Feature Selection for Remaining Useful Life Prediction in Hard Disk Drives with Missing Data

Gabriel L. S. Felix, Francisco L. F. Pereira, Francisco D. B. S. Praciano, João P. P. Gomes, Javam C. Machado

[1]LSBD - Universidade Federal do Ceará, Brazil

{gabriel.felix, lucas.falcao, daniel.praciano}@lsbd.ufc.br,

{joao.pordeus, javam.machado}@lsbd.ufc.br

Abstract. *This paper proposes a two-stage feature selection approach for the problem of Remaining Useful Life (RUL) prediction in Hard Disk Drives (HDDs) with missing data. First, a wrapper method is employed, utilizing a regression estimator to identify the most informative features for RUL prediction. The selected feature set is then evaluated in the second stage using a neural network model, with a focus on assessing the imputation performance for missing data. The goal is to determine a feature subset that enhances RUL prediction accuracy and exhibits robustness in handling missing data scenarios. This approach addresses the challenges of missing data and provides insights into the most relevant features for accurate RUL prediction.*

1. Introduction

Hard disk drives (HDDs) are the primary storage solution in organizational IT environments and in common household usage. Despite advancements in technology, HDD failures remain a significant concern, often leading to severe disruptions, data loss, and costly downtime. The ability to predict and prevent these failures has become a critical aspect of maintaining the reliability and integrity of IT systems [Schroeder and Gibson 2007].

Several machine learning-based techniques for HDD failure prediction have been proposed [Li et al. 2014, Xu et al. 2016], and many have focused on using deep neural networks [Hu et al. 2020, Cahyadi and Forshaw 2021, Pereira et al. 2022]. Most models use Self-Monitoring, Analysis, and Reporting Technology features (SMART) as input, a set of sensor data and counters from within the disk. However, a manufacturer can report dozens of attributes, which can increase the complexity of a machine-learning model based on it.

One of the challenges in analyzing HDD failures is missing data. In real-world scenarios, IT environments and household disks often lack comprehensive and continuous monitoring systems, resulting in incomplete datasets that weaken failure prediction models. Therefore, addressing the issue of missing data is crucial for building robust models that can reliably predict HDD failures.

The present work aims to focus on a feature selection approach to SMART attributes that deals with predicting the imminent failure of HDDs while also taking into account the missing aspect of its data.

2. Background

2.1. SMART

The conventional method for diagnostic monitoring in hard disk drives involves the use of SMART technology [Ottem and Plummer 1995]. This technique relies on a set of attributes defined by the HDD manufacturer, which include sensors and error counters. By setting thresholds for these attributes, the manufacturer can identify an imminent failure when those thresholds are reached. In this technology, each attribute/metric has at least the following information: id, name, raw value, normalized value, and the threshold value. The first two, id and name, are used to uniquely identify each of the metrics, the name being a human-readable description to indicate what each metric represents.

In the collected information, the raw value is the metric's current value as reported by the drive. Depending on the metric, it can be a numerical value or a specific code. The raw value often represents a specific count, such as the number of reallocated sectors or seek errors. On the other hand, the normalized value is derived from the raw value. It provides a scaled representation of the metric current status. Hence, lower values indicate potential problems or degraded performance. In this sense, we have the threshold value that is used as an indication of the tipping point by manufacturers; that is, if the raw or normalized value exceeds the threshold, it indicates a potential issue or a warning condition.

In [Pinheiro et al. 2007], the authors presented an analysis of the HDDs in operation at Google and concluded that the following metrics are good predictors: **5, 187, 197 and 198**. In another study, [Amram et al. 2021] performed an approach to try to interpret and understand the underlying failure mechanisms. In a similar way, they concluded that SMART **3, 5, 7, 187, 197** and **198** are good predictors.

2.2. HDD Failure Prediction

Although the emergence of SMART has brought a breakthrough in HDD failure prediction, studies conducted by [Murray et al. 2005] and [Pinheiro et al. 2007] have indicated that relying on these thresholds for predictions leads to low accuracy. These studies have concluded that SMART data has limited usefulness in anticipating disk failure. Consequently, the main shortcomings of the SMART threshold approach are its limited ability to detect faulty disks and its narrow focus on near-term failures.

To address these issues and improve failure prediction, researchers have approached the problem in different ways. One approach involves treating the problem as a classification task, where day intervals from the Remaining Useful Life (RUL) are considered distinct health levels or classes [Lima et al. 2021]. Another approach involves directly treating the problem as a regression on the RUL [Lima et al. 2018]. In this work, we will adopt the second approach. Specifically, our model will predict the remaining days until each hard drive fails on a given day.

