# 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 process starts by loading the raw SMART attribute dataset, which contains daily health records of hard disk drives from multiple manufacturers. Each data point includes numerical SMART values such as reallocated sector count, uncorrectable error count, power-on hours, and temperature, along with a binary failure label (0 for healthy, 1 for failure).

### B. Data Preprocessing Block

The experiments were performed with the SMART dataset having different Self-Monitoring, Analysis, and Reporting Technology (SMART) attributes of hard disk drives. The dataset was read from an Excel file. For the regression analysis, a single attribute of dataset was chosen as the independent variable, and another attribute was used as the dependent variable. To preserve data integrity, the records with missing values in either of these fields were filtered out through a concatenation process. The cleaned dataset was next divided into training and test subsets in an 8:2 ratio using the split function by fixing a random seed to ensure reproducibility across experiments.

### C. Entropy and Gini Calculations

After cleaning and discretising the SMART dataset, the next step was to characterise the distribution of the target attribute (`failure`). The entropy of the target was calculated and yielded a value of 0.4262, which already indicates a class imbalance (there are far more "0" instances than "1"). To confirm this, the Gini index was also computed and resulted in a value of 0.1588.

### D. Information Gain and Attribute Selection

Information gain was calculated for every SMART attribute in order to determine which feature should be used as the root node of the decision tree. After removing all identifier attributes such as `serial_number`, the feature that produced the highest information gain was `smart_1_raw`. This attribute therefore represents the most informative split and was selected as the starting point for the tree construction.

### E. Decision Tree Training

A decision tree classifier (with entropy as impurity measure) was trained on the discretised SMART dataset. To avoid overfitting, the maximum depth of the tree was limited to 3. After training, the final model was exported using the `joblib` library so that it could be re-used for visualisation and further evaluation.

### F. Tree Visualisation

The trained model was reloaded and visualised in the form of a full decision tree. The diagram clearly shows `smart_1_raw` at the root node, followed by additional splits using other SMART attributes with smaller information gain values. The leaf nodes indicate the resulting class (0 = healthy, 1 = failure).

### G. Decision Boundary Plot (Two Features)

To provide an intuitive interpretation of the classifier, two of the most informative SMART attributes were used to retrain a simplified decision tree. A two-dimensional decision boundary plot was then generated. In this figure, only a small region of the feature space is classified as "failure", which is consistent with the fact that failures in the Backblaze dataset are relatively rare compared to non-failure cases.

## IV. Result and Analysis

The entropy of the target class (`failure`) was computed as 0.4262 and the Gini index as 0.1588, confirming that the dataset is highly skewed toward healthy drives.

Information gain was calculated for all SMART attributes after removing the identifier fields. Among all the evaluated attributes, `smart_1_raw` achieved the highest information gain and therefore became the root node of the decision tree.

Figure 1 shows the full decision tree produced by the ID3 algorithm. The model first splits the data based on the value of `smart_1_raw` and then performs additional splits based on other SMART parameters. Each leaf contains only a single class, which means the decision tree is able to perfectly separate the failure and non-failure classes on the training data.

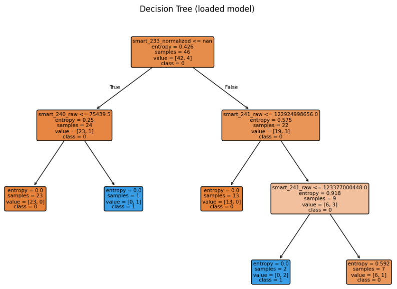

Fig. 1. Visualisation of the trained decision tree using entropy and information gain

Figure 2 illustrates the decision boundary using the two most informative SMART attributes. Only a narrow region of the feature space is classified as a failure, which is consistent with the limited number of failure samples present in the dataset.

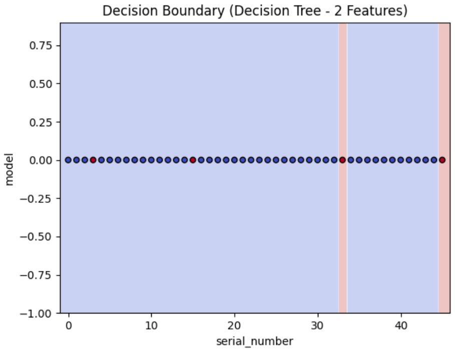

Fig. 2. Decision boundary obtained from a decision tree using the two most informative SMART attributes

## V. Conclusion

This work applies an entropy-based decision tree model for HDD failure prediction using SMART attribute data from the Backblaze dataset. After removing identifier attributes and discretising the continuous SMART values using equal-width binning, the entropy and Gini index confirmed that the dataset is highly imbalanced. Information gain calculations identified `smart_1_raw` as the most informative attribute for building the decision tree. The trained classifier was able to successfully separate failure cases from non-failure cases within the training data, and a two-dimensional decision boundary plot clearly illustrated the region associated with drive failure.

Although the current model provides interpretable rules for identifying failure cases, the limited number of failure samples indicates that future work should investigate imbalance handling strategies (e.g. SMOTE or cost-sensitive learning) to improve the detection of rare failure events.

## 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.