Lifelong Disk Failure Prediction via GAN-based Anomaly Detection

Tianming Jiang[†], Jiangfeng Zeng[‡]*, Ke Zhou[†]*, Ping Huang[§], Tianming Yang[††],
[†] Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System,
Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China,
School of Computer Science & Technology, Huazhong University of Science & Technology, China
[‡] School of Information Management, Central China Normal University, China
[§]Temple University, USA, [§]Huanghuai University, China
Email: jiangtianming@hust.edu.cn, jfzeng@mail.ccnu.edu.cn,
k.zhou@hust.edu.cn, templestorager@temple.edu, ytm@huanghuai.edu.cn

***Abstract*—As a classical technique in storage systems, disk failure prediction aims at predicting impending disk failures in advance for high data reliability. Over the past decades, taking as input the SMART (Self-Monitoring, Analysis and Reporting Technology) attributes, many supervised machine learning algorithms have been proven to be effective for disk failure prediction. However, these approaches heavily rely on the availability of substantial annotated failed disk data which unfortunately exhibits an extreme data imbalance, i.e., the number of failed disks is much smaller than that of healthy ones, resulting in suboptimal performance and even inability at the beginning of their deployment, i.e., cold starting problem.**

Inspired by the significant success achieved in GAN (Generative Adversarial Network) based anomaly detection, in this paper, we translate disk failure prediction into an anomaly detection problem. Specifically, we develop a novel Semi-supervised method for lifelong disk failure Prediction via Adversarial training, called SPA. The distinguishing feature of SPA from existing supervised approaches is that SPA is only trained on healthy disks, which avoids the traditional limitations of imbalance in datasets and eliminates the cold starting problem. Furthermore, a novel 2D image-like representation technique is proposed to enable the deployment of deep learning techniques and the automatic feature extraction. Experimental results on real-world SMART datasets demonstrate that, compared with the state-of-the-art supervised machine learning based methods, our approach predicts disk failures at a higher accuracy for the entire lifetime of models, i.e., both the initial period and the long-term usage.

***Index Terms*—disk failure; data reliability; SMART; adversarial training; anomaly detection**

I. Introduction

Since the advent of computers and the Internet, there has been an explosive growth of the volume of data, and over 90% of data is stored in hard disk drives [1]. It has been estimated that hard disk drive failures account for 78% of the hardware replacements in data centers [2]. Disk failures have become the normality rather than the exception as the size of storage continually grows [3], [4]. Moreover, disk failures may bring about catastrophic consequences if not properly handled, for they not only cause service downtime but also lead to data loss if no data redundancy schemes are deployed [5], [6].

To ensure high data reliability and availability of storage systems, some data redundancy schemes, e.g., replication [3] and erasure code [7], [8], have been proposed and deployed in storage systems as remedy methods in response to disk failure occurrences. However, these approaches are reactive fault-tolerant techniques used to reconstruct data when disk failures occur, and thus they are storage space inefficient and bandwidth demanding [9]. For this reason, disk failure prediction is proposed to predict disk failures before they actually happen. The key idea behind disk failure prediction is that if disk failures have been predicted, users could be informed to take precautions, e.g., data migration, which can significantly reduce the maintenance costs.

Nowadays, the vast majority of modern hard disk drives have been equipped with Self-Monitoring, Analysis and Reporting Technology (SMART), which monitors individual disk and outputs attributes that contain the information regarding the evolution of disk states. Therefore, SMART itself provides the ability to predict impending disk failures. Specifically, before a disk fails, a binary warning (will fail) will be issued if any attribute exceeds its threshold defined by the manufacturers [10]. However, this threshold-based method can only reach a failure detection rate (FDR) of 3-10% with 0.1% false alarm rate (FAR) [11] due to its simplicity and the conservative settings of thresholds [12].

Targeting at improving the FDR, many supervised machine learning methods have been introduced [11]–[20]. These approaches take the SMART attributes as input, followed by classifiers implemented using supervised machine learning algorithms. Although some approaches [12], [19], [20] have achieved satisfactorily high FDRs and low FARs, they suffer heavily from the data imbalance issue, i.e., the number of failed disks is much smaller than that of healthy ones. Unfortunately, balanced labeled data, where the number of samples in each class is roughly equal, is essential for classifiers [21]. In other words, the data imbalance issue undermines the accuracy of disk failure prediction. What is worse, the training data is gradually gathered instead of being given in advance [12]. As a result, the training data collected within the initial period may be insufficient and could result in an inability of the predictor

*Corresponding authors

978-1-5386-6648-7/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCD46524.2019.00033

at the beginning of its deployment, i.e., cold starting problem.

In the field of computer vision, anomaly detection is confronted with a similar problem as there is a wide gap in the ratio between normal samples and abnormal samples. Recent research results have demonstrated that Generative Adversarial Networks (GANs) are able to capture the data distribution and thus have been investigated for anomaly detection. Considering the superiority of anomaly detection in dealing with the heavily unbalanced dataset and inspired by the significant success achieved by GAN-based approaches in anomaly detection, in this paper, we propose a novel end-to-end semi-supervised deep learning method for detecting impending disk failures. Specifically, a **S**emi-supervised method for lifelong disk failure **P**rediction via **A**dversarial training, SPA, is applied to learn a manifold of healthy disk SMART attributes.

