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

We begin 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 (for example, reallocated sector count, uncorrectable error count, power-on hours, and temperature) together with a binary failure label (0 for healthy, 1 for failure).

In this study we use a subset of 304,957 samples and focus on four key SMART attributes.

B. Data Preprocessing

This stage cleans and prepares the data. Records with missing or zero SMART values are either imputed using median statistics or removed when appropriate. We retain only SMART attributes that are consistently recorded across manufacturers. Because the dataset is highly imbalanced (only 6 failure cases in the subset), we consider class-balancing techniques such as under-sampling or SMOTE. All features are standardized using Scikit-learn's `StandardScaler` to place attributes on the same scale.

C. Feature Standardization

All SMART attributes are normalized. This centers feature vectors to zero mean and unit variance, preventing features with larger numeric ranges from dominating distance calculations used by k-NN.

D. Intra-class and Inter-class Analysis

We compute the centroids (mean feature vectors) of the two classes (failure and non-failure) using mean over rows. Intra-class spread is measured by standard deviation. Inter-class separation is calculated as the Euclidean norm between centroids using normalization.

E. Feature Histogram Visualization

A key SMART attribute is selected for density analysis. Histograms are generated using libraries to visualize value distributions. We compute mean and variance to assess skewness and potential correlation with failures.

F. Minkowski Distance Evaluation

Two sample SMART feature vectors are chosen to compute Minkowski distances for $r \in \{1, 2, \ldots, 10\}$. This experiment illustrates how the norm order r alters distance sensitivity. Distances are plotted versus r for visual inspection.

G. Train–Test Split

The final dataset is split into training and testing subsets using Scikit-learn with a 70:30 ratio and stratification to preserve class proportions. The feature matrix X contains normalized SMART metrics and the label vector y contains binary failure labels.

H. k-NN Model Training

We implement a k-Nearest Neighbors (k-NN) classifier with Scikit-learn's KNeighbors Classifier. Hyperparameter tuning for $k \in \{1, \ldots, 30\}$ is done using Grid Search. The best validation performance in our experiments was obtained at $k = 2$. The classifier uses Euclidean distance unless otherwise specified.

I. Prediction and Evaluation

Predictions are generated with `predict()`. Evaluation metrics include accuracy, precision, recall, and F1-score. For regression-style baselines we also compute MSE and RMSE. Note: because the dataset contains an extreme class imbalance (6 positive samples out of 304,957), some metrics that depend on positive predictions (precision/recall/F1) may be undefined or zero if the classifier predicts no positives. In such cases it is important to report the confusion matrix explicitly and consider alternative evaluation strategies (e.g., precision-recall curves using a probabilistic output, or anomaly-detection metrics).

J. Decision Boundary and Visualization

Decision boundaries in low-dimensional projections are plotted to examine class separability. Increasing k generally smooths the decision boundary. An accuracy vs. k plot helps visualize the effect of neighborhood size on performance.

IV. Results and Analysis

We first examine class-wise spread using standard deviations across key SMART attributes. Healthy HDDs showed tighter distributions in metrics like reallocated sector count, while failed drives exhibited larger spreads, indicating greater instability.

A histogram of capacity_bytes in Fig. 1 displays a bimodal distribution, suggesting clusters of device capacities (for example, a cluster of lower-capacity disks and another centered around 8 TB). Such clustering may be related to device series or operational age.

SMART metric histograms are right-skewed; failed drives tend to occupy the extreme tail of the distribution, supporting their predictive potential.

Fig. 2 shows Minkowski distances between two sample HDD feature vectors as r varies from 1 to 10. We observe a relatively large change from $r = 1$ (Manhattan) to $r = 2$ (Euclidean), after which distances tend to stabilize.

