2022 IEEE 18th International Conference on Automation
Science and Engineering (CASE)
August 20-24, 2022. Mexico City, Mexico

2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) | 978-1-6654-9042-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CASE49997.2022.9926720

Classification Based Hard Disk Drive Failure Prediction: Methodologies, Performance Evaluation and Comparison

Ruiyu Xu, Xinming Wang and Jianguo Wu*

Abstract— **Considering the reliability of the data storage system, it is essential to accurately and timely predict impending failures of hard disk drives (HDDs) so as to prevent data loss and reduce recovery cost. Over the past decades, taking as input the SMART (Self-Monitoring, Analysis and Reporting Technology) attributes, many supervised machine learning based methods have been proposed for HDD failure prediction. However, these methods are conducted on different datasets or different preprocessing treatments and thus lack comparative analysis. To fill this gap, we provide a systematic study in this paper on three key steps of the failure prediction, i.e., feature selection strategies, data preprocessing treatments and classification models. A feature selection strategy is proposed by testing the significance of difference between healthy and failed samples. Data relabeling, together with some other data preprocessing treatments are applied and proven to be effective in the case study. The performance of seven classification models are compared, among which the Random Forest model achieves the best performance with 53.95% failure detection rate (FDR) and 6.0% false alarm rate (FAR). Moreover, the Gini importance of SMART attributes is calculated, where two attributes, SMART 197 and SMART 187 are found closely related to the HDD failures.**

I. INTRODUCTION

With the rapid development of social informatization, the data storage volume has been growing explosively, which puts forward higher requirements for the reliability of data storage system. Among all components of the storage system, hard disk drives (HDDs) have a higher annual failure rate e.g., 2% to 4%, and thus are replaced more frequently [1], [2]. For instance, about 78% of the hardware replacements result from the hard drive failures in the data centers of Microsoft [3]. The hard drive failures could cause temporal unavailability or even permanent loss of stored data, posing a great threat to the security and reliability of storage system. If a potential failure of the drive could be predicted, timely data backup and hardware replacement can be done to prevent data loss and reduce recovery cost.

A standard monitoring system named SMART (Self-Monitoring, Analysis and Reporting Technology) has been widely applied to HDDs since 1995 [4]. In order to predict potential hardware failures, SMART detects and reports various attributes of drive's health condition. Typical attributes include power-on hours, temperature, read error rate, reallocated sectors count and current pending sectors count. In analysis stage, a will-fail warning would be issued if any single attribute value exceeds its predefined threshold. These thresholds are set conservatively to achieve a low false alarm rate (0.1%). As a result, the failure detection rate is extremely low (3%-10%) [5], which is far from satisfactory.

To simultaneously improve the failure detection rate and keep a low false alarm rate, several statistical and machine learning based methods have been proposed. In this paper, we focus on the supervised classification methods. It is worth mentioning that other failure prediction methods, such as the remaining useful life (RUL) prediction models in [6], [7], could also be applied to HDD failure prediction problem.

Hughes et al. [4] were the first to focus on the hard drive failure prediction problem. Since they found that many of the SMART attributes are non-parametrically distributed, they proposed two algorithms using distribution-free statistical hypothesis tests, namely multivariate rank-sum test and ORing single attribute rank-sum test. These two algorithms both achieved a failure detection rate around 60% with a 0.5% false alarm rate on a dataset containing 3744 drives. Since then, many machine learning methods were applied to improve the prediction performance, including the Bayesian method NBEM (Naive Bayes-Expectation-Maximization) and supervised naive Bayes classifier in [4], rank-sum method and support vector machines (SVMs) in [9], multiple-instance naive Bayes (mi-NB) in [5], hidden Makov models (HMM) and hidden semi Markov models (HSMM) in [10], a two-step parametric method in [11] and Gaussian Mixture Model in [12].

However, the above methods are designed for a small amount of drives, which are not suitable to today's big data environment. More recent researches focus on establishing prediction models on large datasets. Zhu et al. [13] collected the data of 23395 drives from Baidu data center, and proposed a model based on a neural network, whose failure detection rate could be raised up to 95% with a reasonable low false alarm rate of 0.48%. Li et al. [14] compared Decision Tree (DT) model and Gradient Boosted Regression Tree (GBRT) model on the same dataset. Yang et al. [15] further expanded the size of the dataset, on which a simple logistic regression method achieved a high failure detection rate. Based on other large datasets, Li et al. [16] developed a classification and regression tree model. Xu et al. [17] introduced a novel method based on Recurrent Neural Networks (RNN) to estimate a health degree of the HDDs. Xu et al. [18] developed a ranking-based model, learning characteristics in the past and estimating the error proneness in the near future. Ganguly et al. [19] combined decision tree model and logistic regression model.