However, it is not straightforward but very challenging to apply GAN-based anomaly detection for disk failure prediction. The first challenge is how to transfer non-image SMART data into 2D image-like representation data for the deployment of deep learning techniques. We tackle this challenge with a novel data construction method which segments time series SMART data via a sliding window. The merit of the image-like representation data lies in that it enables the deployment of deep learning techniques and the automatic feature extraction, killing two birds with one stone. Another critical challenge is how to deal with the model aging problem that the prior trained model will lose validity on the new coming SMART data [12]. The model aging problem is due to the dynamic characteristic of SMART, i.e., the underlying distribution of SMART will gradually change with the sequentially collected data. To solve the problem, we exploit the fine-tuning technique within deep learning for model updating. Specifically, we fine-tune the old model on new coming data periodically.

To sum up, we make the following contributions in this paper:

- To the best of our knowledge, we pioneer the use of GAN-based anomaly detection for disk failure prediction to deal with the data imbalance issue and cold starting problem.
- We propose a novel data construction method, which transforms non-image SMART data into 2D image-like representations and enables the deployment of deep learning techniques and the automatic feature extraction.
- To deal with the model aging problem, we propose using the fine-tuning technique within deep learning to do efficient model updating.
- Conducting experiments on real-world datasets, we simulate the use of our model in both the initial period and the long-term use and validate its applicability for the small- and large-size datasets. The experimental results demonstrate the effectiveness of our model.

The remainder of this paper is organized as follows: In Section II, we introduce the motivations of our work. Section III describes the proposed approach. Section IV discusses our experimental results. The related work is presented in Section V, followed by conclusion in Section VI.

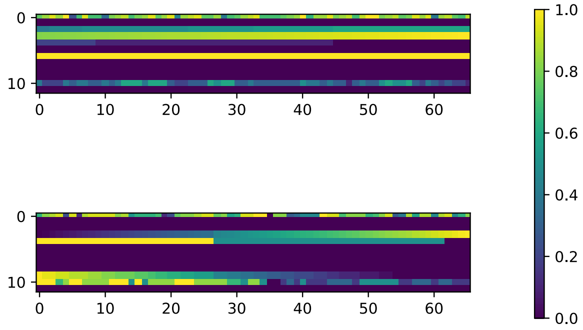

Fig. 1: SMART attribute values along with the time of exemplary disks, increasing a healthy disk (top) and a failed disk (bottom). The x-axis is the time of days and the y-axis is SMART attribute values which are normalized to the global mean and standard deviation for the convenience of display. Obviously, the SMART attribute values of the failed disk have more transitions.

II. Motivations

In this section, we introduce the challenges of disk failure prediction which motivate our work. There are two main challenges needed to be adequately addressed for high prediction performance, i.e., the data imbalance issue and the time series feature extraction from SMART data.

A. Data Imbalance Issue

Since disk failure is relatively a rare event and failed disks only account for a tiny part of all disks [12], there exists a data imbalance phenomenon, i.e., the number of healthy disks is much larger than that of failed ones. However, traditional supervised machine learning methods work well only for balanced datasets [21]. When working on an imbalanced dataset, these methods biasedly predict all disks to be healthy for overall high accuracy [21].

To this end, some re-balancing techniques [12], [18], i.e., oversampling and downsampling, are proposed to obtain a balanced training set for these supervised methods. Botezatu et al. [18] downsample the healthy samples of the entire training set to an amount that is close to the size of the failed samples. In [12], online bagging technique is used to sample the sequentially arrived data for online learning. Besides the downsampling scheme, there are also works that address the data imbalance issue from the cost-sensitive learning perspective [20], assigning different costs to the false positive (FP) and false negative (FN).

B. Time Series Features

1D-SMART attributes (the SMART attributes of one disk at a specific timepoint) are fed into classifiers implemented by supervised machine learning algorithms while the changes of SMART attributes over time are ignored. Previous research has shown that the changes in SMART attributes over time, i.e.,

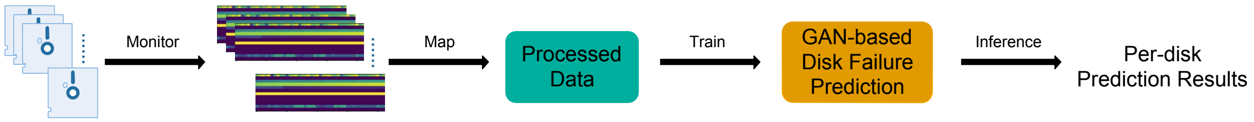
Monitor
Map
Processed Data
Train
GAN-based Disk Failure Prediction
Inference
Per-disk Prediction Results

Fig. 2: Overview of the SPA and its two main components, i.e., the data processing and the training of GAN-based method for disk failure prediction. After the SMART attributes are monitored, they are aggregated and mapped to processed data which are image-like representations. Then in the training phase, the healthy image-like representations are used to train a GAN-based method for disk failure prediction via adversarial training. In the inference phase, the GAN-based disk failure prediction takes image-like representations as input and outputs per-disk prediction results.

temporal locality, are beneficial to distinguishing disk failures. In [19], Zhu et al. calculate the absolute differences between the current SMART attributes values and their corresponding values six hours ago as time series features. Formally, given a time window size of w, the absolute difference at time stamp t is noted as $Diff$ and calculated as $Diff(x, t, w) = x(t) - x(t - w)$ [19], [20]. Besides $Diff$, $Sigma$ means the variance of attribute values within a period and is calculated as $Sigma(x, t, w) = E[(X - \mu)^2]$, and Bin means the sum of attribute values within a period and is calculated as $Bin(x, t, w) = \sum_{j=t-w+1}^{x(j)}$ [20].

