National College of Ireland

SSD and HDD Failure Detection using Advance Deep Learning Algorithms

MSc Research Project
Cloud Computing

Sandesh Muralidhar
Student ID: x20195737

School of Computing
National College of Ireland

Supervisor: Shreyas Setlur Arun

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sandesh Muralidhar
Student ID:	x20195737
Programme:	Cloud Computing
Year:	2022
Module:	MSc Research Project
Supervisor:	Shreyas Setlur Arun
Submission Due Date:	20/12/2022
Project Title:	SSD and HDD Failure Detection using Advance Deep Learning Algorithms
Word Count:	XXX
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sandesh Muralidhar
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

SSD and HDD Failure Detection using Advance Deep Learning Algorithms

Sandesh Muralidhar
x20195737

Abstract

Data centers are under demand to provide ever-more-efficient services due to the growing need for data processing and storage. But these services effectiveness may be effected by their dependence on hard drives—in particular, solid-state and magnetic drives, which are now among the most widely used types of data storage. Occasionally, these devices may fail and cause permanent data loss, which would violate contractual service level commitments and cause financial harm to both the customer and the hosting provider. This research aims to increase the cloud storage service quality by predicting SSD (Solid-State Drive) and HDD(Hard disk drives) failures using machine learning methods. Hard disks failure is a significant cause of application failures which leads to potential data loss and downtime. This research explores the application of deep learning methods and ensemble algorithms for detecting Hard drives failure prediction. Blackblaze dataset which has details of SMART parameters has been used for Research. This Research also explores on finding top paramters effecting the drive failure. It explores deep learning techniques, especially Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and a hybrid Conv-GRU model. The study takes into account both the spatial and temporal facets of the dataset to resolve the urgent demand for trustworthy forecasting algorithms as a solution. To assure effectiveness, the research employs a complete methodology that begins with thorough feature engineering and includes the extraction of the most important characteristics. A detailed correlation analysis of SMART parameters has also been performed in this research. Three deep learning models —Convolutional Neural Network (CNN), Gated Recurrent Unit(GRU), and a hybrid CNN-GRU—are applied to study it's efficiency in prediction of failures. The Conv-GRU model is the best-performing algorithm of the ones that were looked into; it performs better in terms of accuracy, precision, recall, and F1-score metrics. Its superior performance over CNN and GRU equivalents is largely due to its capacity to combine spatial and temporal information. The outcomes of this research would contribute to improve efficient storage management, hence the overall availability of datacenters and quality of cloud storage service.

1 Introduction

Increasing adoption of cloud storage systems and the expansion of large-scale data centers, cloud computing has experienced tremendous growth. Frequent storage system failures are becoming a significant problem to data dependability and service quality as

the cloud storage system expandsZhou et al. (2021). Hard disk drives (HDDs) and solid-state drives (SSDs) are two types of storage devices that are essential to maintaining service quality. Drives have one of the greatest failure rates among hardware components in cloud storage systems, and as a result, their Mean Time Between Failure (MTBF) is extremely high. It poses significant challenges for data centers, including the possibility of data loss, system failures, and higher management expensesZhou et al. (2023). Survey conducted indicates that about 28% of data center outages are attributable to storage component problems. The failure rate is accounted by the bulk of HDD and SSD failures Gu et al. (2023).A different study claims that in 2016, the reported downtime costs for 63 data centers increased from $5,617 per minute in 2010 to $8,851 per minute Zhang et al. (2023). Increased storage dependency and availability are achieved through the use of passive tolerance techniques, such as erasure codes and Redundant Arrays of Independent Disks (RAID). The drawbacks of this approach are long recovery time and challenges in maintaining dependability as cloud storage starts to scale Zhou et al. (2021). Implementation of Fault tolerance cannot guarantee System accessibility as the action is taken after the failure incident. A proactive tolerance mechanism is required to predict the failure and migrate the data before the failure of the disk. Proactive tolerance should help in increasing system reliability and availability Zhang et al. (2023).

The growth of extensive use of data, hard disk drive (HDD) manufacturers are continuously evolving their approach to self-monitor technology with their manufactured products. The enhancement of this approach has introduced an effort to predict failures in hard disks at an early stage to enable users to access the data backup facility. In recent years, with advancement to extensive data utilization, hard disks have played an essential role in data storage and serve as a primary technology to store backup data for a long time. In the recent decade, an emerging trend - "Solid State Drive" (SSD), which functionalizes as semiconductor storage, has surpassed in application over hard disk drives (HDDs) regarding response time as well as high-throughput performance. However, Djordjevic (2021) explained that HDDs had been identified as a cheaper storage medium per byte than SSDs. Therefore, it is considered an integral and predominant medium for data storage available at an industrial level and in the consumer market. According to the potential research evidence, hard disk drive (HDD) renders a more susceptible condition to failure in stored data protection compared to other storage components in the user's computer system.

