HDD Failure Prediction

P Sankeerthan Reddy, G Lohith Venkat Reddy, M Vivek Chara, Dr.Peeta Basa Pati
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse23038@bl.students.amrita.edu, bl.en.u4cse23011@bl.students.amrita.edu,
bl.en.u4cse23026@bl.students.amrita.edu, bp_peeta@blr.amrita.edu

***Abstract*—Hard Disk Drives (HDDs) remain at the forefront of contemporary storage infrastructure but are prone to mechanical failure resulting in catastrophic data loss. Predictive failure prediction is crucial to data integrity and system reliability. The aim of this project is to develop a failure prediction model for HDD from the publicly available Backblaze dataset containing SMART (Self-Monitoring, Analysis, and Reporting Technology) attributes of thousands of production disks. With the assistance of state-of-the-art machine learning techniques, including decision trees, ensemble methods, and recurrent neural networks, the target system will attempt to predict failing disks with low false positives. The project also investigates feature engineering methods for extracting valuable knowledge from SMART parameters and aims to achieve a trade-off between model interpretability and efficacy. This solution has the potential to enhance predictive maintenance of data centers, enabling timely intervention and reduced downtime.**

I. Introduction

Hard disk drive (HDD) is the ubiquitous data storage unit in data centers and desktops. Regardless of the sophistication of the storage technology, HDDs are prone to electronic and mechanical failure, resulting in heavy usage and economic loss. Predictive maintenance programs that predict impending failure can minimize such risks through pre-emptive replacement before catastrophic failure.

SMART attributes are valuable sensor information that reflect HDDs' health. Reallocated Sector Count, Spin Retry Count, and Current Pending Sector Count are effective predictors of failure. It is challenging, however, to develop effective prediction models due to the nature of SMART data, the skewed proportion of failed to healthy drive samples, and the dynamic nature of storage technology.

This project will construct a machine learning predictor of HDD failure from the Backblaze dataset, providing historical SMART attribute values for thousands of drives under representative workloads. Classification and Regression Trees (CART), ensemble approaches like Random Forests, and time models like Long Short-Term Memory (LSTM) networks will be employed to detect at-risk drives. System performance will be approximated from precision, recall, F1-score, and early warning ability with low false alarms. With regard to considering the interpretability of tree models and time sensitivity of deep learning approaches, this project will provide a strong and scalable predictive maintenance solution for storage systems.

II. Literature Survey

Li et al. (2014) [1] implemented data-driven method in the form of Classification and Regression Trees (CART) to forecast hard drive failure based on SMART attributes. CART models were trained on historical SMART data for healthy or failing drive prediction. CART models' greatest strength is interpretability along with acceptable performance relative to non-linear variable relationships. The authors went on to say that the strength of decision tree-based models in real applications is that they are interpretable, but more compromising in accuracy compared to more sophisticated models.

Chaves et al. (2018) [2] proposed Bayesian Network-based failure prognosis model wherein probabilistic interaction among SMART parameters was used to quantify failure probability. The methodology is best suited for parameter interaction and uncertainty modeling and hence operated even in the absence of sensor data. Simulation experiments with real data showed Bayesian inference was able to give consistent and interpretable predictions with detection and identification of the cause of failure in HDDs.

Wang et al. (2014) [3] also proposed a two-stage prediction model with screening using threshold-based values and followed by logistic regression. The drives with high SMART attributes exceeding pre-specified thresholds were selected in the first stage for subsequent testing. In the second phase, failure probability was predicted through a regression model. The hybrid scheme improved specificity and sensitivity and provided early warning ability with little false alarm, a necessity in large-scale data center roll-out.

Züfle, et al. (2021) [4] contrasted holistic data-driven methodologies for incrementally retraining machine learning models to predict HDD failure over some future time horizon and contrasted the model obsolescence phenomenon. They demonstrated how ever-green models deteriorate with changing drive technologies and usage patterns over a period of time. By incremental learning and iteratively retraining, they testified the predictability of accuracy and generalizability to balloon and that necessitates periodic updating of models.

Tomer et al. (2021) [5] created feature analysis of SMART attributes and determination of where they are most significant to failure prediction using classic machine learning methods. They validated Reallocated Sector Count, Spin Retry Count, and Current Pending Sector Count as primary metrics in their

work. The work also tackled issues with feature selection of significant features and model fine-tuning in order to obtain a balance between high accuracy and low false positives, particularly useful in sensitive storage applications.