To validate the importance of time series features in SMART data, we conduct two preliminary examinations. First, we calculate the variance of each SMART attribute within a period and sum these variances up. The results reveal that failed disks score bigger than healthy disks in terms of the sum of variances. Second, to have an intuitive understanding, we visualize the changes of SMART attributes over time for both one healthy disk and one faulty disk as an example, as shown in Fig. 1. It can be observed that for faulty disks, some (maybe not all) SMART attributes fluctuate drastically over time while for healthy disks, the SMART attributes remain stable or change little over time. In summary, time series features are indicative of disk failure and could play an essential role in predicting disk failures.

For SPA, to enable the deployment of deep learning techniques, 2D-SMART attributes are taken as input. 2D-SMART attributes refer to SMART attributes of one disk within a period and are obtained by stacking several 1D-SMART attributes within a specified period. This construction renders image-like representations which are suitable for 2D convolution operations of deep learning. Besides enabling deployment of deep learning, therefore, 2D-SMART attributes also bring the benefit of automatic feature extraction. We illustrate the construction of 2D-SMART attributes in detail in Section III-A.

III. The Proposed Method

Our goal is to predict whether a disk will fail within a given time interval, using the SMART data that the disk reported. In the following discussion, we constrain the period to seven days before a faulty event for the sake of simplicity. We formulate the prediction problem as an anomaly detection problem instead of a classical binary classification problem.

TABLE I: The 12 selected SMART attributes

Attribute ID	Attribute Name	Attribute Type
1	Real_Read_Error_Rate	Normalized
4	Start_Stop_Count	Raw
5	Reallocated_Sector_Count	Raw
7	Seek_Error_Rate	Normalized
9	Power_On_Hours	Normalized
10	Spin_Retry_Count	Normalized
12	Power_cycle_Count	Raw
187	Reported_Uncorrect	Normalized
194	Temperature_Celsius	Normalized
197	Current_Pending_Sector	Raw
198	Offline_Uncorrectable	Raw
199	UltraDMA_CRC_Error_Count	Raw

In this section, we will describe the proposed method SPA and techniques adopted for training. As shown in Figure 2, SPA comprises two main components: (1) the data processing, (2) the training of GAN-based disk failure prediction.

A. Data Processing

1) Feature Selection: Feature selection aims to remove redundant and irrelevant features and select relevant features. This preprocessing not only reduces the time of model training and predicting, but also enhances the prediction performance [11].

For each disk in our datasets, it reports 24 SMART attributes. For each attribute, it contains two values of interest, including a raw value and a normalized value. Treating each SMART attribute value as a feature, therefore, we have 48 features to choose from. We first use Pearson's correlation coefficients [22] to measure the feature-to-label information and obtain the feature ranking. Then we build random forests (RF) models on different numbers of top k features and determinate the value of k by comparing the FDRs of RF models. The selected 12 most correlated features are shown in Table I.

2) Construction of 2D-SMART attributes: To support the use of GAN-based model for disk failure prediction, we first reformat the input format of SMART attributes. Inspired by [23], [24] where a 2D chunk, i.e., image-like representation, is adopted as the input of Convolutional Neural Network (CNN), we reformat the 1D-SMART attributes into 2D input chunks which maintain the temporal locality of the time series

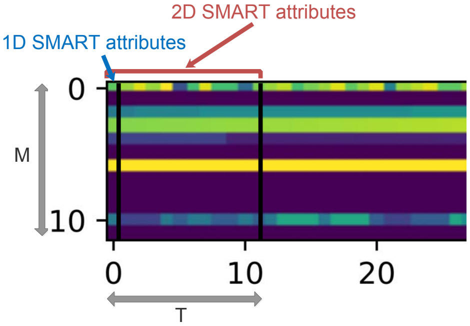

Fig. 3: Construction of 2D-SMART attributes. For each individual disk, we first stack its continual 1D-SMART attributes (the SMART attributes of one disk at a specific timepoint) and then segment data using a sliding window with the same size of selected features, resulting in a 2D-SMART attributes (M selected features within time range T, i.e., the size is $M * T$).

SMART data. As shown in Figure 3, 1D-SMART attributes refer to the M selected SMART features of one disk at a specific timepoint, and 2D-SMART attributes indicate a group of 1D-SMART attributes within T time range. The construction, also denoted as 1Dto2D in the following discussion, of 2D-SMART attributes exploits CNN's advantage of feature extraction which is automatically done and adapted by the deep learning model itself. Note that, since the construction only involves stacking 1D-SMART attributes along with time, the construction is light-weight and straightforward. After constructing the 2D-SMART attributes input, we use only the samples from healthy disks to train the GAN-based anomaly detection model.

3) Normalization: Since different SMART attributes have diverse value intervals, to ensure a fair comparison among them, we apply data normalization. The normalization used in our approach is calculated as follows [12], [20]:

$$x^{'} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is the original value of a feature, and max and min are the maximum value and the minimum value of the feature in our dataset, respectively.

B. GAN-based method for Disk Failure Prediction

1) GAN for Anomaly Detection: Known as one of the most outstanding deep generative models, Generative Adversarial Networks (GANs) were first put forward by Goodfellow et al. [25]. GANs consist of two components, i.e., the generator G that aims to generate synthetics looking like true training data and the discriminator D that aims to distinguish training samples from synthetics produced by G. The key idea behind GAN is that the generator G and the discriminator D compete with each other like two players of one game and are optimized alternatively using stochastic gradient descent (SGD).