Traditional Deep Neural Networks are powerful tools to make good HDD failure predictions; however, they need the data to be continuously sampled at regular intervals, which doesn't always happen. In addition to the fact that individual users are unlikely to run frequent checkups, these tests can be quite costly, forcing checkups to be less frequent, which results in datasets with irregular sample frequency and missing data [Pereira et al. 2022].

3. Methodology

This paper aims to find a set of features that can be used to properly predict the RUL of HDDs while also considering the natural missing aspect of the data. Following the trend of research of recent years [Pereira et al. 2022], this study aims to find a set of features to be used by a Recurrent Neural Network (RNN) on time series data.

However, performing the usual feature selection techniques, such as wrappers, on an RNN would be hugely time-consuming. Therefore, we are proposing to perform a two-phase feature selection in which we first select feature sets using other predictive models and then evaluate their performance to predict the Remaining Useful Life (RUL) and the missing data in the final RNN.

For the first stage, we performed a sequential wrapper feature selection on the regression task of predicting the exact RUL of a disk in days. In this approach, the algorithm iteratively finds a new feature that maximizes a cross-validated prediction score when an estimator is trained on this chosen feature.

For the second stage, we employed a RNN model to predict, again, the RUL, and the data that will be used for imputation in missing data. This model takes into account the history of SMART values to make predictions of future values.

4. Experiments and results

4.1. Dataset

To compare the performance of each subset of SMARTs chosen by feature selection, we used a publicly available dataset provided by the Backblaze Company [Backblaze 2023]. This dataset comprises daily records of SMART attributes for multiple HDDs from April 2013 to December 2022. These records include the HDD's serial number, model, SMART attributes, and a label indicating whether it has failed or displayed any indicators of potential failure.

For this particular study, we specifically chose the Seagate ST4000DM000 model, which had the largest volume of data among all HDD manufacturers. We focused solely on serial numbers corresponding to failed HDDs. Our analysis was limited to the final 360 days of the HDDs' lifespan, and we excluded serial numbers that had incomplete daily records without a faulty label or disks with data measurements recorded after being identified as faulty. After implementing these processes, our dataset comprised 1,631,802 daily records distributed across 4,936 serial numbers.

4.2. Experimental procedure

For each stage of feature selection, we use raw values of SMARTS from a set of HDDs in which every disk is a time series. Of these disks, 60% are for the train, 20% for the validation, and 20% for the test.

Upon analyzing the preprocessed dataset, we observed that the vast majority of the samples, although not all, consist of a 360-day observation period. In order to ensure balance within our dataset, we randomly divided each training sample into a maximum of three parts, ensuring that each part contained a minimum of 30 days relative to its size. At intervals of 120 days, we introduced an additional partition and combined them to

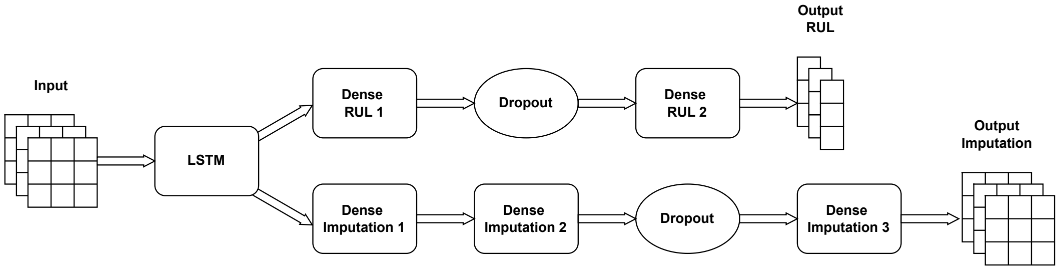

Figure 2. Predictive model. The input is the time series for each HDD.

generate new samples, as illustrated in Figure 1. Therefore, the dataset now has 16,199 training samples, 976 validation samples, and 1032 test samples. The RNN model utilizes the validation set as a criterion for early stopping.

For the first stage of our feature selection, we have chosen the ARD, Bayes Ridge, Lars, Lasso Lars IC, ordinary least squares Linear regression, Lasso, Lasso Lars, Quantile, Gamma, Huber, Elastic, OMP, Ransac, SGD, Poisson, Tweedie, Decision Tree (DT), and Extra Tree from the Scikit Learn library. This stage used the models with their default hyperparameters and the Sequential Feature Selector wrapper method.

