International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Hard Disk Failure Prediction using Federated Learning: A Privacy Preserving Approach

*Vishal Pandey, Lokendra Gaur, Akhilesh A. Waoo**

AKS University Satna, Satna, Sherganj, Satna Madhya Pradesh

ABSTRACT

In the era of artificial intelligence and big data, storage and computation resources are become very important as they are the soul of data analysis and prediction strategies. In this context, the role of the hard disk increases to the very extent as it stores all the data & pieces of information and any interruption in it may cause a big issue. So, it is necessary to deal with the issues occurring with hard disks. In the current scenario, hard disk failure is a big issue, and any prior prediction of hard disk failure would be very useful. Traditional machine learning models are used to predict the failure of a Hard Disk using the data associated with it to be stored on the central server. But due to increased rules and regulations, the sharing of data is not so much easy and safe now. So, it is very necessary to create such a predictive model in the privacy-preserving approach. Hence this work proposes a Federated Learning (FL) for failure prediction, in which the training of machine learning models will be done in such a way that none of the data is

required to be shared with a central server for training. In this work, the FL is used for the training of the hard disk failure prediction model by using Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) Machine Learning models. The predictive model achieved maximum accuracy of 90.66%, which shows its significance in predicting the failure more accurately in the privacy-preserving approach.

Keywords: Hard disk drive, failure prediction, Federated Learning, machine learning

1. Introduction

The importance of hard disk drives (HDDs) in the modern era of data-driven applications cannot be overstated. These devices are the primary storage media for data in most computer systems, ranging from personal computers to large data centers. As the amount of data being generated and stored continues to increase exponentially, the reliability and durability of HDDs have become a critical concern. The failure of an HDD can result in the loss of valuable data, which can have significant financial and operational consequences for individuals and organizations alike. Therefore, the ability to predict HDD failures accurately can be essential for taking proactive measures to prevent data loss and minimize downtime.

The exponential growth of digital data in the modern era has led to a tremendous increase in demand for storage solutions. Hard disks, being one of the most widely used storage devices, are critical to the functioning of modern computer systems. However, hard disk failures remain a persistent problem that can lead to significant data loss and system downtime. To mitigate the risks associated with hard disk failure, there has been a growing interest in developing predictive models that can anticipate and prevent such failures before they occur.

With the advent of Big Data and Artificial Intelligence (AI), there has been a significant improvement in the ability to predict hard disk failures. These technologies have enabled the collection, processing, and analysis of vast amounts of data to identify patterns and make accurate predictions.

Predicting hard disk failures is a critical task in maintaining the reliability and availability of storage systems. However, traditional approaches often involve collecting sensitive data from hard disks, raising concerns about privacy and data security. In recent years, there has been growing interest in privacy-preserving approaches for hard disk failure prediction. Privacy-preserving methods aim to protect sensitive data while still allowing for accurate prediction of disk failures. One growing approach to privacy-preserving machine learning is federated learning, where models are trained collaboratively across multiple devices or entities without sharing raw data. These privacy-preserving approaches enable organizations to predict hard disk failures while preserving the privacy of sensitive information. By implementing federated learning to predict the failure of the disk, organizations can strike a balance between maintaining the reliability of storage systems and protecting the privacy of individual disk data. Such approaches provide a promising avenue for advancing hard disk failure prediction in a privacy-conscious manner.

Federated learning is a kind of machine learning setting that enables training models on decentralized data sources without the need to directly transfer or centralize the data. It allows multiple devices or entities to collaboratively learn a shared model while keeping their data local and private. In traditional machine learning, data is typically collected from various sources and centralized in a single location for model training. However, this approach may raise concerns about data privacy, security, and bandwidth usage. Federated learning addresses these challenges by distributing the model training process across multiple devices or entities, such as mobile phones, IoT devices, or edge servers.