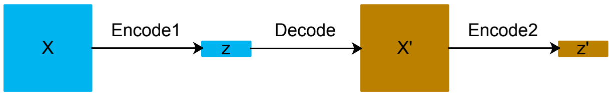

Fig. 4: $Encode1$ learns the input data representation z and $Decode$ reconstructs the input. $Encode2$ learns the representation $z^{'}$ of reconstructed data. In the training phase, the model learns both the normal data distribution and minimizes the output anomaly score $A(X)$, calculated as $||z - z^{'}||_1$. In the test phase, the anomaly score $A(X)$ is compared with a threshold ϕ, and an anomaly will be alarmed if $A(X) > \phi$.

Formally, the competition between the generator G and the discriminator D can be computed as follows:

$$\begin{aligned} \min_G \max_D \mathbf{E}_{x \sim p_d(x)} [\log D(x)] + \\ \mathbf{E}_{z \sim p(z)} [\log(1 - D(G(z)))] \end{aligned} \tag{2}$$

where x is a training sample following the true data distribution $p_d(x)$, and z is a latent vector sampled from a prior distribution $p(z)$ (e.g., standard normal distribution).

Recently, Akcay et al. [26] establish a generic GAN-based anomaly detection framework which is made up of encoder-decoder-encoder sub-networks and achieves satisfactory performance for anomaly detection problem. In their work, they take normal samples as input and use autoencoder in standard GAN to generate samples that are as close as possible to normal samples. The details of their GAN-based anomaly detection is shown in Figure 4. The autoencoder network for image generation can learn the feature representation z of the input sample X. To detect anomalies, they add an encoder to learn the representation $z^{'}$ of the reconstructed sample $X^{'}$. The difference between z and $z^{'}$ is used to measure the effectiveness of sample generation. The smaller the difference is, the better the sample generation is. Because only the normal samples are used for training, the model learns the distribution of normal samples, making the difference for normal samples smaller. The abnormal sample deviates from the normal sample distribution, so the difference is more significant. Therefore, the difference $A(X)$, calculated as $||z - z^{'}||_1$, is used to signify anomaly detection, i.e., when its value is larger than a certain threshold, it means that the sample is abnormal.

2) GAN for Disk Failure Prediction: In our disk failure prediction scenario, the underlying distribution of SMART attributes gradually change with time [12], [16]. As a result, we are facing the model aging problem, i.e., the prior trained model will lose validity on the new coming SMART data.

To deal with the model aging issue, we use the fine-tuning feature of CNN to do model updating. Fine-tuning is a common technique to transfer information from one dataset to another one. Unlike the 1-month replacing strategy [16], which discards the old model and trains a brand-new model using new coming data, fine-tuning belongs to the accumulation

Algorithm 1 Model-updating-enabled GAN-based Algorithm

```
Input: Disk identifier: $i$; Current 1D-SMART attributes: $\vec{x}$; Current disk status: $y$
Input: 1D-SMART attributes Dataset: $S$; Constructed 2D-SMART attributes dataset: $S^{'}$
Output: Prediction result: $y^{'}$
//Model update phase
if $y == 1$ then                                  ▷ Disk $D_i$ is failed
    deleteDisk($D_i$)
else                                              ▷ Disk $D_i$ is operating
    if ifFull($Q_i$) then
        $\vec{x}^{'} \leftarrow dequeue(Q_i)$
        enset($S, \vec{x}^{'}$)
    end if
    enqueue($Q_i, \vec{x}$)
    if ifFull($S$) then
        $S^{'} \leftarrow 1Dto2D(S)$              ▷ construct 2D-SMART attributes samples using dataset $S$
        fine-tune(oldGAN, $S^{'}$)                ▷ fine-tune old GAN using dataset $S^{'}$
        emptyset($S$)                             ▷ empty dataset $S$
    end if
    //Prediction phase
    $X \leftarrow 1Dto2D(Q_i)$                    ▷ construct 2D-SMART attributes samples using $Q_i$
    $y^{'} \leftarrow predictGAN(X)$
    if $y^{'} == 1$ then                          ▷ Disk $D_i$ is soon-to-fail
        Trigger an alarm
    end if
end if
```

updating strategy [16] which reuses the old model and retrains it on new coming data.

However, sample labeling is very challenging due to the training samples arriving continuously and the statuses of disks being uncertain [12]. To address this issue, we adopt the automatic online labeling method proposed by Xiao et al. [12]. In detail, a fixed length first-in-first-out queue Q_i is used to store the samples for disk D_i and keep the samples unlabeled. After D_i has failed, all the samples in the queue Q_i will be labeled as positive. If D_i is still in operation, Q_i outputs the oldest samples which are then labeled as negative, and replaces them with new ones.

Unlike the supervised method used in [12] where both healthy and failed samples are used to train the models, our semi-supervised method only uses healthy samples. Another difference is that fine-tuning relaxes the updating frequency. We update our model using batches of samples instead of every single new sample used in online learning [12]. Specifically, we use a dataset S to maintain annotated data within a constant time interval and update models using dataset S. When dataset S is full, we construct them into 2D-SMART attributes chunks, i.e., image-like representations, as shown in Figure 3. Note that in our implementation, the model updating interval is not equal to the prediction time interval, and we predict for each sample currently collected. The proposed model-updating-enabled GAN-based algorithm for disk failure prediction is illustrated in Algorithm 1.

TABLE II: Overview of dataset

Dataset	Disk Model	Class	No. Disks
STA (large-size)	ST4000DM000	good	33,701
		failed	1,508
STB (small-size)	ST8000DM002	good	9,887
		failed	92

IV. Experimental Results

A. Datasets

To evaluate the proposed method, we use publicly available real-world datasets from Backblaze [27], which span a period of 12 months ranging from January 2017 to December 2017. The datasets contain daily snapshots of all SMART attributes for each operational disk. From the datasets, we select two disk models, Seagate's ST4000DM000 and ST8000DM002, which are a large-size dataset and a small-size dataset respectively. A failure event in our work is defined as follows: a disk is considered to have failed if it was replaced as part of a repair procedure, which is inherited from Pinheiro et al. [1]. As shown in Table II, failed disks indicate disks that are replaced in the year of 2017.

