# Failure Forecasting in Multi-Brand HDDs Using Feature-Driven Supervised Learning

Rochit Madamanchi[1], Ritu Solanki[2], Ganavi S P[3], Peeta Basa Pati [4]
Department of Computer Science and Engineering, Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Bengaluru, India
[1]bl.en.u4cse23148@bl.students.amrita.edu,
[2]bl.en.u4cse232165@bl.students.amrita.edu,
[3]bl.en.u4cse23124@bl.students.amrita.edu,
[4]bp_peeta@blr.amrita.edu

*Abstract*—**Hard disk drive (HDD) failure within large-scale data centers can lead to catastrophic data loss, service interruptions, and maintenance costs. Anticipating such failures ahead of time is an essential process towards adopting proactive maintenance and system reliability. HDDs today are equipped with Self-Monitoring, Analysis, and Reporting Technology (SMART), which produces internal health indicators that can be used to predict failure. This project proposes a data-driven methodology for predicting HDD failure based on the publicly available Backblaze dataset, which contains daily SMART logs from thousands of drives from many manufacturers.**

*Index Terms*—**Hard disk drive (HDD), SMART attributes, failure prediction, machine learning, predictive maintenance, data imbalance, feature selection**

## I. Introduction

As the demand for data-intensive applications and cloud-based storage solutions continues to grow exponentially, ensuring the dependability and longevity of storage infrastructure—particularly hard disk drives (HDDs)—has emerged as a critical concern for data centers. In modern IT environments, an unexpected hard drive failure can lead to serious consequences, including prolonged service downtime, irreversible data loss, reputational damage, and significant financial losses. Despite advances in monitoring and alert systems, traditional drive health assessment techniques often struggle to detect impending failures, especially within large-scale deployments comprising thousands of drives from various vendors.

Contemporary HDDs are equipped with Self-Monitoring, Analysis, and Reporting Technology (SMART), which continuously monitors the internal health and performance parameters of the drive. These SMART attributes provide vital information about the drive's operational behavior, including indicators such as the Reallocated Sector Count, Uncorrectable Error Count, Power-On Hours, Temperature, Spin Retry Count, and several others. If analyzed correctly, these metrics can serve as precursors to failure, enabling proactive detection and preventive action before a critical breakdown occurs.

This project focuses on leveraging machine learning (ML) techniques to build an effective predictive model capable of forecasting HDD failures using SMART data. For this purpose, we utilize the Backblaze publicly available dataset, which contains daily SMART logs of tens of thousands of HDDs from multiple manufacturers, collected over several years. The core objective is to identify and rank the most informative SMART attributes that correlate strongly with failure patterns and then train supervised learning classifiers—such as Logistic Regression, Decision Trees, Random Forests, Gradient Boosting Machines, and Support Vector Machines—to distinguish between healthy and failing drives.

A major challenge encountered in this task is the highly imbalanced nature of the dataset, where instances of drive failure are exceedingly rare compared to normal operating records. This imbalance poses a significant hurdle for standard machine learning algorithms, which tend to favor the majority class. To address this, the project incorporates a variety of data resampling strategies, including oversampling (e.g., SMOTE), undersampling, and hybrid methods. Moreover, performance evaluation metrics such as Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC) are emphasized over mere accuracy, to ensure robust detection of rare but critical failure events.

By enabling precise and timely prediction of HDD failures, the proposed model aims to improve proactive maintenance strategies within data centers. Such advancements could significantly reduce unplanned outages, minimize data loss risks, and optimize the overall operational efficiency of storage infrastructures. Furthermore, the techniques and insights derived from this study can serve as a foundation for broader applications in predictive maintenance across diverse industrial, commercial, and IT systems.

Ultimately, this research not only addresses an urgent practical need in modern computing environments but also contributes to the growing field of intelligent fault detection and maintenance using machine learning and big data analytics.

## II. Literature Survey

Recent advancements in HDD failure prediction have shifted focus from threshold-based diagnostics to sophisticated machine learning and domain generalization techniques. Wang et al. [1] proposed a Heuristic-Invariant Risk Minimization (HIRM) framework that enhances generalization to out-of-distribution (OoD) data using SMART attributes, enabling better predictive accuracy for HDDs operating in novel or evolving environments. The HIRM method builds upon traditional IRM theory by incorporating heuristic strategies into domain division and training pipelines.

Han et al. [2] introduced ALAE (Adversarial LSTM-Autoencoder), a model designed for predicting failures in newly deployed HDD models that lack historical SMART data. By combining LSTM-based temporal modeling with adversarial domain generalization and a multi-domain MMD module, ALAE effectively aligns data distributions across different drive models and demonstrates improved cross-model prediction performance.

Liu et al. [3] proposed DCGAN-QP (Deep Convolutional GAN with Quadratic Potential), an unsupervised anomaly detection method for HDD failure prediction. The model uses adversarial learning with residual modules and a novel quadratic potential divergence loss to identify outliers in SMART attribute sequences, achieving strong results even under severe label sparsity and class imbalance.