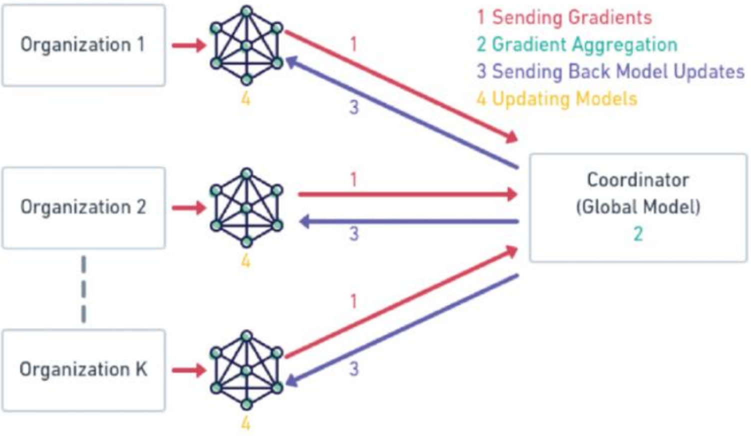


Fig. 1: Architecture of Federated Learning [6]

Figure 1 represents the architecture of federated learning in which the process starts with the initialization of a global model on a central server, which is then shared with selected devices/entities. Each device independently performs local training on the global model using its own data, updating the model parameters. The updated models or gradients are sent back to the central server, which aggregates them to create an improved 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 desired performance is achieved. Finally, the trained global model can be deployed for inference or analysis. This approach preserves data privacy, allows collaboration on distributed data, and enables efficient and secure machine learning on decentralized devices.

Federated learning can significantly contribute to hard disk failure prediction by leveraging the collective knowledge of multiple devices or servers without compromising data privacy. By leveraging federated learning, organizations can harness the collective intelligence of their distributed hard disks while preserving 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.

The main objective of this work is to utilize a federated learning approach for predicting hard disk failures. The privacy-preserving approach adopted in this study aims to strike a balance between achieving accurate predictions and safeguarding the confidentiality of the data. By implementing privacy-preserving techniques, the research work aims to prevent unauthorized access or disclosure of the data, protecting sensitive information related to hard disk performance and health. The significance of this research lies in its potential to provide a solution that enables organizations to effectively predict hard disk failures while mitigating the risks associated with data privacy. By addressing the research gap, this work aims to contribute to the development 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 brief overview of related work in HDD failure prediction. Section 3 describes the methodology , dataset and preprocessing techniques used in this study. Section 4 discusses the results of our experiments and compares them with the existing literature. Finally, Section 5 concludes the paper and highlights the implications of our study.

2. Literature Review

The authors in [6] explore the use of Federated Learning (FL) in the Life Insurance Industry for risk prediction. It emphasizes the importance of evaluating customer applications while addressing data privacy concerns. The study utilizes the Kaggle Prudential Life Insurance Assessment dataset and employs the Dirichlet Process to simulate diverse data distributions among clients. The results validate the effectiveness of FL for risk prediction in the Life Insurance Industry. Overall, this review highlights the practical application of FL in this domain, showcasing its benefits for collaborative machine learning while ensuring data security. While in [7], authors focused on developing a privacy-preserving CNN model for Iris recognition using a federated learning approach. Biometric identifiers, such as Iris recognition, offer secure and unique authentication. The sensitivity of Iris data necessitates secure handling. The proposed approach ensures privacy by avoiding data sharing with a central server, unlike traditional machine learning methods. The implemented CNN model is evaluated for various combinations of participating clients and global rounds, providing performance insights. This research contributes to the field of privacy-preserving Iris recognition in federated learning, addressing the growing concerns of security and identification in advanced technology domains.

In [4], the authors have taken a unique approach by combining two different and complex data sources, namely S.M.A.R.T data and Windows performance counters, to develop and deploy a predictive model for hard disk drive failure. The authors have detailed the process of parsing and transforming the data to create a classification problem and have tested different machine learning and statistical modelling techniques to arrive at the best-performing two-stage ensemble model. The authors believe that their approach can be applied to other hardware components as well, and the successful execution of the hard disk failure prediction model can improve the economics of running a large-scale cloud service.