B. Metrics

We use the *failure detection rate* (FDR) and the *false alarm rate* (FAR) metrics, which have been widely used to evaluate the effectiveness of disk failure prediction [12], [19]. FDR is defined as the ratio of correctly predicted failed disks to the total failed disks:

$$FDR = \frac{\text{\#true positives}}{\text{\#true positives + \#false negatives}} \tag{3}$$

FAR is defined as the ratio of mis-predicted good disks to the total good disks:

$$FAR = \frac{\text{\#false positives}}{\text{\#false positives + \#true negatives}} \tag{4}$$

Note that a disk is predicted as failed if any of the samples from it is predicted as failed. Moreover, a failed disk is correctly detected only when any of the samples collected within the last week before failure is predicted positive, and a healthy disk is mis-predicted if any of the samples collected outside the latest week is predicted as positive [12]. We calculated FDRs and FARs in each model updating interval.

C. Comparisons with Existing Methods

To evaluate our SPA method, we divide disks in each dataset randomly into training set and test set in the proportion of 7:3. To demonstrate the effectiveness of our model, we compare it with three commonly used classification algorithms: RF (random forests), SVM (support vector machines), and BP (backward propagation neural networks). Note that RF is

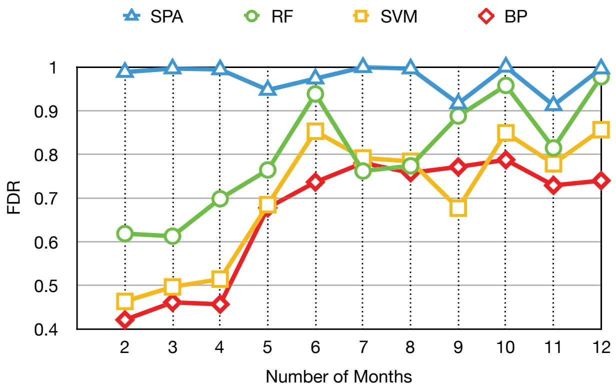

Fig. 5: Comparisons of FDRs with existing methods on dataset STA (large-size dataset).

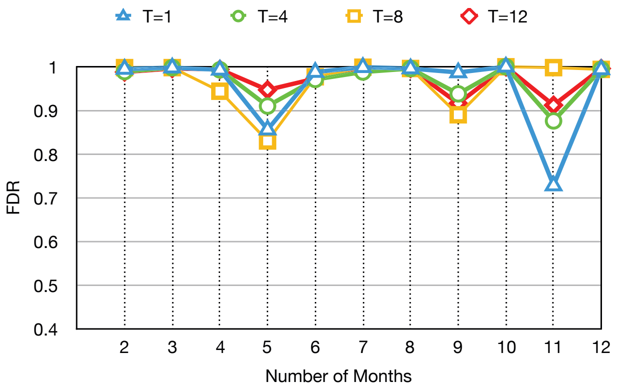

Fig. 7: FDRs under different time range T on dataset STA (large-size dataset).

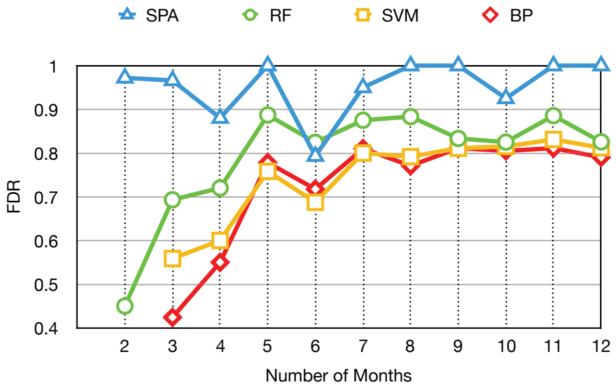

Fig. 6: Comparison of FDRs with existing methods on dataset STB (small-size dataset).

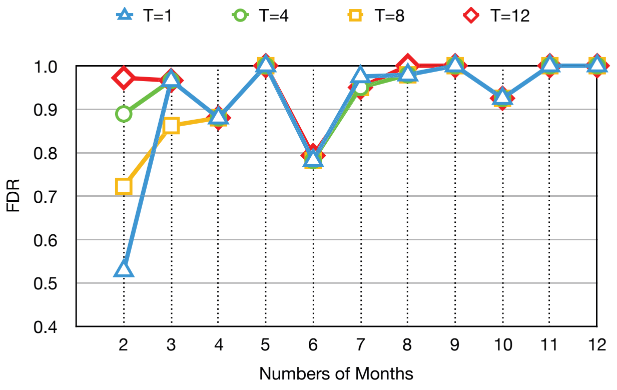

Fig. 8: FDRs under different time range T on dataset STB (small-size dataset).

demonstrated to be able to deliver the state-of-the-art performance for disk failure prediction [12], [20]. For RF, we experiment with different numbers of trees, and settle on using 150 trees because of its superiority in performance. For SVM, we use the LIBSVM library [28], and experiment with three different kernels: polynomial, sigmoid and linear kernels. For BP, we use three layers BP with 64 nodes in the hidden layer and use *relu* as the activation function. We set the maximum number of iterations to 1000, the learning rate to 0.01 and adopt Adam [29] for optimization. For our method, we build our scheme upon the code from [26] and set the size of z as 100. For the square image-like representation is the commonly used input shape for CNN, we choose to set T as same as the value of M, i.e., 12.