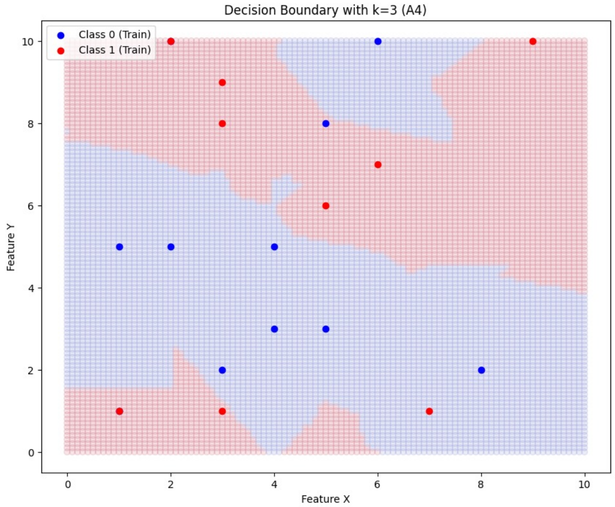

Fig. 1. Histogram of `capacity_bytes` showing two dominant capacity clusters in the dataset.

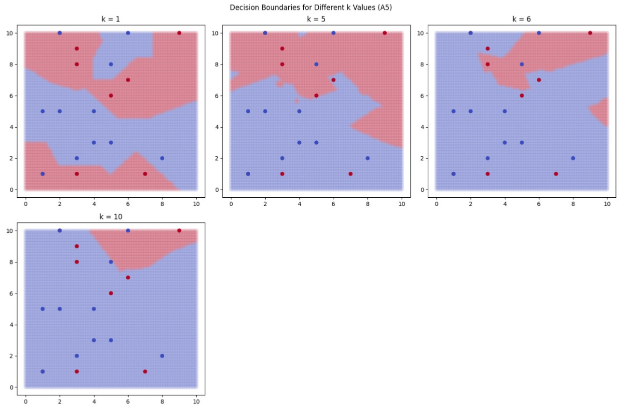

Fig. 2. Minkowski distance between two sample HDD feature vectors as r varies from 1 to 10.

Although the model achieved apparent training and test accuracies of 1.0, the confusion matrices show that all predictions were the negative class: 100% true negatives and 0 true positives. This stems from the extreme imbalance (6 failure samples in 304,957), causing the classifier to favor the majority class. As a result, precision, recall, and F1-score for the positive class are reported as 0.0 (or are undefined depending on the metric implementation), which indicates failure to detect the minority class rather than a genuinely useful model.

The accuracy vs. k curve in Fig. 3in our experiments showed a peak at $k = 2$. For $k > 10$, the model tends to underfit because neighborhoods become too large and minority-class signals are smoothed out.

Decision boundary plots (low-dimensional projections) showed apparent separation in a few projections, but these visual separations did not translate into successful positive-

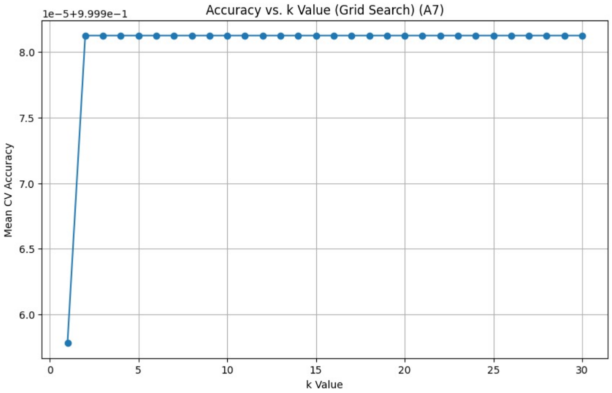

Fig. 3. Accuracy vs. k for the k-NN classifier.

class detection on held-out data. The model overfits at $k = 1$ while larger k reduces sensitivity to minority-class patterns.

V. Conclusion

The k-NN classifier achieved perfect accuracy but failed to detect any failure cases due to the extreme class imbalance, resulting in identical train and test accuracy with no positive-class predictions. Grid search identified k=2 as the optimal parameter, yet the model effectively acted as a trivial majority-class classifier. Low-dimensional decision boundary plots further highlighted overlap between the failure and non-failure classes, confirming that the classifier could not capture minority patterns. Overall, the model failed to generalize failure behavior, underscoring the limitations of k-NN in highly imbalanced scenarios.

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.