The authors in [8] present a comprehensive analysis, and the remarkable findings derived from an extensive study focused on disk failure prediction. This study stands as one of the largest endeavors of its kind, encompassing a staggering cohort of 380,000 hard drives. Over a span of two months, these drives were carefully observed and monitored across 64 sites, all belonging to a leading data center operator. The primary objective of this research was to shed light on the enigmatic nature of disk drive failures. By harnessing the power of machine learning, the researchers aimed to develop robust models capable of predicting such failures with utmost precision. The proposed models were evaluated based on two key performance metrics: the F-measure, which measures the balance between precision and recall, and the Matthews correlation coefficient (MCC), a statistic that captures the overall classification performance. The results obtained from this extensive study showed good performance while predicting the failure of the disk.

Predicting HDD failure has become possible by utilizing attributes such as Self-Monitoring and Reporting Technology (SMART), which are collected by HDD manufacturers during normal operations. The use of SMART attributes and threshold values has become a common method for predicting disk drive failures. However, relying solely on threshold values often results in high false positive rates, leading to the unnecessary replacement of healthy disks, which is both costly and inefficient. SMART attributes, such as the number of command timeouts, scan errors, and reallocation counts, provide valuable insights into HDD health statistics. To accurately and proactively predict HDD failures, this study explores the application of Machine Learning techniques, as they are well-suited for solving learning and rare event-based problems. Through the evaluation of Random Forest, Decision Tree, and Naive Bayes algorithms, the study aims to identify the most accurate model for HDD failure prediction [11].

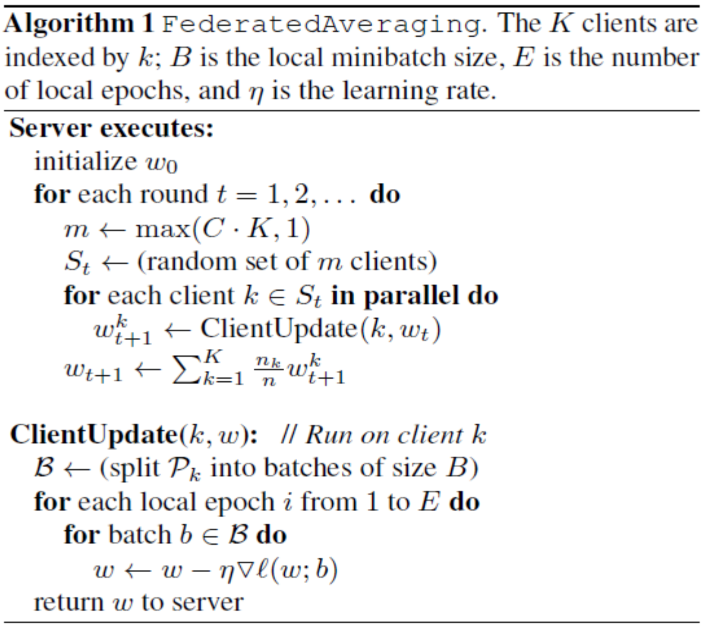
The study [5] proposes a method to identify disks with media failures in a production environment and utilizes supervised machine-learning techniques to predict disk failures. A crucial aspect of the proposed approach is the automated labelling of disks as "healthy" or "at-risk" during the training and validation stages. The paper presents a detailed description of this labelling stage and outlines a tuning strategy to optimize the hyperparameters of the associated machine learning classifier. To evaluate the effectiveness of the approach, it is trained and validated using a large dataset consisting of 65,000 hard drives from the CERN computer center. The achieved results are thoroughly discussed, providing insights into the performance and efficacy of the proposed method.

After conducting a literature survey for works [4]-[8], [11], a research gap has been identified in the area of data privacy. As the authors of all the previous works have focused on the training of ML models to get good performance, it is also very necessary to preserve the privacy of data being used in the training. Due to strict rules and regulations like GDPR [2] the data are not easily shared, causing a big issue in the training process. Hence with the increasing importance of data privacy and the sensitivity of the information involved in predicting hard disk failures, it is crucial to address the privacy issue of data. Therefore, this research work aims to develop a solution that ensures the privacy of the data such that none of the data is shared while the training process, along with accurately predicting hard disk failures.

Methodology

This section of this study presents the key aspects of the research design, including the dataset, methods, and models employed in predicting hard disk drive (HDD) failures. This section provides an overview of the data collection process, the selection and preprocessing of the dataset, the implementation of machine learning algorithms, and the construction of the predictive model.