The balanced training dataset is essential to supervised machine learning methods because the class imbalance issue undermines the prediction performance [21]. To relieve the data imbalance issue for these supervised methods, we apply the commonly used under-sampling method [30] to under-sample the majority class, resulting in different ratios of failed to healthy samples ranging from 1:1 to 1:50. In the final training set, this ratio is set to 1:5 which leads to superior prediction accuracy. Note that, we use the same training set to train SPA, but only healthy samples from the data set are used.

To deal with the model aging problem, we fine-tune our model with the training data collected in the latest month and evaluate the prediction performance of our model on the test set monthly. The reason why we set the model updating interval as one month is to ensure a fair comparison with accumulation update strategy used in [12], [16], in which, once a month, all the data collected from the beginning are used to update the offline models. That is to say, for each month, we build offline models with all the training data collected so far and investigate their performance on the same test set.

Figure 5 and Figure 6 depict the FDRs of these methods on large-size dataset STA and small-size dataset STB, respectively. For the convenience of comparison, as in [12], we measured the FDRs under the constraint that the FARs are around 1.0%. At the beginning of both figures, we observe a similar phenomenon shown in [12] that all the supervised methods exhibit poor performance due to the fact of lacking valid samples, i.e., the cold starting problem. Interestingly, our method achieves high FDRs even in the very beginning, which demonstrates our model is effective in the initial period. The reason is that our model is only trained on healthy

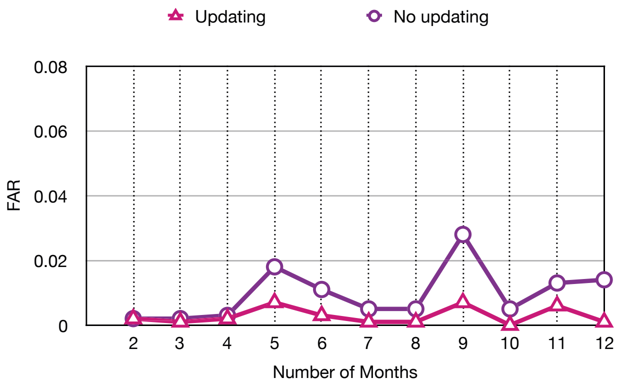

Fig. 9: FARs of model updating and no updating on dataset STA (large-size dataset).

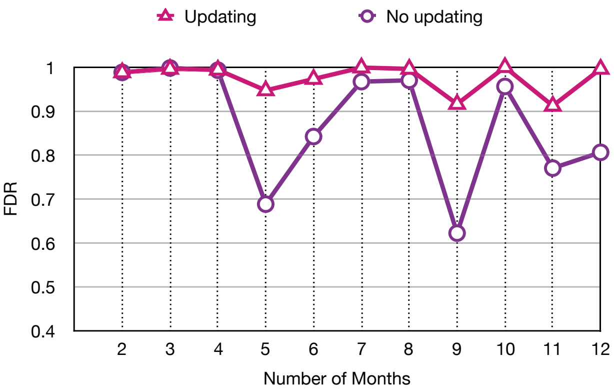

Fig. 11: FDRs of model updating and no updating on dataset STA (large-size dataset).

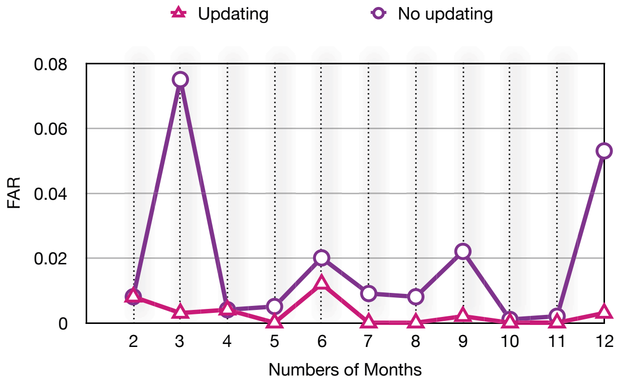

Fig. 10: FARs of model updating and no updating on dataset STB (small-size dataset).

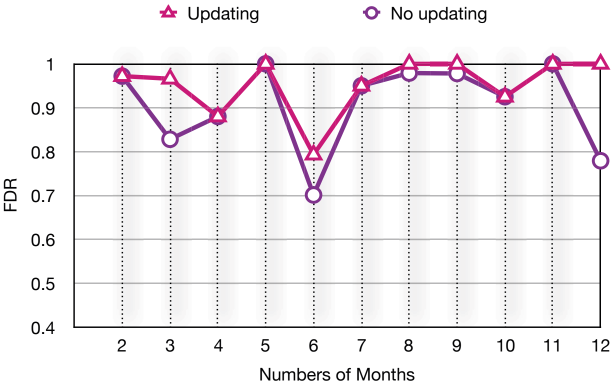

Fig. 12: FDRs of model updating and no updating on dataset STB (small-size dataset).

samples which are sufficient even at the beginning of the model deployment. For the long-term use, our method also outperforms the supervised counterparts. This is because, contrary to supervised learning methods, anomaly detection scheme is able to detect the unknown/unseen anomaly case [26]. In summary, our SPA method outperforms the supervised machine learning based counterparts which use 1D input data, demonstrating its effectiveness in both the initial period and the long-term use, and on both small- and large-size datasets.