*Corresponding author: Jianguo Wu
The authors are with the Department of Industrial Engineering and Management, Peking University, Beijing 100871, China (xuruiyu@pku.edu.cn, wang-xm20@stu.pku.edu.cn, j.wu@pku.edu.cn)

TABLE I

THE DATASETS, ALGORITHMS, PREDICTION PERFORMANCE OF EXISTING RESEARCHES.

Dataset size	Existing researches	Algorithms	Performance (FAR,FDR)
1936 drives	Hughes et al. (2000)	Rank-sum tests	(0.5%, 60%)
	Hamerly and Elkan (2001)	NBEM	(0.67%, 33%)
		naive Bayes classifier	(0.82%, 56%)
369 drives	Murray et al. (2003)	Rank-sum tests	(0.6%, 43.1%)
	Murray et al. (2005)	SVM	(0%, 50.6%)
	Zhao et al. (2010)	HMM joint model	(0%, 52%)
	Wang et al. (2014)	two-step parametric method	(0%, 68.4%)
	Queiroz et al. (2018)	FDGE	(0%, 80.24%)
23395 drives	Zhu et al. (2013)	ANN	(0.48%, 95%)
		SVM	(0.3%, 80.0%)
	Li et al. (2017)	DT	(0.01%, 93%)
		GBRT	(0%, 90%)
220,022 drives	Yang et al. (2015)	logistic regression	(0.3%, 97.82%)
25792 drives	Li et al. (2014)	classification and regression tree	(0.09%, 95.5%)
71619 drives	Xu et al. (2016)	RNN	(0.06%, 97%)
Not given	Xu et al. (2018)	Ranking model	(0.1%, 25-45%)
Not given	Ganguly et al.(2016)	DT and logistic regression	(0%, 40-50%)
Around 36531 drives	Xiao et al. (2018)	online random forests	(0.66%,98.08%)
	Shen et al. (2018)	part-voting random forest	(0.44%,94.89%)
	Jiang et al. (2019)	GAN	(0%,85%)
	Shen et al. (2021)	LSTM	(2.43%,77.33%)

In general, these methods could be divided into three categories: non-parametric statistical tests [4], [5], [9], [11], [12], [18], supervised classification [5], [8], [9], [13]–[16], [19] and time-series prediction [10], [17]. By comparison, the methods in the latter two categories perform better in larger datasets. Details of the mentioned models are summarized in Table I.

Although recent researches have reached extraordinary high failure detection rate, two serious problems are neglected. Firstly, it is difficult to make a valid comparison of the prediction performance among the proposed models. The scale and quality of the training dataset is different for most works. In addition, varying feature selection strategies and data preprocessing treatments could also make different influence on the prediction results. Therefore, it is unreasonable to make a comparison conclusion for these models solely relying on the FAR and FDR. Secondly, the time span of the collected data is much shorter than the average life span of HDDs. There is a 90% chance that a HDD will survive for three years, while the datasets only contain SMART attributes information within less than two months. Supposing the health condition gradually degrades over time, the prediction model may not be applicable to all periods since the training data may belong to only one or several periods that are not representative for the whole period.

In order to address these problems, we use the open dataset from BlackBaze Inc., which contains the SMART attributes of 36514 drives over five years, and provide a systematic study about feature selection strategies, data preprocessing treatments and classification models. Note that similar datasets have been used in several researches [20]–[23]. In this paper, 11 SMART attributes are selected whose probability distribution in the failure state is significantly different from that in the healthy state. Several data preprocessing treatments, including discretization, adding time features, data relabeling, are proposed and demonstrated to be efficient in the case study. Seven prediction models, including decision tree (DT) model, support vector machines (SVM) model, naive Bayes model, Adaboost model, Random Forests (RF) model, Multilayer Perceptron (MLP) neural network model and Long Short Term Memory (LSTM) model are trained, and their capabilities for long-period drive failure prediction can be well evaluated and compared under the same settings. The results show that Random Forest model achieves the best performance. Based on the analysis of the Gini importance of features, two SMART attributes are found closely related to the HDD failures. The analysis of the time in advance (TIA), which describes how long in advance we can detect impending failures, shows that a large part of the failures occurs suddenly, without signs until in the last three days of the life.

The rest of this paper is organized as follows. In Section II, we describe the new SMART dataset in detail and give a brief analysis of the SMART attributes. In Section III, the feature selection strategies, data preprocessing treatments and seven classification models used in this paper are introduced. The results of experiments are presented in Section IV. Section V provides a brief conclusion of this study.

