JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Hard Disk Failure by Support Vector Machine in Federated Learning Setting

[1]Vishal Pandey, [2]Akhilesh A. Waoo, [3]Lokendra Gaur

[1]Student, [2]Professor, [3] Assistant Professor
[1]Computer Science Engineering,
[1]AKS University Satna, Satna, India

Abstract : Traditional approaches to hard disk failure prediction often struggle with the inherent challenge of safeguarding sensitive data when performing centralized analysis. In response to this privacy setback, our study introduces a novel approach by harnessing Federated Learning—a cutting-edge, privacy-preserving machine learning paradigm. In our investigation, we meticulously compare the experimental setups and outcomes of Federated Learning against conventional machine learning methods. Notably, our predictive model of choice is the Support Vector Machine (SVM). Our comprehensive findings underscore that the Federated Learning approach not only rivals traditional methods in terms of predictive accuracy but also substantially bolsters data privacy. This research emphasizes the immense potential of Federated Learning as a transformative solution for addressing privacy concerns, particularly in scenarios where data security is paramount, such as hard disk failure prediction. We delve deeply into the workings of Federated Learning, illustrating its power to securely harness the collective intelligence of decentralized data sources while preserving the confidentiality of sensitive information. The experiment is performed and it has been observed that the proposed approach outperforms the traditional machine learning approach.

IndexTerms **- Hard disk drive, failure prediction, Federated Learning, Support Vector Machine**

I. INTRODUCTION

The modern era's exponential growth in digital data has caused an immense surge in the demand for storage solutions. Among these solutions, hard disks hold a prominent position due to their widespread usage in contemporary computer systems. The significance of hard disk drives (HDDs) in today's data-centric era cannot be overstated. HDDs serve as the primary storage medium for data in various computer systems, ranging from personal computers to large-scale data centers. As the volume of generated and stored data continues to skyrocket, the reliability and durability of HDDs have emerged as crucial considerations. The failure of an HDD can result in the loss of valuable data, leading to substantial financial and operational implications for individuals and organizations. However, the problem of hard disk failures persists, posing risks of significant data loss and system downtime. To address these risks, there has been a growing interest in the development of predictive models that can anticipate and prevent such failures before they manifest.

The emergence of Big Data and Artificial Intelligence (AI) has greatly enhanced the capacity to forecast hard disk failures. These technologies enable the collection, processing, and analysis of vast data volumes, facilitating the identification of patterns and the provision of accurate predictions. The accurate prediction of hard disk failures is crucial for maintaining the reliability and availability of storage systems [1]. However, conventional approaches often involve the collection of sensitive data from hard disks, giving rise to concerns regarding privacy and data security. In recent years, there has been a growing interest in privacy-preserving techniques for predicting hard disk failures. These methods aim to safeguard sensitive data while still achieving accurate failure prediction.

One emerging approach in privacy-preserving machine learning is federated learning. This approach enables the collaborative training of models across multiple devices or entities without the need to share raw data [3]. By utilizing federated learning, privacy-preserving methods can be employed to predict hard disk failures while maintaining the privacy of sensitive information. This allows organizations to strike a balance between ensuring the reliability of storage systems and protecting the privacy of individual disk data. Consequently, privacy-preserving approaches, such as federated learning, offer a promising avenue for advancing hard disk failure prediction in a privacy-conscious manner.

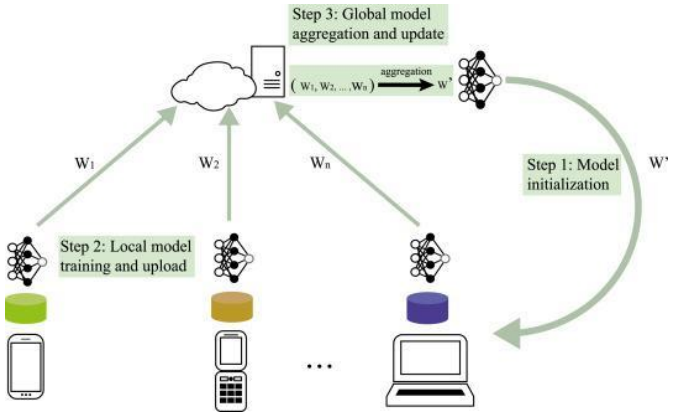

Fig.1: Federated Learning Architecture [2]