For the second stage of our feature selection, we used a model composed of an LSTM, whose output is replicated to two branches, the RUL prediction branch, and the imputation branch. The RUL prediction branch is composed of two dense layers and a dropout layer, and its output is normalized with min-max. The imputation branch is composed of three dense layers and a dropout layer. The model flowchart is shown in Figure 2, and its hyperparameters, which went through a grid search process, can be consulted in Table 1. All experiments were run using Scikit-Learn 1.0.2 version and TensorFlow 2.8.0 version.

Hyperparameter	Value
Epochs	5,000
Learning Rate	0.001
Batch Size	1024
Early Stopping Patience	300
Optimizer	Adam
LSTM Cell Dimension	128
Dense Imputation 1 Output Dimension	128
Dense Imputation 2 Output Dimension	64
Dense Imputation 3 Output Dimension	12
Dense RUL 1 Output Dimension	64
Dense RUL 2 Output Dimension	1
All Dense Imputation Activation	Relu
All Dense RUL Activation	
L1 and L2 Regularization	0.00001
All Dropout Rate	0.1

Table 1. Predictive model hyperparameters.

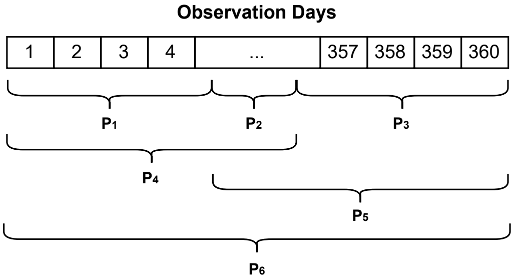

Figure 1. Days partition for the training dataset.

4.3. Results

For the first stage, the results of the models are listed in Table 3. The models ARD, Bayes Ridge, LARS, Lasso Lars IC, and ordinary least squares Linear regression gave the same set of features; therefore, we call them **Set 1**. Also, Lasso, Lasso Lars, Quantile, and Gamma regression are being named **Set 2** for the same reason.

These subsets are tested on the aforementioned neural network to evaluate these attributes in the prediction with missing data problems. To compare and choose the best set of features, the validation loss of the RUL prediction and the imputation prediction for each feature subset is being compared in Figure 3.

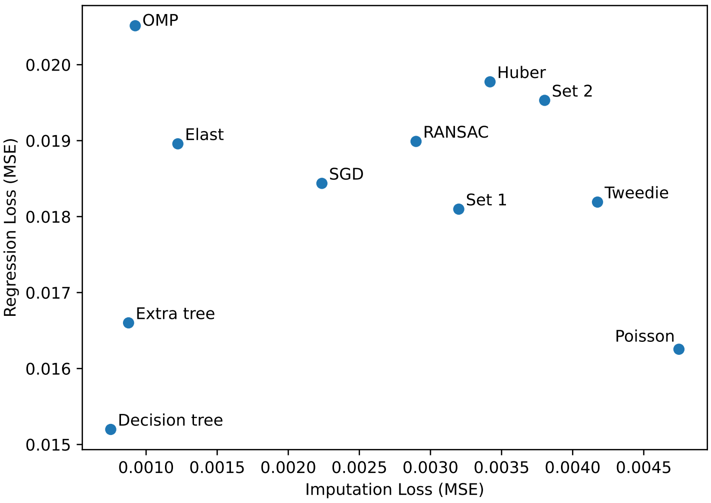

Figure 3. Validation losses for the feature subset discovered by several methods.

Model	MSE Regression Loss	MSE Imputation Loss
Decision Tree	**0.01527**	0.000761
Extra Tree	0.01630	**0.000692**
RANSAC	0.02055	0.000699
Set 1	0.01776	0.000734
SGD	0.01852	0.000967
Huber	0.01976	0.001078
OMP	0.01979	0.006306
Tweedie	0.01937	0.006813
Elast	0.01912	0.007327
Poisson	0.01717	0.013103
Set 2	0.02011	0.018231

Table 2. Test loss for each feature set.

As one can see, the feature subset with the overall best result for the imputation and RUL prediction validation losses is the one defined by the Decision Tree (DT) method. Therefore, the result for our feature selection method are the features **1** (Read

Error Rate), **5** (Reallocated Sectors Count), **9** (Power-On Hours), **10** (Spin Retry Count), **184** (End-to-End error), **188** (Command Timeout), **191** (G-sense Error Rate), **192** (Unsafe Shutdown Count), **193** (Load Cycle Count), **197** (Current Pending Sector Count), **240** (Head Flying Hours), and **242** (Total LBAs Read). Table 2 show the loss results on the test dataset.