D. The Effectiveness of 2D Image-Like Representation

To evaluate the effectiveness of the proposed 2D image-like representation, we train models with different time range T, including 1, 4, 8, and 12 in units of days. Note that SMART attributes are collected on a daily basis. Therefore, $T = 1$ indicates the specific case of 1D-SMART attributes, i.e., no time series data is used. Fig. 7 and Fig. 8 show the FDRs on dataset STA and STB, respectively. As we can see, the model trained with $T = 1$ achieves satisfactory performance, which demonstrates the effectiveness of our adversarial learning strategy. However, it is consistently outperformed by the models trained with other values of T in both figures. These results demonstrate the effectiveness of 2D image-like representations because they exploit the inherent time series features of SMART data. When comparing the performance under different values of T, we observe that the models trained with $T = 12$ consistently outperforms the ones trained with other values of T. Therefore, we set $T = 12$ in the following experiments.

E. The Effectiveness of Model Updating

Although the necessity and effectiveness of model updating have been verified by prior work [12], [19], they are all limited to supervised machine learning models. To estimate the effectiveness of model updating for our semi-supervised method, we compare models trained with updating and without updating in which the previously trained model, i.e., model trained on data from the first month, is used to test the remaining data.

Fig. 9 and Fig. 10 show the FARs of these two models for dataset STA and STB, respectively. These FARs are measured under the constraint that the FDRs are around 85%. It can be seen that when setting FDR around 85%, we can achieve FAR with 0% for models with updating. In other words, 85% failed disks are detected without any false alarms. In addition, Fig. 11 and Fig. 12 show the FDRs of the models. As shown in these figures, although FDRs and FARs of no-updating are acceptable, FDRs and FARs associated with updating are always better. Moreover, the stability of models with updating are also superior to the ones without updating. One possible reason is that the no-updating models are trained with only the samples collected in the first month, which hinders their adaption to the continuous update of forthcoming data [12]. These results suggest that model updating is necessary and effective in anomaly detection. From these results, we also observe fluctuations of FDRs and FARs along with months. This phenomenon is because of the inherent volatility within the data, i.e., there exist significant differences between the number of failed disks and the number of unpredictable failures in each month [12].

V. Related Work

To enhance the failure prediction performance, statistical techniques are proposed based on SMART attributes. Hughes et al. [31] proposed two statistical methods to improve the prediction performance. They regarded the disk failure prediction as an anomaly detection by using Wilcoxon rank-sum test and OR-ed single variate test and achieved a 60% FDR with 0.5% FAR on a dataset composed of 3,744 disks with 36 failed disks. Also from the perspective of anomaly detection, Wang et al. [5], [32] proposed using Mahalanobis distance (MD) for failure detection, which aggregated the input variables into one index and detected failed disks by setting an appropriate threshold. This method delivered a 68% FDR with 0% FAR on the same dataset used by [11], which was a small-size and balanced dataset.

Besides statistical schemes, machine learning methods have also been employed to predict disk failures and demonstrated to outperform the former. Hamerly and Elkan [33] employed two Bayesian approaches named NBEM (Naive Bayes Expectation-Maximization) and supervised naive Bayes classifier, respectively. Both algorithms were tested on a dataset from Quantum Inc., including 1,927 working hard disks and 9 failed disks, and achieved promising prediction performance. Meanwhile, the supervised naive Bayesian classifier was robust against irrelevant attributes. Murray et al. [13] constructed their failure prediction systems based on support vector machine (SVM) and unsupervised clustering respectively and compared them with two non-parametric statistical tests (rank-sum and reverse arrangements test). Experimental results showed that the rank-sum method achieved the best prediction performance with 33.2% failure detection rate at 0.5% FAR. In their subsequent work [11], they designed a novel algorithm by combining multiple-instance learning framework and naive Bayesian classifier, and the results of which showed that the SVM achieved the best performance, 50.6% detection at 0% FAR, with all selected features. However, the rank-sum test outperformed the SVM when using certain small sets of SMART attributes with 28.1% failure detection at 0% FAR. Zhu et al. [19] implemented a backward propagation (BP) neural network model and an improved SVM using SMART attributes as features. Both of these models achieved satisfactory prediction performance with low FARs and the BP neural network model obtained FDR of more than 95%.

All the methods mentioned above dealt with the prediction as an offline training process which suffered from the model aging problem. Recently, [20] and [12] attempted to train prediction models in online mode, and both works achieved high FDRs with low FARs. Moreover, both works demonstrated that random forests achieved superior prediction accuracy. However, these online methods have three drawbacks. First, they all fall into supervised learning framework, which requires a large number of failed samples, limiting their applications in the initial period of model deployment. For instance, in [12], it takes 4-6 months for models to achieve acceptable prediction performance, which is called cold starting problem in this paper. Second, a high prediction performance largely depends on complicated manual feature engineering which is time-consuming and cost-expensive. Last but not least, they need to perform additional operations, i.e., re-sampling [12] or cost-sensitive learning [20], to deal with data imbalance issue.

In this paper, we argue that the cold starting problem is due to data imbalance issue. Therefore, we transform the disk failure prediction into an anomaly detection problem and present a GAN-based semi-supervised method. According to the key idea of anomaly detection, that only healthy disks are need for training, our proposed approach is applicable for small-scale data centers and in the initial period of model deployment. Meanwhile, the use of CNN removes the need of manual feature extraction. Furthermore, our GAN-based model can be fine-tuned to dynamically adapt to new patterns of data and operates without concern of model aging.