Federated learning operates within a decentralized machine learning setting, where models are trained on data from various sources without directly transferring or centralizing the data. It enables multiple devices or entities to collaboratively learn a shared model while keeping their data local and private. In contrast to traditional machine learning approaches that centralize data for training, federated learning addresses concerns related to data privacy, security, and bandwidth usage [1]. By distributing the model training process across multiple devices or entities like mobile phones, IoT devices, or edge servers, federated learning ensures privacy while still achieving effective model training.

Figure 1 illustrates the architecture of federated learning, which begins with the initialization of a global model on a central server. This global model is then shared with specific devices or entities. Each device independently conducts local training on the global model using its own data, updating the model parameters accordingly. The updated models or gradients are sent back to the central server, which aggregates them to create an enhanced global model. This updated model is then distributed to the participating devices, and the iterative process of local training, model update, aggregation, and distribution continues until convergence or the desired performance is achieved. Ultimately, the trained global model can be deployed for inference or analysis. This federated learning approach ensures data privacy, encourages collaboration on distributed data, and enables efficient and secure machine learning on decentralized devices.

Federated learning holds great potential in the realm of hard disk failure prediction by leveraging the collective knowledge of multiple devices or servers without compromising data privacy. By employing federated learning, organizations can tap into the collective intelligence of their distributed hard disks while upholding data privacy. The collaborative nature of federated learning enhances the accuracy and robustness of hard disk failure prediction models, leading to improved maintenance strategies, reduced downtime, and enhanced data protection.

This work focuses on utilizing a federated learning approach to predict hard disk failures. The adopted privacy-preserving approach aims to strike a balance between achieving accurate predictions and safeguarding the confidentiality of the data. By implementing privacy-preserving techniques, this research aims to prevent unauthorized access or disclosure of the data, thereby safeguarding sensitive information related to hard disk performance and health. The significance of this research lies in its potential to offer a solution that enables organizations to effectively predict hard disk failures while mitigating the risks associated with data privacy. By addressing this research gap, this work contributes to the advancement of privacy-preserving methods in the field of hard disk failure prediction.

The remainder of this paper is organized as follows: Section 2 provides a concise overview of related work in HDD failure prediction. Section 3 outlines the methodology, dataset, and preprocessing techniques employed in this study. Section 4 discusses the results of our experiments and compares them with existing literature. Finally, Section 5 concludes the paper, highlighting the future scope of our study.

II LITERATURE REVIEW

Significant variations exist in the characteristics and degradation processes associated with different types of HDD failures [5]. These discrepancies stem from the complex structure and inherent mechanisms inherent in HDDs. An HDD comprises four critical components: the head-disk interface, the head stack assembly, spindle motors/bearings, and the electronics module. Because these components function differently, the underlying causes of their failures also differ. For instance, the head-disk interface and spindle motors/bearings are mechanical components susceptible to wear-and-

tear failures, whereas the control board is an electronic component prone to short circuits. Wear-and-tear failures typically exhibit a significantly prolonged degradation process compared to other failure types, rendering them more predictable.

Numerous research efforts have been dedicated to the classification of HDD failures. In a study conducted by [6], HDD failures were categorized based on their mode, cause, and underlying mechanisms. In contrast, [5] divided HDD failures into three primary groups: logical failures, bad sector failures, and read/write head failures. Dedicated classifiers were developed for each of these groups, yielding impressive accuracy. It was suggested that incorporating a form of differential prediction, utilizing the similarity between healthy HDD samples, could enhance the precision and efficiency of failure prediction. However, it is worth noting that empirical evidence indicates that relying solely on failure classification may not provide the necessary accuracy to fully optimize the effectiveness of failure prediction. To address this challenge, this work proposes an alternative approach by considering the utilization of the similarity between healthy HDD samples to implement a coarse-grained form of differential prediction. This approach draws inspiration from similarity measures, which have previously proven successful in recommendation systems [7,8], and offers a potentially viable solution to improve the accuracy of HDD failure prediction.