SMART	Estimators										
	Set 1	Set 2	Huber	Elastic	OMP	Ransac	SGD	Poisson	Tweedie	DT	Extra Tree
1		X		X	X					X	X
3		X		X	X	X					
4		X		X	X						X
5	X	X	X	X	X	X	X	X	X	X	
7	X	X	X	X	X						X
9	X	X	X	X	X		X	X	X	X	X
10		X		X	X	X				X	
12	X	X	X	X	X	X	X	X	X		X
183		X		X	X			X	X		
184	X	X	X		X	X	X	X	X	X	
187	X	X	X			X	X	X	X		
188		X			X	X				X	X
189					X						X
190	X		X				X	X	X		
191						X				X	X
192						X				X	X
193	X		X				X	X	X	X	X
194						X		X	X		
197	X		X				X	X		X	X
198							X				
199			X			X					
240	X		X	X			X	X	X	X	X
241	X		X	X			X	X	X		
242	X			X		X	X		X	X	

Table 3. SMARTs selected by each estimator.

5. Conclusion

In this article, we used a new approach to obtain the best set of features for the RUL prediction and imputation in HDDs with missing data. The losses comparison for each proposal allowed us to conclude that the set composed by features **1**, **5**, **9**, **10**, **184**, **188**, **191**, **192**, **193**, **197**, **240**, and **242** has the best trade-off between the analyzed tasks.

In future work, we are going to study the generality of the selected features for others HDDs manufacturer models.

Acnowledgment

This research was partially funded by Lenovo, as part of its R&D investment under Brazilian Informatics Law, by CAPES grants 88887.609134/2021-00 and 88887.609129/2021, and CNPQ grants 307323/2022-6 and 316729/2021-3.

References

Amram, M., Dunn, J., Toledano, J. J., and Zhuo, Y. D. (2021). Interpretable predictive maintenance for hard drives. *Machine Learning with Applications*, 5:100042.

Backblaze (2023). Hard drive data and stats. https://www.backblaze.com/b2/hard-drive-test-data.html. Accessed: 2023-02-13.

Cahyadi and Forshaw, M. (2021). Hard disk failure prediction on highly imbalanced data using lstm network. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3985–3991.

Hu, L., Han, L., Xu, Z., Jiang, T., and Qi, H. (2020). A disk failure prediction method based on lstm network due to its individual specificity. *Procedia Computer Science*, 176:791–799. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

Li, J., Ji, X., Jia, Y., Zhu, B., Wang, G., Li, Z., and Liu, X. (2014). Hard drive failure prediction using classification and regression trees. In *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 383–394. IEEE.

Lima, F. D. S., Pereira, F. L. F., Chaves, I. C., Gomes, J. P. P., and Machado, J. C. (2018). Evaluation of recurrent neural networks for hard disk drives failure prediction. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 85–90. IEEE.

Lima, F. D. S., Pereira, F. L. F., Chaves, I. C., Machado, J. C., and Gomes, J. P. P. (2021). Predicting the health degree of hard disk drives with asymmetric and ordinal deep neural models. *IEEE Transactions on Computers*, 70(2):188–198.

Murray, J. F., Hughes, G. F., and Kreutz-Delgado, K. (2005). Machine learning methods for predicting failures in hard drives: A multiple-instance application. *J. Mach. Learn. Res.*, 6:783–816.

Ottem, E. and Plummer, J. (1995). Playing it smart: The emergence of reliability prediction technology. Technical report, Technical report, Seagate Technology Paper.

Pereira, F. L. F., Bucar, R. C. B., Brito, F. T., Gomes, J. a. P. P., and Machado, J. C. (2022). Predicting failures in hdds with deep nn and irregularly-sampled data. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part II*, page 196–209, Berlin, Heidelberg. Springer-Verlag.

Pinheiro, E., Weber, W.-D., and Barroso, L. A. (2007). Failure trends in a large disk drive population. In *5th USENIX Conference on File and Storage Technologies (FAST 07)*, San Jose, CA. USENIX Association.

Schroeder, B. and Gibson, G. A. (2007). Understanding disk failure rates: What does an mttf of 1,000,000 hours mean to you? *ACM Transactions on Storage (TOS)*, 3(3):8–es.

Xu, C., Wang, G., Liu, X., Guo, D., and Liu, T.-Y. (2016). Health status assessment and failure prediction for hard drives with recurrent neural networks. *IEEE Transactions on Computers*, 65(11):3502–3508.