De Santo et al. [4] addressed the challenge of estimating the Remaining Useful Life (RUL) of hard drives using a temporal LSTM-based architecture. Unlike binary classification models, their approach automatically identifies HDD health levels using regression trees and models sequential SMART attribute dependencies. Their framework successfully predicted failures up to 45 days in advance using both hourly and daily sampled datasets.

Pinciroli et al. [5] presented a comparative study on HDD and SSD reliability using six years of operational logs. They found that failure triggers vary between devices and proposed interpretable Random Forest classifiers. They demonstrated that data partitioning—based on drive-specific features such as head flying hours—can improve predictive accuracy, especially in highly imbalanced settings.

In the context of unlabeled and imbalanced data, Wang et al. [6] proposed a multi-instance LSTM model, treating HDD lifespan as a bag of instances rather than labeling each SMART reading. This approach improved failure prediction accuracy by exploiting sequence data while addressing issues related to short degradation periods and data imbalance.

To address uncertainty in degradation patterns, Wang et al. [7] introduced an adaptive Rao–Blackwellized particle filter method, modeling HDD degradation with a hybrid jump Markov process. The model dynamically adjusts prediction thresholds based on residual errors, outperforming traditional classifiers in early failure detection and reducing false alarms.

Miller et al. [8] provided a large-scale industry perspective from Meta, highlighting the roles of drive age and workload in HDD failure trends. Using XGBoost models trained on SMART metrics collected over 30-day windows, they identified that temporal changes in SMART attributes significantly improve failure prediction performance.

Hai et al. [9] addressed the complexity and imbalance issues of LSTM by introducing a GRU-based prediction framework with a TimeGAN module. The GRU architecture simplifies temporal modeling while TimeGAN generates synthetic samples to mitigate class imbalance, resulting in superior detection performance.

Finally, Pinciroli et al. [10] extended their previous work with a detailed six-year comparative study on HDDs and SSDs. Their machine learning models not only achieved high recall and low false positives but also provided interpretable insights by leveraging workload-aware partitioning of datasets, improving predictive reliability.

## III. Methodology

The following methodology describes the complete pipeline for HDD failure prediction using SMART attribute data from the Backblaze dataset

### A. HDD SMART Dataset Input

The experiments were conducted using the SMART dataset, which contains daily operational records of hard disk drives along with a binary failure indicator. Each record includes multiple numerical SMART attributes such as reallocated sector count, uncorrectable error count, power-on hours, and drive temperature. The binary target variable (*failure*) was used to distinguish healthy drives (0) from failing drives (1).

### B. Data Preprocessing Block

The dataset was imported from an Excel file and cleaned before analysis. Non-numeric and timestamp attributes were removed to retain only useful SMART statistics. Columns containing only missing values were discarded, and the remaining missing entries were imputed using the mean strategy. To preserve class balance information during training and evaluation, the dataset was stratified and split into training and test subsets with an 8:2 ratio, fixing a random seed for reproducibility.

### C. Model Selection and Training

A diverse set of machine learning classifiers was employed to capture the complex failure patterns. The models included Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost, Naïve Bayes (GaussianNB), and Multi-Layer Perceptron (MLP). Additionally, XGBoost and CatBoost were tested when available. Each classifier was embedded in a pipeline with feature scaling using StandardScaler to normalize numerical attributes. Hyperparameters for the models were tuned using `RandomizedSearchCV` with 5-fold stratified cross-validation, and accuracy was chosen as the optimization criterion.

### D. Evaluation Metrics

To comprehensively assess classification performance, several evaluation metrics were computed on both training and test subsets. These included accuracy, precision, recall, F1-score, and ROC-AUC. Given the strong class imbalance in the dataset, particular emphasis was placed on recall and F1-score, as they capture the ability to detect rare failure events without overemphasizing the dominant healthy class.

### E. Visualization and Result Reporting

The experimental framework produced summary tables for all classifiers, reporting cross-validation accuracy, training set performance, and test set performance across the chosen metrics. The results were sorted by test F1-score to prioritize models that balance sensitivity and precision in detecting HDD failures.