Furthermore, the SMART (Self-Monitoring, Analysis, and Reporting Technology) feature integrated within HDDs plays a crucial role in this context. SMART is designed to collect performance data through sensors or counters that correspond to record counts or physical units [9]. These SMART attributes encompass around 30 internal drive metrics, including reallocated sector count (RSC), spin-up time (SUT), seek error rate (SER), temperature in Celsius (TC), and power-on hours (POH). These attributes serve as valuable indicators of the HDD's overall health status and its internal operating conditions. For instance, the RSC value signifies the number of defective disk sectors and can provide insights into the health of the disk media, while changes in SUT and TC are closely linked to the operational state of the spindle motor. This rich source of data contributes significantly to the understanding and prediction of HDD failures.

SMART attributes demonstrate a strong association with HDD failures. Hamerly and Elkan [23] introduced a predictive approach based on three attributes: grown defect count, read soft errors, and seek errors, achieving superior accuracy compared to using all available attributes. This underscores the notion that not all SMART attributes hold equal efficacy in predicting failures. In their study [9], researchers identified significant correlations between scan errors, RSC (reallocated sector count), offline RSC, and HDD failures. Moreover, [10] highlighted latent sector errors as a primary cause of HDD failure, with RSC emerging as the most crucial attribute for predicting impending failures. In [6], distinctive indicators for the failure of various internal HDD components were observed, leading to the establishment of SMART attribute priorities within a prediction model based on the frequency and severity of relevant failures. Furthermore, [5] demonstrated that SMART attributes exhibit varying degrees of correlation with different types of HDD failures. Consequently, the successful prediction of HDD failures hinges on the use of SMART attributes strongly associated with each specific failure type, rather than a uniform combination of attributes. To achieve this, we employ suitable sub-classifiers within the ensemble, utilizing these selected sub-classifiers to contribute to the prediction outcome for each set of healthy samples. The amalgamation of sub-classifiers for different sample groups adheres to non-uniform classification rules.

In the realm of statistical approaches, common methods include the rank-sum test and Bayesian approaches. In [11], it was discovered that many SMART attributes have non-parametric distributions, leading to the application of a multivariate rank-sum test, along with OR-ed single variate tests, to a dataset of 3744 drives, including 36 failures. Their approach achieved a false discovery rate (FDR) of 60% with a false alarm rate (FAR) of 0.5%. Subsequent studies frequently utilized the rank-sum test in feature selection.

Based on an extensive review of the literature on hard disk failure prediction, a notable gap emerges within the domain concerning data privacy considerations. The existing research landscape appears to overlook the critical aspect of data privacy when developing models for predicting disk failures. Hence, a model should be developed which could predict the disk failure in a privacy preserving approach.

III. METHODOLOGY

The work proposes to develop a machine learning model using Federate Learning and Support Vector Machine to predict the failure of hard disks. SVMs are well-suited for binary classification tasks, where the objective is to distinguish between healthy and failing hard disks, a common goal in failure prediction. Support Vector Machines (SVM) and Federated Learning can be combined effectively for hard disk failure prediction in a distributed and privacy-preserving manner. SVM is a machine learning algorithm that can be used for classification tasks, such as predicting hard disk failures. In the context of hard disk failure prediction, SVM can be trained on historical data that includes various attributes of hard disks (e.g., temperature, read/write errors, SMART data) and their corresponding labels (failed

or not failed). SVM aims to find the hyperplane that best separates these two classes while maximizing the margin between them. This hyperplane can then be used to classify new, unlabeled hard disks as either at risk of failure or not. SVM's ability to work well in high-dimensional spaces and handle non-linear data via kernel functions makes it a suitable choice for this task [12].