II. DATASET DESCRIPTION

The dataset contains time series of SMART attributes from 36514 HDDs running in the data center of Backblaze Inc. All the HDDs are of the same model named "ST4000DM000". Denote all the attributes of the HDD i at the time t as a

sample x_{it}. The data of one HDD $x_{i1}, x_{i2}, \ldots, x_{in_i}$ includes n_i samples in a chronological order. The samples were collected once a day over 5 years from 2014 to 2018 and the running time n_i is different for different drives. If one HDD failed or was judged to be about to fail, the data collection stopped. The last collected sample was labeled as failed and the previous samples were labeled as healthy. As shown in Table II, 3335 HDDs finally failed and the remaining 33179 HDDs maintained healthy in our dataset. It should be noticed that the judgement for the failing HDDs is not entirely correct, i.e., the HDDs labeled as failed may still maintain healthy in reality. Therefore, the dataset itself determines that there exists an upper limit of prediction accuracy for all the prediction models.

TABLE II
OVERVIEW OF THE DATASET.

Dataset	Disk model	Class	No. Disks
ST	ST4000DM000	healthy	33179
		failed	3335

Each sample contains raw value and normalized value for each SMART attributes. The raw value refers to the actual value of the SMART attributes. The normalized value is obtained through compressing the raw value to a range of 0-100 or 0-120 by an unclear normalization strategy. Generally, the HDD is more likely to fail with a small normalized value.

SMART attributes can be classified into three categories, the error count attributes, whose values maintain zero at most of the time; the volatile attributes, whose values fluctuate with time; and the cumulative attributes, whose values increase over time. Apart from these attributes, each sample contains the information of the drive's serial number, model, capacity bytes and the recording date as well.

III. METHODOLOGY DEVELOPMENT

In this section, we will describe the training process of failure prediction models in detail. The training process consists of three steps. Firstly, 11 features are selected from the SMART attributes. Secondly, some effective data preprocessing treatments are conducted to establish the training dataset. Finally, seven popular classification models are trained based on the processed dataset.

A. Feature Selection

For some SMART attributes, the data are either missing or unchanging over time. These two types of attributes are eliminated and there remain 21 SMART attributes with each containing raw or normalized values. The 21 SMART attributes include 10 error count attributes, 3 volatile attributes and 8 cumulative attributes. We propose several novel feature selection strategies for these three types of attributes, aiming at making the data distribution of failure state and healthy state significantly discriminative.

1) Error count attributes: The data are simply divided into two categories, stable state data and unstable state data. A data is classified as a stable state data if it is at the stable value, and an unstable state data otherwise. Assume that the probability of a recorded value deviating from the stable value is constant for both healthy samples and failed samples of one attribute. The samples of stable state and unstable state could be considered following two binomial distributions. Hypothesis testing could be used to find out whether the data distribution of failure state and healthy state are significantly different. The confidence intervals of two binomial distributions' parameters are estimated with 95% confidence. If the two confidence intervals have no intersection, it can be regarded that the two distributions are different and the deviation of this attribute is more related to the failure of HDDs.

After hypothesis testing, as shown in Table III, we find that all the 10 SMART error count attributes show significant difference on the distribution at least in the raw data. Meanwhile, if the results of raw data and normalized data are similar, normalized data is chosen as the selected feature for a simpler value range.

TABLE III
THE CONFIDENCE INTERVALS OF BINOMIAL DISTRIBUTIONS' PARAMETERS FOR THE ERROR COUNT ATTRIBUTES.

Attribute ID	Confidence Intervals		Difference
	Failed	Healthy	
SMART 5 Raw	(0.207,0.235)	(0.002,0.006)	T
SMART 5 Normalized	(0.068,0.086)	(0.000,0.002)	T
SMART 183 Raw	(0.263,0.293)	(0.146,0.170)	T
SMART 183 Normalized	(0.263,0.293)	(0.146,0.170)	T
SMART 184 Raw	(0.020,0.031)	(0.000,0.003)	T
SMART 184 Normalized	(0.020,0.031)	(0.000,0.003)	T
SMART 187 Raw	(0.393,0.427)	(0.001,0.016)	T
SMART 187 Normalized	(0.393,0.427)	(0.001,0.016)	T
SMART 188 Raw	(0.085,0.105)	(0.051,0.067)	T
SMART 188 Normalized	(0.000,0.000)	(0.000,0.000)	F
SMART 189 Raw	(0.412,0.445)	(0.300,0.331)	T
SMART 189 Normalized	(0.412,0.445)	(0.300,0.331)	T
SMART 192 Raw	(0.468,0.502)	(0.340,0.373)	T
SMART 192 Normalized	(0.000,0.000)	(0.000,0.000)	F
SMART 197 Raw	(0.492,0.525)	(0.003,0.008)	T
SMART 197 Normalized	(0.088,0.108)	(0.000,0.000)	T
SMART 198 Raw	(0.492,0.525)	(0.003,0.008)	T
SMART 198 Normalized	(0.088,0.108)	(0.000,0.000)	T
SMART 199 Raw	(0.029,0.042)	(0.019,0.029)	T
SMART 199 Normalized	(0.000,0.000)	(0.000,0.000)	F

TABLE IV
THE MANN-WHITNEY U TEST AND THE K-S TEST RESULTS FOR THE VOLATILE ATTRIBUTES.