## IV. Results and Analysis

The dataset used for HDD failure prediction was highly imbalanced, with only a small number of failure samples (Train: 32 non-failures, 2 failures; Test: 8 non-failures, 1 failure). Several classification algorithms were evaluated using cross-validation, and their performance was measured on both training and test sets. To ensure readability in the two-column format, the performance results are divided into two tables (Table I and Table II). graphicx

TABLE I
Performance Comparison of Classifiers for HDD Failure Prediction (Part I)

| Model | CV Best | Train Acc | Train Prec | Train Recall | Train F1 | Train ROC-AUC |
|---|---|---|---|---|---|---|
| SVC | 0.943 | 0.971 | 1.000 | 0.500 | 0.667 | 0.000 |
| DecisionTree | 0.914 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| RandomForest | 0.943 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| AdaBoost | 0.943 | 0.971 | 1.000 | 0.500 | 0.667 | 1.000 |
| XGBoost | 0.943 | 0.941 | 0.000 | 0.000 | 0.000 | 0.500 |

TABLE II
Performance Comparison of Classifiers for HDD Failure Prediction (Part II)

| Model | Test Acc | Test Prec | Test Recall | Test F1 | Test ROC-AUC |
|---|---|---|---|---|---|
| SVC | 0.778 | 0.000 | 0.000 | 0.000 | 0.625 |
| DecisionTree | 0.778 | 0.000 | 0.000 | 0.000 | 0.438 |
| RandomForest | 0.778 | 0.000 | 0.000 | 0.000 | 0.188 |
| AdaBoost | 0.778 | 0.000 | 0.000 | 0.000 | 0.438 |
| XGBoost | 0.889 | 0.000 | 0.000 | 0.000 | 0.500 |
| CatBoost | 0.889 | 0.000 | 0.000 | 0.000 | 0.375 |
| GaussianNB | 0.667 | 0.000 | 0.000 | 0.000 | 0.375 |
| MLP | 0.778 | 0.000 | 0.000 | 0.000 | 0.125 |

Tables I and II summarize the performance of different classifiers for HDD failure prediction during the training and testing phases. Table I shows the cross-validation and training results, where models like DecisionTree and RandomForest achieved perfect scores across most metrics. In contrast, Table II presents the test results, highlighting the generalization performance of the same classifiers on unseen data, where the accuracies were moderate but precision, recall, and F1-scores were notably low, indicating challenges in correctly identifying failure cases.

## V. Conclusion

From the results, it is evident that most models achieved strong cross-validation performance (CV scores above 0.91) and nearly perfect training accuracy. However, this success did not translate to the test set: every model failed to identify the minority class (failures), resulting in zero recall and F1-scores. The tree-based models (Decision Tree, Random Forest, AdaBoost, and MLP) clearly overfitted, achieving perfect training scores but failing to generalize. The class imbalance dominated learning, causing models to default to the majority class. Overall, the experiments reveal that traditional classifiers are ineffective under extreme class imbalance. Despite high accuracy, they predict only the majority class, ignoring rare failures. This highlights the need for imbalance-aware strategies such as resampling, cost-sensitive methods, or anomaly detection for reliable HDD failure prediction.

## References

[1] J. Wang, R. Zhang, G. Qi, and L. Hong, "A Heuristic-IRM Method on Hard Disk Failure Prediction in Out-of-distribution Environments," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manag. (IEEM)*, 2021.

[2] G. Han, Y. Wang, Q. Hai, Z. Yang, and S. Peng, "Adversarial Domain Generalization Network for New Enabled Hard Disk Drive Failure Prediction," *IEEE Trans. Magn.*, early access, 2025.

[3] A. Liu, Y. Liu, Y. Deng, and S. Li, "DCGAN with Quadratic Potential for Hard Disk Failure Prediction in Operational Data Center," in *Proc. 6th Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, 2023.

[4] A. De Santo, A. Galli, M. Gravina, V. Moscato, and G. Sperlí, "Deep Learning for HDD Health Assessment: An Application Based on LSTM," *IEEE Trans. Comput.*, vol. 71, no. 1, pp. 69–82, Jan. 2022.

[5] R. Pinciroli, L. Yang, J. Alter, and E. Smirni, "Machine Learning Models for SSD and HDD Reliability Prediction," in *Proc. Annu. Rel. Maintainab. Symp. (RAMS)*, 2022.

[6] Y. Wang et al., "A Multi-Instance LSTM Network for Failure Detection of Hard Disk Drives," in *Proc. IEEE Int. Conf. Industrial Informatics (INDIN)*, 2020.

[7] Y. Wang, L. He, S. Jiang, and T. W. S. Chow, "Failure Prediction of Hard Disk Drives Based on Adaptive Rao–Blackwellized Particle Filter Error Tracking Method," *IEEE Trans. Ind. Inf.*, vol. 17, no. 2, pp. 913–925, Feb. 2021.

[8] Z. Miller, O. Medaiyese, M. Ravi, A. Beatty, and F. Lin, "Hard Disk Drive Failure Analysis and Prediction: An Industry View," in *Proc. IEEE/IFIP Int. Conf. Dependable Syst. Netw. - Supplemental Volume (DSN-S)*, 2023.

[9] Q. Hai, S. Zhang, C. Liu, and G. Han, "Hard Disk Drive Failure Prediction Based on GRU Neural Network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2022.

[10] R. Pinciroli, L. Yang, J. Alter, and E. Smirni, "Lifespan and Failures of SSDs and HDDs: Similarities, Differences, and Prediction Models," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 1, pp. 256–270, Jan./Feb. 2023.