McLean and Sterritt (2025) [6] carried out a comparative study among various supervised machine learning models such as Support Vector Machines, Random Forests, and Neural Networks. They experimented on benchmark SMART datasets and compared their performance based on precision, recall, and F1-score. The conclusion of the study was that ensemble approaches, i.e., Random Forests, gave the best performance in terms of strength with interpretability. Their study is suggestive towards the choice of suitable algorithms keeping system specifications in mind.

Yang et al. (2015) [7] described a massive data-supported HDD failure prediction system that is highly effective in handling gigantic volumes of SMART data. On distributed computing systems like Hadoop, they demonstrated the effectiveness of executing machine learning algorithms on a very large number of drives. Their solution focused on scalability and efficiency needed in today's cloud and enterprise storage platforms.

Zhao et al. (2020) [8] presented a SMART customization method for early failure detection. Through feature engineering on raw SMART values, they obtained faint patterns most likely leading to disk failure but perhaps not detectable by general SMART thresholds. Their method served as an early warning and reduced false alarms with detection of device-specific patterns leading to failure. The customized method worked well with various HDD models under varied workloads.

Lima et al. (2018) [9] also tested the ability of Recurrent Neural Networks (RNNs), in this case, Long Short-Term Memory (LSTM) networks, to learn sequential temporal trends from SMART data. This approach yielded better results than standard classifiers when time-series behavior was an issue, e.g., degrading over time. This paper was useful in emphasizing the role of learning temporal dynamics towards predictive maintenance action.

Strom et al. (2007) [10] described one of the early statistical HDD reliability models with focus on failure rate, life distributions, and environmental. They had also used reliability engineering methods earlier to quantify failure behavior in various HDD models. The paper did not include machine learning but set the stage for future data-driven methodology on the basis of defining leading factors and trends of HDD failure and baseline models for comparison with.

III. Methodology

We focused on hard disk drive (HDD) failure prediction using the Backblaze dataset and its SMART (Self-Monitoring, Analysis, and Reporting Technology) attributes in this study. Since the dataset is very large, containing records of many HDD models across several years, it was necessary to prefilter the data to make it more manageable while retaining important information needed to train predictive models. The filtering process targeted two of the most frequently occurring HDD models in the dataset: **ST12000NM0007** and **ST14000NM0138**. These models were selected because they provide a sufficient number of instances for reliable failure prediction, including both healthy and failed states.

A Python script was created to automate the data filtering process. This script scanned the working directory for daily SMART log files in CSV format. For each file with a `.csv` extension, it attempted to load the contents into memory using the `pandas` library. The parameter `low_memory=False` was specified to ensure that large files could be read without causing `dtype` inconsistencies.

After successfully reading each file, the script checked for the presence of a column named `model`, which is necessary for identifying the HDD models. If this column was present, the script filtered rows corresponding to the two selected target models. Feedback was provided during execution by printing the number of matching rows found in each file, allowing progress tracking. Files that lacked the `model` column or did not contain relevant records were discarded.

The cleaned rows from each file were appended to a list, which was later combined into a single DataFrame. This unified DataFrame represented the merged data for the chosen HDD models across all dates. Finally, the aggregated data was exported to a new CSV file named `top_2_hdd_models.csv`, which served as the input for subsequent steps involving preprocessing, feature engineering, and model development.

This filtering approach allowed us to significantly reduce the dataset size while retaining a representative and balanced subset containing sufficient failure instances. This was crucial for developing a robust and accurate machine learning model for HDD failure prediction.

IV. Result Analysis and Discussion

a. Filtering of Backblaze Dataset for Target HDD Models

To ensure efficient model development, the original Backblaze SMART dataset was filtered to include only the two most frequently occurring HDD models: `ST12000NM0007` and `ST14000NM0138`. A Python script was used to iterate through multiple daily CSV logs and extract only the rows containing these models. After processing, the filtered dataset contained approximately **[insert number]** rows spanning multiple time periods and device IDs. This reduced dataset served as the foundation for failure prediction modeling. The filtering step not only decreased computational overhead but also ensured a balanced distribution of failure and non-failure records, which is critical for training reliable classification models.

b. Rank of the SMART Feature Matrix

The SMART attribute matrix, formed by extracting relevant features for each HDD over time, was evaluated for rank. A full-rank observation matrix indicates that all feature vectors are linearly independent, which is important for stable model training. If the matrix were rank-deficient, it would imply multicollinearity or redundancy among features, potentially

leading to instability in model coefficients and poor predictive performance. Ensuring that the SMART matrix has full rank is essential for interpretable and accurate machine learning results, especially when using algorithms sensitive to feature overlap like logistic regression or SVM.

c. Creating a Predictive Model for HDD Failure

Objective: Build a classification model capable of predicting HDD failure in advance, using SMART attribute trends over time.