Attribute ID	Mann-Whitney U test		K-S test	
	statistic	p value	statistic	p value
SMART 1 Raw	8.41e8	0.426	0.0124	0.679
SMART 1 Normalized	7.56e8	3e-25	0.0825	3e-20
SMART 190 Raw	8.02e8	9e-7	0.0321	0.002
SMART 190 Normalized	8.02e8	7e-7	0.0324	0.002
SMART 194 Raw	8.02e8	9e-7	0.0321	0.002
SMART 194 Normalized	8.02e8	9e-7	0.0321	0.002

2) Volatile attributes: The distributions of healthy samples and will-fail samples are compared through two non-parametric statistical tests, the Mann-Whitney U test and the Kolmogorov–Smirnov test, which give similar results. With 95% confidence, we cannot reject that the two volatile attributes show significant difference. Thus, these two attributes are chosen as the selected features. More details are shown in Table IV.

3) Cumulative attributes: The value of the cumulative attributes grows steadily over time and the attributes are highly correlated with running time. Thus, these attributes are not directly related to the failure of HDD, and only one cumulative attribute is selected as the time feature.

For the SMART attributes with high correlations, i.e., sharing the same value most of the time, only one of the attributes will be kept as selected features. Eventually, 11 features are selected, which are shown in Table V, including 9 error count attributes, 1 volatile attribute and 1 cumulative attribute. The histograms of four representative attributes are illustrated in Figure 1. The error count attribute SMART 197 and SMART 187 show great difference between healthy samples and failed samples.

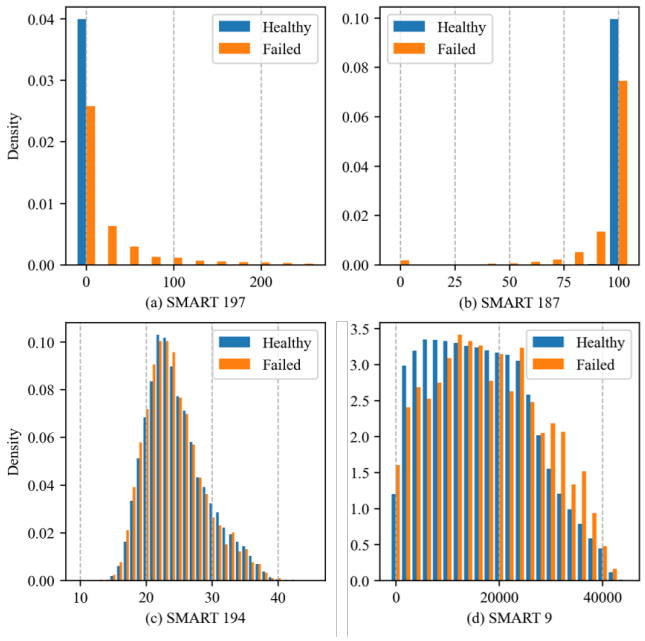

Fig. 1. Histograms of four representative attributes, (a-b) from the error count attributes, (c) from volatile attributes and (d) from cumulative attributes.

B. Data Preprocessing

1) Missing data processing: There remain a few samples with missing data for the features selected above. These samples are found only existed at the beginning of recorded data in certain hard disk, and have little effect on the subsequent prediction. Therefore, we simply delete the samples with missing data.

2) Deleting short time series: Since the HDDs will stop recording after the failure and then new HDDs are added halfway, the running time of each HDD is not the same.

TABLE V
THE ELEVEN SELECTED FEATURES.

Attribute ID	Attribute name	Date type
SMART 5	Reallocated Sectors Count	Raw
SMART 9	Power-On Hours	Raw
SMART 183	SATA Downshift Error Count	Normalized
SMART 184	End-to-End error / IOEDC	Normalized
SMART 187	Reported Uncorrectable Errors	Normalized
SMART 188	Command Timeout	Raw
SMART 189	High Fly Writes	Normalized
SMART 192	Power-off Retract Count	Raw
SMART 194	Tempreature	Normalized
SMART 197	Current Pending Sector Count	Raw
SMART 199	UltraDMA CRC Error Count	Raw

When the running time is too short, it is difficult to extract enough effective information. The prediction under such circumstance is very difficult and it is not common in practical applications. Therefore, we delete the data of HDDs whose running time is less than 90 days.

3) Discretization: Nine of the eleven selected features are of the error count attributes. Once these features deviate from the stable value, it is regarded that the hard disk has a high probability of failure. To expand the difference between deviating and stable values, we divide the value region of these features into several classes by numerical discretization and separate the stable values into one class. If the data is divided into two classes, it will be further converted into binary data. In this paper, binary discretization is taken for the nine error count attributes.