VI. Conclusion

We propose a novel GAN-based anomaly detection approach for lifelong disk failure prediction, called SPA, which unlike traditional supervised machine learning methods does not require failed disks to build prediction model. Instead, SPA use only healthy disks for model building, thus detouring the data imbalance issue and eliminating the cold starting problem confronted by supervised counterparts. Moreover, our model is trained end-to-end by leveraging CNN's powerful feature extraction characteristic which captures the temporal locality contained in constructed image-like 2D-SMART attributes. The fine-tuning technique is used for efficient model updating. We evaluate our approach using two real-world datasets. The results confirm that the proposed approach is effective and outperforms supervised counterparts in both the initial period and the long-term use of model deployment.

Acknowledgement

The research is supported by the Innovation Group Project of the National Natural Science Foundation of China under Grant No. 61821003, the National Key Research and Development Program of China under Grant No.2016YFB0800402, and the Innovation Fund of Wuhan National Laboratory for Optoelectronics.

REFERENCES

[1] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population." in *Proceedings of FAST*, vol. 7, 2007, pp. 17–23.

[2] K. V. Vishwanath and N. Nagappan, "Characterizing cloud computing hardware reliability," in *Proceedings of SoCC*, 2010, pp. 193–204.

[3] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS operating systems review*, vol. 37, no. 5, 2003, pp. 29–43.

[4] B. Schroeder and G. A. Gibson, "Disk failures in the real world: What does an mttf of 1, 000, 000 hours mean to you?" in *Proceedings of FAST*, vol. 7, 2007, pp. 1–16.

[5] Y. Wang, Q. Miao, E. W. Ma, K.-L. Tsui, and M. G. Pecht, "Online anomaly detection for hard disk drives based on mahalanobis distance," *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 136–145, 2013.

[6] W. Jiang, C. Hu, Y. Zhou, and A. Kanevsky, "Are disks the dominant contributor for storage failures?: A comprehensive study of storage subsystem failure characteristics," *ACM Transactions on Storage*, vol. 4, no. 3, p. 7, 2008.

[7] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, S. Yekhanin *et al.*, "Erasure coding in windows azure storage." in *Proceedings of ATC*, 2012, pp. 15–26.

[8] Z. Huang, H. Jiang, K. Zhou, C. Wang, and Y. Zhao, "Xi-code: A family of practical lowest density mds array codes of distance 4," *IEEE Transactions on Communications*, vol. 64, no. 7, pp. 2707–2718, 2016.

[9] J. Li, R. J. Stones, G. Wang, Z. Li, X. Liu, and K. Xiao, "Being accurate is not enough: New metrics for disk failure prediction," in *Proceedings of SRDS*, 2016, pp. 71–80.

[10] B. Allen, "Monitoring hard disks with smart," *Linux Journal*, no. 117, pp. 74–77, 2004.

[11] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Machine learning methods for predicting failures in hard drives: A multiple-instance application," in *Journal of Machine Learning Research*, 2005, pp. 783–816.

[12] 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 ICPP*, 2018, pp. 1–10.

[13] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Hard drive failure prediction using non-parametric statistical methods," in *Proceedings of ICANN/ICONIP*, 2003.

[14] Y. Zhao, X. Liu, S. Gan, and W. Zheng, "Predicting disk failures with hmm-and hsmm-based approaches," in *Proceedings of ICDM*, 2010, pp. 390–404.

[15] M. Goldszmidt, "Finding soon-to-fail disks in a haystack." in *Proceedings of HotStorage*, 2012.

[16] J. Li, X. Ji, Y. Jia, B. Zhu, G. Wang, Z. Li, and X. Liu, "Hard drive failure prediction using classification and regression trees," in *Proceedings of DSN*, 2014, pp. 383–394.

[17] W. Yang, D. Hu, Y. Liu, S. Wang, and T. Jiang, "Hard drive failure prediction using big data," in *Proceedings of SRDS Workshops*, 2015, pp. 13–18.

[18] M. M. Botezatu, I. Giurgiu, J. Bogojeska, and D. Wiesmann, "Predicting disk replacement towards reliable data centers," in *Proceedings of KDD*, 2016, pp. 39–48.

[19] B. Zhu, G. Wang, X. Liu, D. Hu, S. Lin, and J. Ma, "Proactive drive failure prediction for large scale storage systems," in *Proceedings of MSST*, 2013, pp. 1–5.

[20] Y. Xu, K. Sui, R. Yao, H. Zhang, Q. Lin, Y. Dang, P. Li, K. Jiang, W. Zhang, J.-G. Lou *et al.*, "Improving service availability of cloud systems by predicting disk error," in *Proceedings of ATC)*, 2018, pp. 481–494.

[21] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley and Sons, 2013.

[22] J. Xu, B. Tang, H. He, and H. Man, "Semisupervised feature selection based on relevance and redundancy criteria," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 9, pp. 1974–1984, 2016.

[23] S. Ha, J.-M. Yun, and S. Choi, "Multi-modal convolutional neural networks for activity recognition," in *Proceedings of SMC*, 2015, pp. 3017–3022.

[24] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Proceedings of IJCNN*, 2016, pp. 381–388.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of NIPs*, 2014, pp. 2672–2680.

[26] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," *arXiv preprint arXiv:1805.06725*, 2018.

[27] BACKBLAZE, "The backblaze hard drive data and stats." https://www.backblaze.com/b2/hard-drive-test-data.html, 2018, [Online; accessed 1-february-2018].

[28] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[31] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, "Improved disk-drive failure warnings," *IEEE Transactions on Reliability*, vol. 51, no. 3, pp. 350–357, 2002.

[32] Y. Wang, E. W. Ma, T. W. 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.

[33] G. Hamerly, C. Elkan *et al.*, "Bayesian approaches to failure prediction for disk drives," in *Proceedings of ICML*, 2001, pp. 202–209.