Data Preprocessing:

- Filter only relevant SMART attributes and drop columns with excessive missing values.
- Label data points as 1 (failed) or 0 (healthy), optionally applying a window to label imminent failures.
- Apply rolling statistics (e.g., moving average, delta, standard deviation) for each feature to capture trends.
- Normalize or scale data to stabilize model convergence.

Model Selection:

- Binary classification models such as Logistic Regression, Random Forest, XGBoost, and Support Vector Machines were considered.
- Time-series-aware models like LSTM may be used to incorporate sequential behavior of SMART attributes.

Evaluation Metrics:

- Due to class imbalance, metrics such as Precision, Recall, F1-Score, and ROC-AUC were prioritized over Accuracy.
- Confusion matrices were used to visualize performance on failure prediction.

Model Adaptability: Given that SMART behavior may change across HDD models and manufacturers, the model should be retrainable or fine-tuned to accommodate evolving failure patterns in new datasets.

REFERENCES

[1] J. Li, et al., "Hard Drive Failure Prediction Using Classification and Regression Trees," Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Atlanta, GA, USA, pp. 383–394, 2014. doi: 10.1109 DSN.2014.44.

[2] I. C. Chaves, M. R. P. de Paula, L. G. M. Leite, J. P. P. Gomes, and J. C. Machado, "Hard Disk Drive Failure Prediction Method Based On A Bayesian Network," Proceedings of the International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, pp. 1–7, 2018. doi: 10.1109/IJCNN.2018.8489097.

[3] Y. Wang, E. W. M. Ma, T. W. S. Chow, and K.-L. Tsui, "A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives," IEEE Transactions on Industrial Informatics, vol. 10, no. 1, pp. 419–430, 2014. doi: 10.1109/TII.2013.2264060.

[4] M. Züfle, F. Erhard, and S. Kounev, "Machine Learning Model Update Strategies for Hard Disk Drive Failure Prediction," Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, pp. 1379–1386, 2021. doi: 10.1109/ICMLA52953.2021.00223.

[5] V. Tomer, V. Sharma, S. Gupta, and D. P. Singh, "Hard Disk Drive Failure Prediction Using SMART Attribute," Materials Today: Proceedings, vol. 46, no. 20, pp. 11258–11262, 2021. doi: 10.1016/j.matpr.2021.03.229.

[6] A. McLean and R. Sterritt, "Hard Disk Drive Reliability: A Comparative Study of Supervised Machine Learning Algorithms for Predicting Drive Failure," Proc. of The 21st Int. Conf. on Autonomic and Autonomous Systems, Lisbon, Portugal, pp. 8–14, Mar. 2025.

[7] W. Yang, D. Hu, Y. Liu, S. Wang, and T. Jiang, "Hard Drive Failure Prediction Using Big Data," Proceedings of the IEEE 34th Symposium on Reliable Distributed Systems Workshop (SRDSW), Montreal, QC, Canada, pp. 13–18, 2015. doi: 10.1109/SRDSW.2015.15.

[8] J. Zhao, et al., "Disk Failure Early Warning Based on the Characteristics of Customized SMART," Proceedings of the 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), Orlando, FL, USA, pp. 1282–1288, 2020. doi: 10.1109/ITherm45881.2020.9190324.

[9] F. D. S. Lima, F. L. F. Pereira, I. C. Chaves, J. P. P. Gomes, and J. C. Machado, "Evaluation of Recurrent Neural Networks for Hard Disk Drives Failure Prediction," Proceedings of the 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, pp. 85–90, 2018. doi: 10.1109/BRACIS.2018.00023.

[10] B. D. Strom, S. Lee, G. W. Tyndall, and A. Khurshudov, "Hard Disk Drive Reliability Modeling and Failure Prediction," IEEE Transactions on Magnetics, vol. 43, no. 9, pp. 3676–3684, 2007. doi: 10.1109/TMAG.2007.902969.