Federated Learning: Federated Learning is a privacy-preserving machine learning approach that enables model training across multiple decentralized devices or data sources without sharing sensitive data. In the context of hard disk failure prediction, federated learning can be utilized to leverage data from different hard drives located in various locations while keeping the data localized and private. The following steps represent how a model training will take place by combining SVM and Federated Learning for prediction of hard disk failure.

1**. Data Distribution:** Data from various hard drives is kept on their respective devices or data centers. Each device holds information about the performance and health of its hard disk but does not share this data centrally.

2. **Model Initialization:** Initially, a global SVM model is created with random or predefined parameters. This model will serve as the starting point for the federated learning process.

3. **Local Model Updates:** On each device, the local SVM model is trained using the data available on that device. This involves computing support vectors and the optimal hyperplane for that local dataset.

4. **Model Aggregation:** The updated local models (hyperplanes) are not shared directly. Instead, only the model updates (changes in support vectors and weights) are shared with a central server or aggregator.

5. **Global Model Update:** The central server aggregates these model updates from all devices, combining the knowledge learned from different hard drives while keeping the raw data decentralized and private.

6. **Iteration:** Steps 3 to 5 are repeated iteratively for a specified number of rounds. Each round allows the global model to improve based on the collective knowledge from all devices.

7. **Final Model:** After the desired number of iterations, the final global SVM model is used for making predictions. This model benefits from the collective intelligence of all devices without sharing sensitive data centrally.

Combining SVM with federated learning in hard disk failure prediction offers the advantage of utilizing distributed data sources without compromising privacy and security. It allows organizations to harness insights from a wide range of hard drives while ensuring that sensitive data remains local and protected. This approach can improve the accuracy and reliability of hard disk failure predictions, ultimately reducing the risk of data loss and system downtime.

In this experiment, we aim to assess the performance of a federated learning system utilizing a Support Vector Machine (SVM) model for hard disk failure prediction, involving 10 simulated clients representing separate data sources. We will systematically vary the number of communication rounds as an experimental parameter while maintaining a constant learning rate of 0.01 and conducting 20 training epochs per round. Each client possesses its own dataset, containing historical hard disk performance data, including attributes related to temperature, read/write errors, SMART data, and failure labels. The Backblaze Hard Drive Stats dataset [12] is taken for the entire experiment. The federated learning process entails initializing a global SVM model, iteratively updating it through communication rounds, aggregating model updates from clients, and evaluating performance metrics such as accuracy and F1-score to understand the trade-offs between communication frequency and model efficacy.

IV. RESULTS AND DISCUSSION

The experiment is performed for the proposed approach in which SVM is applied in a federated learning approach to predict the hard disk failure. The corresponding obtained result is shown in table I. The accuracy and F1 score is taken as a performance matrix to validate the proposed work. Accuracy is a fundamental metric in classification, quantifying the ratio of correctly predicted instances to the total predictions. It offers an overall assessment of a model's performance but may not suit imbalanced datasets. In contrast, the F1 score combines precision and recall into a single value, making it particularly valuable for imbalanced data or when striking a balance between precision (correct positive predictions) and recall (capturing all positives) is crucial.

Table I: Performance Comparison Data

Experiment	Approach	Model	Round	Accuracy	F1 Score	Loss
1	Federated Learning	SVM	10	70.36	5.558	0.062
2	Federated Learning	SVM	30	80.55	1.345	0.510
3	Federated Learning	SVM	50	85.03	1.517	0.467
4	Federated Learning	SVM	100	92.14	2.474	0.290
5	Traditional ML	SVM	NA	91.26	1.745	0.418

From the results in Table I it can be analyzed that, as the number of communication rounds increases in the Federated Learning approach (from 10 to 100), the accuracy of the SVM model also increases significantly. This suggests that more communication and collaboration among the participants in Federated Learning leads to improved model performance. The accuracy achieved in the Federated Learning approach with 100 communication rounds (92.14%) is higher than the accuracy achieved in the Traditional ML approach (91.26%). This indicates that Federated Learning outperforms the Traditional ML approach in this context.