4) Adding time features: Since the recorded data is time series data, the failure of HDD may be not only related to the data information of the current sample. The previous state of a HDD also provides critical information in inferring the health condition. Therefore, the changes of data along time are added as new features. The first order difference could represent the change rate and the second order difference could represent the change trend. Specifically, three change rate features and three change trend features with intervals being 1 sample, 5 samples and 15 samples respectively are added.

5) Data relabeling: In our SMART dataset, only the last sample of a failed HDD is marked as failed. However, some HDDs show signs of imminent failure a few days before the failure date. In order to expand the data size under the failure state, the last 30 samples of the failed HDDs are labeled as failed instead.

6) Datasets partition: The preprocessed dataset is divided into training set, validation set and test set. The training set is used to train the models. The validation set is used to adjust the parameters to optimize the prediction effect. And the test set is taken to test the proposed model finally. Samples of the same HDD cannot be put into these three datasets concurrently considering that there exists a great correlation among them. Therefore, the training set, the validation set and the test set are divided according to the ID of the HDD. After preprocessing, the data set contains the data of 36,670 drives, among which 24,937 drives are classified into the

TABLE VI
THE STRUCTURE OF THE MLP MODEL AND THE LSTM MODEL.

	Layer	Nodes	Number of parameters
MLP model	Dense 1	128	9984
	Dense 2	32	4128
	Dense 3	1	33
LSTM model	LSTM 1	32	15232
	Dense 1	128	4224
	Dense 2	32	4128
	Dense 3	1	33

training set, 6233 drives are classified into the validation set and 5500 drives are classified into the test set (500 failed drives and 5000 healthy drives).

7) Data sampling: We do not take all healthy samples into our training set, as the training would be too complex and time-consuming. Instead, we randomly select three samples from the healthy samples for each HDD, which also balances the amount of samples belonging to the two states.

C. Classification methods

In this section, we briefly introduce some of the classification models used in our work. Some basic settings of these models are also provided.

1) AdaBoost based model and Random Forest based model: AdaBoost model is a representative of ensemble learning models, whose main steps are as follows. Firstly, a base classification algorithm is used to train the first classifier. Then, based on the predicted results of the previous classifier on the training set, the weight of each sample is adjusted, i.e., assigning a greater weight to the samples that are not accurately classified. This process continues iteratively until the number of classifiers reaches the limitation or the performance of the new classifier is too bad. Different with AdaBoost model, Random Forest model train multiple classifiers simultaneously. In order to widen the distinction between classifiers, bootstrap sampling is used to create different training sets for those classifiers. For both the ensemble learning models, Decision Tree (DT) models are used as base estimators in our paper.

2) Multilayer Perceptron based model and Long Short Term Memory based model: These two models are both originated from Artificial Neural Network (ANN), whose main idea is to construct a network structure composed of numerous artificial neurons. Multilayer Perceptron (MLP) model contains several neural layers and each layer consists of many neurons. Layers are linked by the dense connections between neurons. Long short-term memory (LSTM) model is a variant of artificial recurrent neural network architecture, which further takes the time correlation of data into consideration. In the following experiments, three dense layers are used in the MLP model. One LSTM layer and two dense layers are used in the LSTM model. The structure of the MLP model and the LSTM model is shown in Table VI.

IV. PERFORMANCE EVALUATION AND COMPARISON

A. Performance metrics

The prediction effect of the model can be measured by two types of performance metrics. The first type is measured in terms of single sample. Specifically, each sample has two kinds of real states and two kinds of predicted states classified by the model. Then, all samples could be divided into four categories named true positive (tp), true negative (tn), false positive (fp) and false negative (fn), where true and false represent the real state, positive and negative represent the predicted state. We aim to get both a higher True Positive Rate (TPR) and a lower False Positive Rate (FPR).

$$TPR = tp/(tp+fn), FPR = fp/(fp+tn).$$

The second type of performance metrics is measured in terms of single HDD. Since each HDD will eventually fail, we aim to correctly predict the drive failure in advance, e.g., less than certain days before the failure, throughout its life. Denote γ as the predefined upper bound for the time in advance (TIA), which is set 30 in this paper. For those finally failed drives, the failure is considered correctly predicted if it is classified as failure at any time with remaining useful life $RUL \leq \gamma$, and vice versa. For those drives maintaining healthy, it means that all samples are classified as healthy. Similar to the first type of performance metrics, all the detection could be divided into four categories named TP, TN, FP and FN. We aim to get both a higher Failure Detection Rate (FDR) and a lower False Alarm Rate (FAR).

$$FDR = TP/(TP+FN), FAR = FP/(TN+FP).$$

In the dataset, the label of the samples are not completely accurate. Only the last sample of the failed HDDs are marked as failed. It is difficult for us to know exactly when the HDD fails. As a result, the first type of performance metrics are greatly affected. In addition, the second type of performance metrics are more important and more meaningful in real applications. Therefore, we choose FDR and FAR as the final performance metrics in the case study.