A. **Federated Learning Algorithm:** The federated learning employs FedAvg aggregation algorithm to aggregates model updates of all participating clients. The Algorithm 1 provides the FedAvg [9] algorithm which is used in the proposed work.

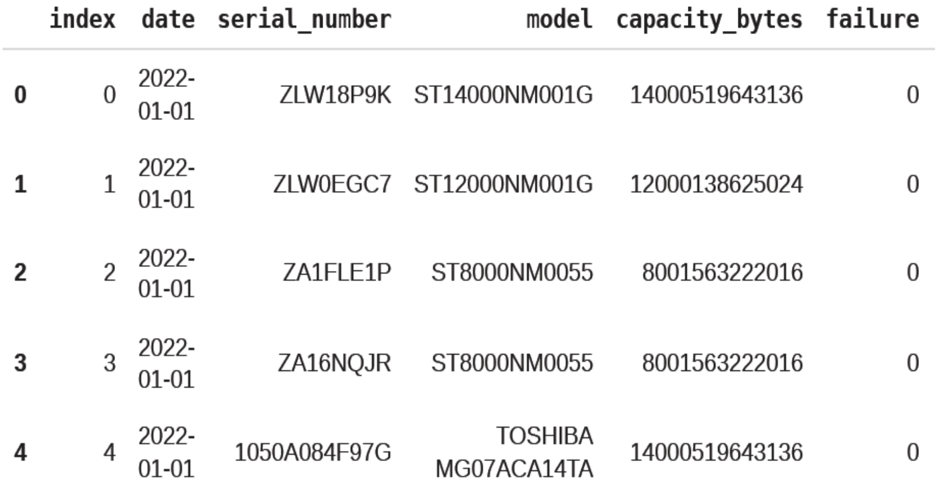


Algorithm 1 FederatedAveraging. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate($k, w$):   // Run on client $k$
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

B. **Dataset:** The dataset used in the proposed work is taken from Backblaze [1], which provides the dataset related to hard disks. It has published data on the failure rates of the hard drives they have used over the years. The data covers various brands and models of hard drives and includes information on the number of drives they have tested, how long the drives were used before failure and the reasons for the failures. The size of the entire dataset is 4139075, which consists of two classes. Each record in the dataset has a meta-data entry that comprises the following information represented in figure 2.



	index	date	serial_number	model	capacity_bytes	failure
0	0	2022-01-01	ZLW18P9K	ST14000NM001G	14000519643136	0
1	1	2022-01-01	ZLW0EGC7	ST12000NM001G	12000138625024	0
2	2	2022-01-01	ZA1FLE1P	ST8000NM0055	8001563222016	0
3	3	2022-01-01	ZA16NQJR	ST8000NM0055	8001563222016	0
4	4	2022-01-01	1050A084F97G	TOSHIBA MG07ACA14TA	14000519643136	0

Fig. 2: Dataset Overview

C. **The CNN Architecture:** A CNN is made up of three layers: a conv layer, a pooling layer and a fully connected layer. The conv layer is the foundation of CNN. It is mainly accountable for the calculation load of the network. The pooling layer replaces the output of the network at certain points using summaries of neighboring outcomes. This decreases the size of the depiction, reducing the number of calculations and loads required. This layer's neurons have perfect connectivity among all neurons in the previous and next layers, just like in a regular FCNN. It may be computed using matrix multiplication accompanied by a bias effect [12]. Figure 3 represents the architecture of CNN.