In terms of F1 score and loss, the Federated Learning approach with 100 communication rounds also outperforms the other experiments, suggesting that it strikes a good balance between precision and recall and converges well during training. In summary, the table shows that Federated Learning with a Support Vector Machine model, especially with a higher number of communication rounds, outperforms the Traditional ML approach in terms of accuracy, F1 score, and loss, indicating its effectiveness in this particular experiment.

V. Conclusion and Future Scope

In this study, we employed Federated Learning with SVM to predict hard disk failures and compared its performance with traditional machine learning methods. The results clearly indicate that Federated Learning outperforms traditional ML in the context of hard disk failure prediction. This achievement can be attributed to the collaborative and privacy-preserving nature of Federated Learning, which allows for model training on decentralized data sources without compromising data privacy. While the present study establishes the effectiveness of Federated Learning with SVM for hard disk failure prediction, the future work can Investigate the applicability of other advanced machine learning algorithms within the Federated Learning framework to further enhance prediction accuracy and robustness.

REFERENCES

[1] S. Huang, S. Fu, Q. Zhang and W. Shi, "Characterizing Disk Failures with Quantified Disk Degradation Signatures: An Early Experience," 2015 IEEE International Symposium on Workload Characterization, Atlanta, GA, USA, 2015, pp. 150-159, doi: 10.1109/IISWC.2015.26.

[2] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, *216*, 106775.

[3] Yang, Y., Lu, Y., & Mei, G. (2023). A federated learning based approach for predicting landslide displacement considering data security. *Future Generation Computer Systems*, *149*, 184-199.

[4] Shen J, Wan J, Lim S-J, Yu L. Random-forest-based failure prediction for hard disk drives. International Journal of Distributed Sensor Networks. 2018;14(11). doi:10.1177/1550147718806480

[5] Huang S, Fu S, Zhang Q, et al. Characterizing disk failures with quantified disk degradation signatures: an early experience. In: 2015 IEEE international symposium on workload characterization (IISWC), Atlanta, GA, 4–6 October 2015, pp.150–159. New York: IEEE.

[6] Wang Y, Miao Q, Pecht M. Health monitoring of hard disk drive based on Mahalanobis distance. In: Prognostics and system health management conference (PHM-Shenzhen), Shenzhen, China, 24–25 May 2011, pp.1–8. New York: IEEE.

[7] Xu Y, Jianwei Y. Collaborative recommendation with user generated content. *Eng Appl Artif Intel* 2015; 45: 281–294

[8] Xu Y, Jianwei Y, Shuiguang D, et al. Context-aware QoS prediction for web service recommendation and selection. *Expert Syst Appl: Int J* 2016; 53(C): 75–86.

[9] Pinheiro E, Weber W-D, Barroso LA. Failure trends in a large disk drive population. In: FAST'07 proceedings of the 5th USENIX conference on file and storage technologies, San Jose, CA, 13–16 February 2007, vol. 7, pp.17–23. Berkeley, CA: USENIX Association.

[10] Ma A, Douglis F, Lu G, et al. RAIDShield: characterizing, monitoring, and proactively protecting against disk failures. In: Proceedings of the 13th USENIX conference on file and storage technologies, Santa Clara, CA, 16–19 February 2015, pp.241–256. Berkeley, CA: USENIX Association.

[11] Hughes GF, Murray JF, Kreutz-Delgado K, et al. Improved disk-drive failure warnings. *IEEE T Reliab* 2002; 51(3): 350–357.

[12] Backblaze Hard Drive Stats — backblaze.com. https://www.backblaze.com/b2/hard-drive-test-data.html. [Accessed 15-May-2023].