining. The hybrid model is deliberately designed to align with existing production monitoring stacks (e.g., Prometheus, Grafana).

The paper provides a full, reproducible pipeline (code, Docker image, pre-processed dataset) that extends the FSA-Benchmark. This enables other researchers to replicate, compare, and build upon the results, fostering healthy scientific discourse and accelerating progress in storage-system anomaly detection. By offering a plug-and-play solution that outputs alerts and explainable dashboards, the research closes the gap between academic prototypes and operational tools used by system operators.