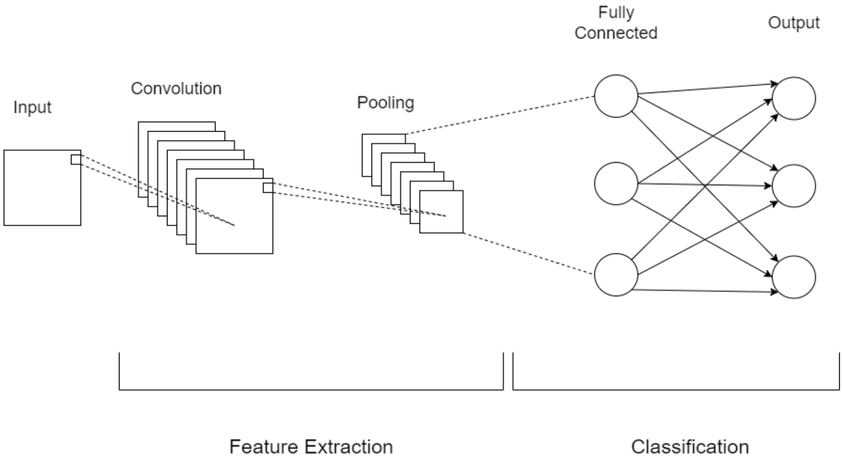


Fig. 3: CNN Architecture

D. **Recurrent Neural Networks:** RNNs are a class of neural networks that are particularly effective in handling sequential data, such as text, speech, or time series data. RNNs have a recurrent internal connection that allows them to maintain the memory of past information and process sequences of arbitrary length. Figure 4 represents the architecture of RNN. The key feature of RNNs is their ability to capture dependencies and patterns in sequential data by propagating information from previous steps to the current step. This makes them suitable for tasks such as language modelling, machine translation, sentiment analysis, and text classification. The basic building block of an RNN is the recurrent cell, which is typically a simple recurrent unit like the Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU). These units allow RNNs to address the vanishing gradient problem by selectively retaining or forgetting information [10].

E. **Experimental Setup:** The experimentation is carried out using TensorFlow Federated library [3]. To carry out federated learning, an environment with 5 clients is adopted, and the number of rounds is 5. In this experiment, the data is equally shared among all the clients. The FedAvg algorithm is used for the aggregation of all the model updates. The CNN and RNN modal were employed for the training of local models. In the experiment, the batch size is taken as 80, momentum is taken as 0.5, while the learning rate was set to 0.01.

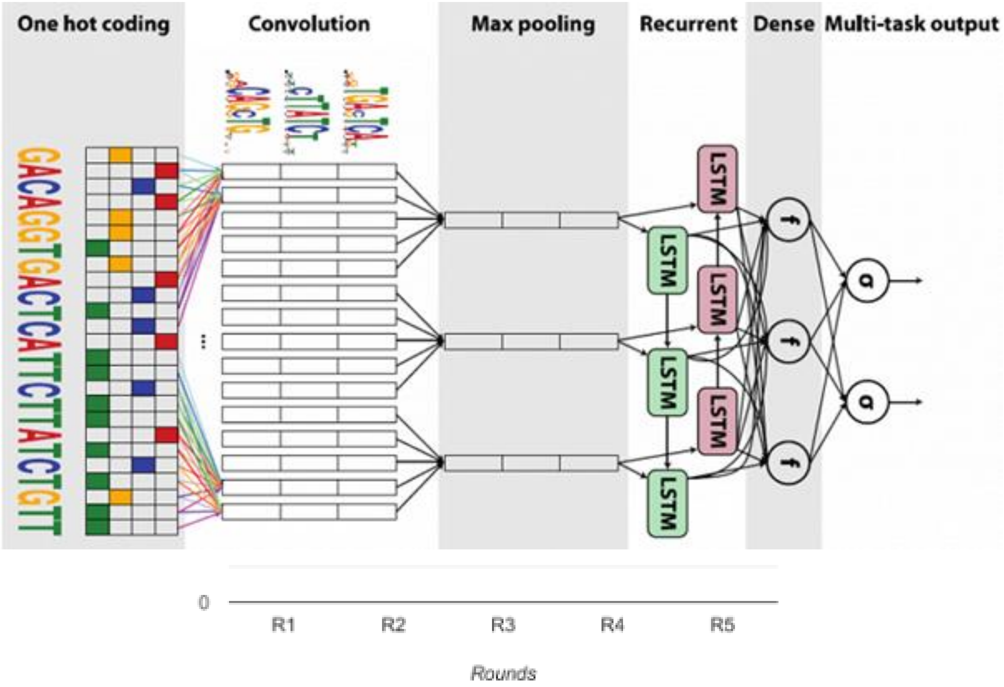


Fig. 4: Recurrent Neural Network Architecture

F. **Performance Metric:** The most basic categorization performance metric is accuracy. It can handle binary as well as multi-class classification issues. Accuracy is calculated as the ratio of true outcomes to total number of instances analyzed. It is a legitimate method of evaluation for well-balanced and non-skewed classification issues. The equation 1 represents the mathematical formulation of the accuracy metric.