B. Experiment results and comparison

We assume that all samples after preprocessing are independent during the training. For the LSTM model, our input is a 30-day time series and the output is the predicted state of the last sample. For the other six models, the input is a single sample and the output is its predicted state. Seven models are compared based on the second type of performance metrics at last. The FDR-FAR curves of the seven models are illustrated in Figure 2. Details of the performance are summarized in Table VII. The bold values indicate the best performance among these models with respect to the corresponding metric.

The results demonstrate that Random Forest model possesses the best prediction performance with $FAR = 6.0\%$ and $FDR = 53.95\%$. The performance of SVM model, Adaboost model and MLP model are slightly worse. Decision Tree

TABLE VII

THE FAILURE PREDICTION PERFORMANCE OF SEVEN MODEL.

Model	FDR(%) FAR=4%	FDR(%) FAR=6%	Area Under FDR-FAR Curve	Running time(s)
DT	2.36	3.54	0.5456	2308.9
SVM	36.79	53.09	0.5553	12700.8
Bayes	33.74	37.76	0.5211	2256.1
Adaboost	39.35	46.13	0.5735	2470.7
RF	**40.76**	**53.95**	0.5579	**1839.4**
MLP	38.33	49.20	**0.6074**	4312.9
LSTM	16.89	23.20	0.4498	27063.3

model, naive Bayesian model and LSTM based model do not perform well in this prediction problem.

In terms of the training time, SVM model and LSTM model require much longer training time but do not achieve better performance. Therefore, these two models are considered not suitable for HDD failure prediction problem.

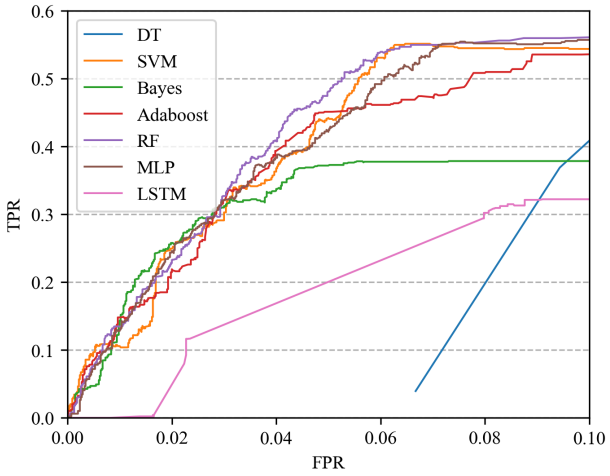

Fig. 2. Comparisons of FDR-FAR curves for seven models.

C. Feature importance analysis

The Gini importance of each feature, which is defined as the total decrease in node impurity brought by that feature, could be calculated in the Random Forest based model. The most important 10 features are listed in Figure 3.

Two error count attributes, named SMART 197 Current Pending Sector Count and SMART 187 Reported Uncorrectable Errors, are most related to HDD failures. Among the ten most important features, six of them (including four time features) are connected to the two attributes. This is consistent with our results in feature selection process, where the deviation ratio from stability of these two attributes, 51.4% and 41.0%, are the largest for failed samples. It suggests that when these two attributes deviate from the stable value, there is a great chance that the HDD will fail in 3 days.

D. Performance improvement by data preprocessing steps

In this section, we re-trained the Random Forest model by applying the aforementioned data preprocessing techniques

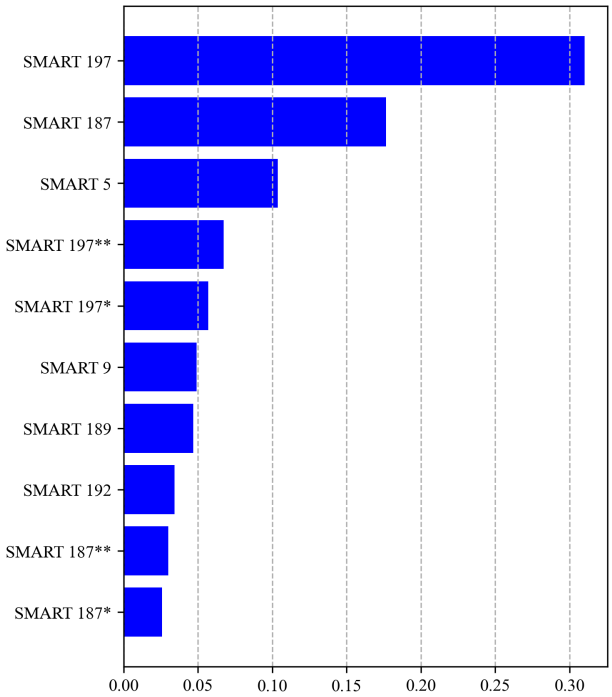

Fig. 3. The most important 10 features in the Random Forest based model. The notation * and ** refer to the change trend features with intervals being 5 samples and 15 samples respectively.