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FP+FN+TN})} \qquad (1)$$

4. Result and Analysis

The experiment with the proposed work was employed, and corresponding results containing accuracy as a performance metric and loss were recorded. These results are shown in the form of a graph with its explanation.

Figure 5 represented the accuracy in every round when the experiment was conducted with the CNN model. In the provided graph, it can be observed that the accuracy of the global model is increasing at every round. At the end of the fifth round, the corresponding final accuracy of the global modal was 89.36\%, while the accuracy for the RNN model was 90.66\%. Similarly, it can also be observed in figure 6, which represents the global loss at every round of training. In this graph, it is crucial to analyze that, as the rounds are increasing, the global model's loss is decreasing, which is caused by the aggregation of the client's model updates and enhanced trained weights in every round.

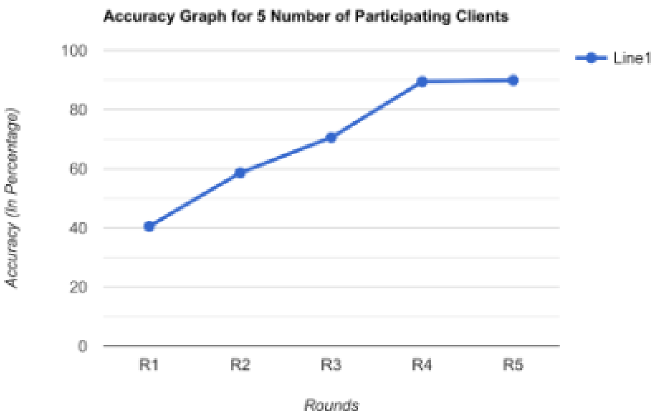


Fig. 5: Aaccuracy Graph for Global Model using 5 clients, 5 rounds and CNN model

Figure 7 represents the accuracy of individual participating clients at the end of every round. Once all the client completes their training, their model updates are sent to the server and get aggregated. The aggregated model is again sent back to all the participating clients for the next round of training. In this graph, it can be observed that the performance of every client is increasing at every round. As in every round, the client gets the aggregated weights which are trained by the data contained by all the clients.

So, after analyzing the results obtained in the performed experiment, it has been observed that the model is performing well as traditional machine learning. But the proposed model provides an additional feature of preserving the privacy of the data, which is not shared with the central server.

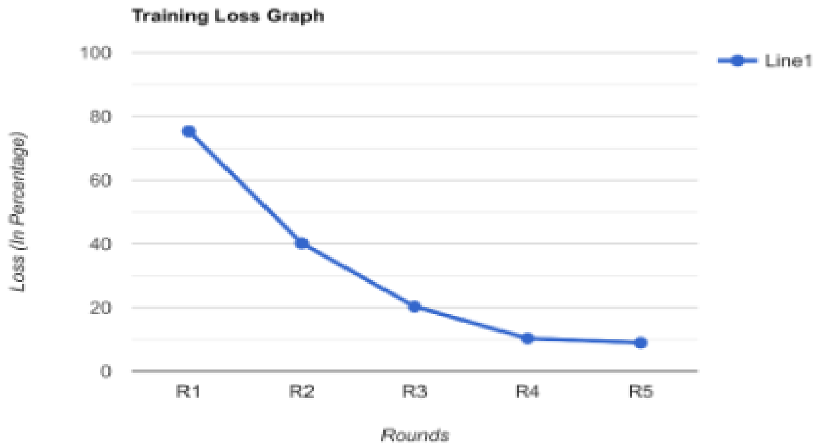


Fig. 6: Global Loss Graph for in experiment with using 5 clients, 5 rounds and CNN model