(adding time features, data relabeling) progressively, and illustrate the effectiveness of our method in Figure 4. After adding time features, the number of features increases from 11 to 77, and the Failure Detection Rate increases from 38.64% to 40.00%. Further relabeling the last 30 samples of the failed HDDs, the Failure Detection Rate reaches at 40.76%. The experimental results show the significance of our data preprocessing process.

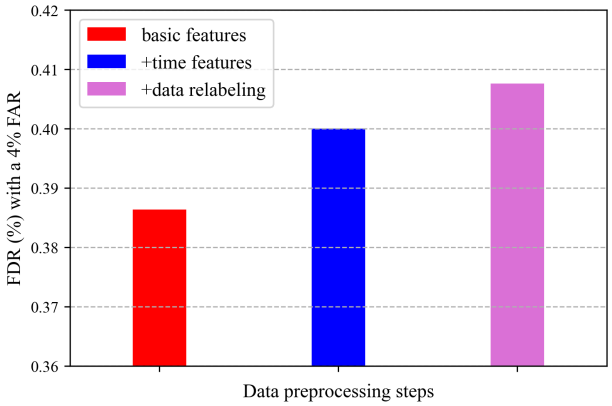

Fig. 4. Improvements in FDR (%) by adding different data preprocessing techniques. We have fixed FAR to 4%.

E. Time in advance analysis

Besides the prediction performance, it is also critical that how much time in advance the failure is predicted before the underlying true failure. Time In Advance (TIA) of the

Random Forest model with a 6% FAR and 53.95% FDR is shown in Figure 5.

In the 191 HDDs whose failure was predicted correctly in advance, 51 failures (26.70%) are predicted on the last sample and 80 failures (41.88%) are predicted on the last three samples, indicating that a large part of the failure occurs suddenly.

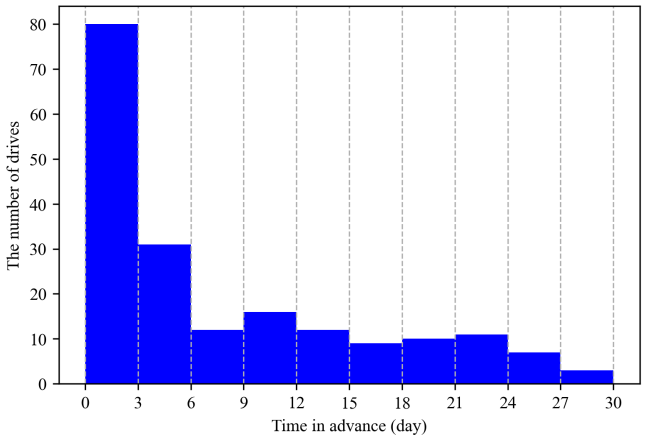

Fig. 5. The histogram of Time In Advance of the Random Forest model with a 6% FAR and 53.95% FDR.

V. CONCLUSIONS

In our study, we build up a long-period data set for HDD failure prediction. The data of 36514 drives are collected daily over 5 years. Seven prediction models are trained and compared with the same feature selection strategies and data preprocessing treatments. The Random Forest model achieves the best performance ($FAR = 6.0\%$ and $FDR = 53.95\%$). Two error count attributes are found closely related to the HDD failures.

The prediction results of most models are very similar. It infers that it is the quality of the data set rather than the model that mainly limit the prediction accuracy. As mentioned in Section II, the marking of a HDD failure is subjective to some degree and not entirely correct. Some healthy drives may be wrongly labeled in the existing data set. In addition, the recording interval is one day in our dataset, which is too long especially considering a large part of the failures occur suddenly in three days. Datasets with longer time spans, shorter record intervals and more accurate labels are demanded.

In addition, we simply predict whether the HDD will fail within 30 days, which may not be flexible in real applications. Condition monitoring and RUL prediction could be a more flexible alternative. Moreover, all the HDDs in our dataset are of the same model named “ST4000DM000”. How to apply our prediction methods to other HDD models needs our further research, where transfer learning could be discussed in depth.

REFERENCES

[1] B. Schroeder and G. A. Gibson, ”Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you?,” ACM Transactions on Storage (TOS), 2007, 3(3): 8-es.