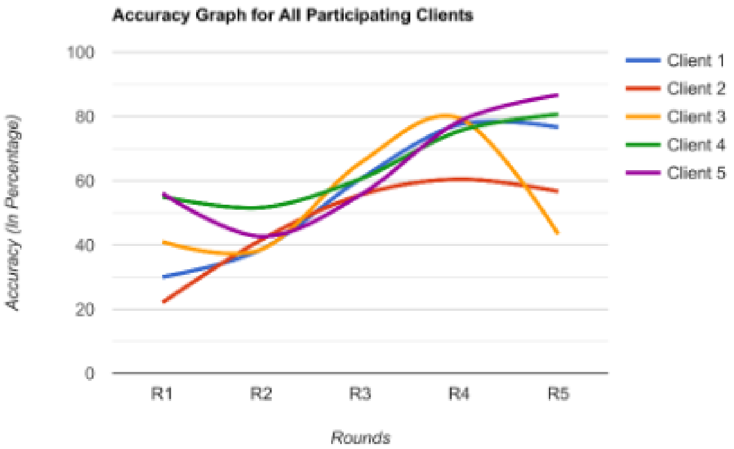


Fig. 7: Accuracy Graph for All Participating Clients in Each Round in experiment with using 5 clients, 5 rounds and RNN model

5. Conclusion and Future Work

The work proposes to train a machine learning model in a privacy-preserving approach to predict the failure of hard disks. As the proposed technique uses federated learning to train the model in a privacy-preserving approach by providing good performance. The obtained results are significant improvements over traditional machine learning training methods and demonstrate the potential of using machine learning algorithms for predicting hard disk failures in a privacy-preserving approach. However, challenges still exist in this area, such as the need for large amounts of high-quality data and the need to develop more sophisticated algorithms that can adapt to changing conditions. These challenges present exciting opportunities for further research and development in this field.

Overall, the success of the research studies reviewed in this paper highlights the importance of training methodology for improving hard disk failure prediction. As data continues to grow and become more complex, the development of more accurate and efficient predictive models will be crucial for ensuring the reliability and stability of computer systems in the future.

References

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

[2] General Data Protection Regulation (GDPR) – Official Legal Text -gdpr-info.eu. https://gdpr-info.eu/. [Accessed 18-May-2023].

[3] TensorFlow Federated — tensorflow.org. https://www.tensorflow.org/federated. [Accessed 17-May-2023].

[4] Sandipan Ganguly, Ashish Consul, Ali Khan, Brian Bussone, Jacqueline Richards, and Alejandro Miguel. A practical approach to hard disk failure prediction in cloud platforms: Big data model for failure management in datacenters. In 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pages 105–116, 2016.

[5] Federico Gargiulo, Dirk Duellmann, Pasquale Arpaia, and Rosario Schiano Lo Moriello. Predicting hard disk failure by means of automatized labeling and machine learning approach. Applied Sciences, 11(18), 2021.

[6] Harshit Gupta, Dhairya Patel, Anurag Makade, Kapil Gupta, O. P Vyas, and Antonio Puliafito. Risk prediction in the life insurance industry using federated learning approach. In 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON), pages 948–953, 2022.

[7] Harshit Gupta, Tarun Kumar Rajput, Ranjana Vyas, O. P. Vyas, and Antonio Puliafito. Biometric iris identifier recognition with privacy preserving phenomenon: A federated learning approach. In Mohammad Tanveer, Sonali Agarwal, Seiichi Ozawa, Asif Ekbal, and Adam Jatowt,editors, Neural Information Processing, pages 493–504, Singapore,2023. Springer Nature Singapore.

[8] Sidi Lu, Bing Luo, Tirthak Patel, Yongtao Yao, Devesh Tiwari, and Weisong Shi. Making disk failure predictions smarter! In FAST, pages 151–167, 2020.

[9] Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. A performance evaluation of federated learning algorithms. In Proceedings of the second workshop on distributed infrastructures for deep learning, pages 1–8, 2018.

[10] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078, 2017.

[11] Vikas Tomer, Vedna Sharma, Sonali Gupta, and Devesh Pratap Singh. Hard disk drive failure prediction using smart attribute. Materials Today: Proceedings, 46:11258–11262, 2021.

[12] Kun Xia, Jianguang Huang, and Hanyu Wang. Lstm-cnn architecture for human activity recognition. IEEE Access, 8:56855–56866, 2020.