[2] E. G. Grochowski and R. F. Hoyt, ”Future trends in hard disk drives,” Asia Pacific Magnetic Recording Conference, vol. 32, no. 3, pp. 1850-1854, 1996.

[3] K. V. Vishwanath and N. Nagappan, ”Characterizing cloud computing hardware reliability,” in symposium on cloud computing, 2010, pp. 193-204.

[4] G. F. Hughes, J. F. Murray, K. Kreutzdelgado, and C. Elkan, ”Improved disk-drive failure warnings,” IEEE Transactions on Reliability, vol. 51, no. 3, pp. 350-357, 2002.

[5] J. F. Murray, G. F. Hughes, and K. Kreutzdelgado, ”Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application,” Journal of Machine Learning Research, vol. 6, pp. 783-816, 2005.

[6] Y. Wen, J. Wu and Y. Yuan, ”Multiple-Phase Modeling of Degradation Signal for Condition Monitoring and Remaining Useful Life Prediction,” in IEEE Transactions on Reliability, vol. 66, no. 3, pp. 924-938, Sept. 2017.

[7] Z. Li, J. Wu and X. Yue, ”A Shape-Constrained Neural Data Fusion Network for Health Index Construction and Residual Life Prediction,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 5022-5033, Nov. 2021.

[8] G. Hamerly and C. Elkan, ”Bayesian approaches to failure prediction for disk drives,” in International Conference on Machine Learning, 2001, pp. 202-209.

[9] G. F. H. J. F. Murray, K. Kreutz-Delgado, ”Hard drive failure prediction using non-parametric statistical methods,” in Proceedings of the International Conference on Artificial Neural Networks, June 2003.

[10] Y. Zhao, X. Liu, S. Gan, and W. Zheng, ”Predicting disk failures with HMM- and HSMM-based approaches,” in International Conference on Data Mining, 2010, pp. 390-404.

[11] Y. Wang, E. W. M. Ma, T. W. S. Chow, and K. 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.

[12] L. P. Queiroz et al., ”Fault Detection in Hard Disk Drives Based on a Semi Parametric Model and Statistical Estimators,” New Generation Computing, vol. 36, no. 1, pp. 5-19, 2018.

[13] B. Zhu, G. Wang, X. Liu, D. Hu, S. Lin, and J. Ma, ”Proactive drive failure prediction for large scale storage systems,” in IEEE Conference on Mass Storage Systems and Technologies, 2013, pp. 1-5.

[14] J. Li, R. J. Stones, G. Wang, X. Liu, Z. Li, and M. Xu, ”Hard drive failure prediction using Decision Trees,” Reliability Engineering System Safety, vol. 164, pp. 55-65, 2017.

[15] W. Yang, D. Hu, Y. Liu, S. Wang, and T. Jiang, ”Hard Drive Failure Prediction Using Big Data,” in 2015 IEEE 34th Symposium on Reliable Distributed Systems Workshop (SRDSW), 2015.

[16] J. Li et al., ”Hard Drive Failure Prediction Using Classification and Regression Trees,” in Dependable Systems and Networks, 2014, pp. 383-394.

[17] C. Xu, G. Wang, X. Liu, D. Guo, and T. Liu, ”Health Status Assessment and Failure Prediction for Hard Drives with Recurrent Neural Networks,” IEEE Transactions on Computers, vol. 65, no. 11, pp. 3502-3508, 2016.

[18] Y. Xu et al., ”Improving Service Availability of Cloud Systems by Predicting Disk Error,” in USENIX Annual Technical Conference, 2018, pp. 481-494.

[19] S. Ganguly, A. Consul, A. Khan, B. Bussone, J. Richards, and A. Miguel, ”A Practical Approach to Hard Disk Failure Prediction in Cloud Platforms: Big Data Model for Failure Management in Datacenters,” in IEEE Second International Conference on Big Data Computing Service & Applications, 2016.

[20] J. Xiao, Z. Xiong, S. Wu, Y. Yi, H. Jin, and K. Hu, ”Disk failure prediction in data centers via online learning,” in Proceedings of the 47th International Conference on Parallel Processing, 2018, pp. 1-10.

[21] J. Shen, J. Wan, S.-J. Lim, and L. Yu, ”Random-forest-based failure prediction for hard disk drives,” International Journal of Distributed Sensor Networks, vol. 14, no. 11, p. 1550147718806480, 2018.

[22] J. Shen, Y. Ren, J. Wan, and Y. Lan, ”Hard disk drive failure prediction for mobile edge computing based on an LSTM recurrent neural network,” Mobile Information Systems, vol. 2021, 2021.

[23] T. Jiang, J. Zeng, K. Zhou, P. Huang, and T. Yang, ”Lifelong disk failure prediction via GAN-based anomaly detection,” in 2019 IEEE 37th International Conference on Computer Design (ICCD), 2019, pp. 199-207: IEEE.