The failure in the hard disk drive generally causes permanent loss of data, which, therefore, marks an expense more than the cost of an HDD. Thus, experts have introduced a research paradigm to explore different technologies that can detect the HDD failure and subsequently prevent data loss through an effective data retention strategy. The mechanism of HDD failure has been identified as "predictable" and "unpredictable" failure. As stated by Djordjevic (2021), the former failure option is mostly showcased with a progressive degradation in drive performance across the HDD operation lifecycle. This degradation is mainly caused by mechanical wear and tear of components as well as the degradation of the storage surface. This degradation in performance is monitored using different parameters that are typically applied to predict the occurrence of failure. On the other hand, unpredictable failures in hard disk drives occur instantaneously, although no previous indications are observed in HDD performance. It is mainly caused due to external forces, but the occurrence is unpredictable. Apart from this, other implic-

ations of unpredictable failures include hidden defects within HDD components typically observed at an "early stage" and also in the "wear-out period."

Predicting HDD component failure has become an area of research interest across multiple industries, such as aerospace, agriculture, technology, energy, and manufacturing. The suitable framework acknowledging the prediction issue typically varies based on the analytical or business goals of the sector and data availability. One of the standard approaches to predicting the problem is regression and estimation of the component's "Remaining Useful Life" (RUL) through the time series method. The suitability of the technique suggestively entailed through its assumptions while there is enough information according to the incremental time steps for the generation of an "RUL" estimate. Thus, the overall understanding of the degradation inducing predictable or unpredictable failure can be determined more smoothly than experiencing sudden failure.

Over the years, especially in the recent decade, the emerging trend of the "Internet of Things" (IoT), along with the availability of "machine sensor" data, has increased feasibility in RUL estimation while using real-world data. In modern enterprises, the prediction of HDD component failure is performed through a "standard self-monitoring system" - "Self-Monitoring, Analysis and Reporting Technology" (SMART). The credibility of this technology has been suitably enhanced through its extensive ability to record "real-time" sensor data that can serve the purpose to detect malfunctioning and also anticipated failure in hard disk drives. The above-specified monitoring system typically monitors different parameters across the HDD lifetime. These parameters store information regarding temperature, hours of operations, and the degree of "on/off" cycles. The values confined to these parameters are then compared to predefined "threshold values," primarily set by an HDD manufacturer. While exploring the principle mechanism of SMART readings, the contribution of the dataset in the prediction process promotes suitable actions.

The Backblaze dataset, which is a public hard disk drive dataset and consists of more than 100,000 active drives, including "hard disk drives" (HDD) and "solid-state drives" (SSD), have been vigorously used in the prediction process. The dataset comprised SMART readings from the above-specified hard drives developed by various brands and of different models. The contribution of this dataset in the prediction of hard disk failure has become a potential aspect to consider due to its availability and feasibility in gaining insights into the malfunctioning and failure of HDD components.

The identification of the failure in HDD components through SMART readings is a common prediction approach in data centers and enterprises since it uses threshold values and enables real-time data backup, preventing data loss upon replacing the failing one. It has been determined that manufacturers set the threshold value at a level significantly higher to prevent false alarm rates and minimize the return of HDD during the "warranty" period Djordjevic (2021). Contributing to this fact, experts have verified numerous attempts to reduce prediction failure, thus integrating threshold-based algorithms into action to increase the accuracy and feasibility of the detection process. While regarding this fact, some advanced approaches in the comprehensive failure detection of hard disks are achieved through machine learning and advanced deep learning algorithms. Thus, emphasis has been given to the critical exploration of the research paradigm, indicating a comprehensive approach to detecting hard drive failures such as HDD and SSD. Feature

Engineering has been applied in this study to find the top features affecting drive failures. This research deploys three different deep-learning models which are CNN (Convolutional Neural network), GRU (Gated Recurrent Unit), and Conv-GRU (Con-volutional Gated Recurrent Unit). To identify the most optimal model for disk failure prediction evaluation of each model is conducted by calculating the Accuracy, Precision, Recall, and F1 score on the test set, and the model with the highest metric scores has been deployed to the web application.

2 Related Work

The contribution of knowledge in this chapter is based on a comprehensive evaluation of information on Disk failure detection using various algorithms and further comparing the prediction outcome to develop insights into the accuracy level of the models. The prediction of failure is further interpreted using the Backblaze dataset and subsequently explores the pre-processing criteria and feasibility of the dataset in the detection process.

2.1 Backblaze Dataset - Application and Implications

With a comprehensive understanding of the importance of hard drives as an effective data storage component, research has typically emphasized the data loss and pertained risk of HDD failure. As explained by Aussel et al. (2017) the prediction of this failure has been enhanced with the application of the SMART method, where operational data is primarily used. In this regard, the Backblaze dataset consisting of large number of hard drives has induced a potential operational implication that exhibits hard drive (HD) heterogeneity with nearly 81 models. Manufacturers, while developing the hard drive, potentially used this dataset in recent times to assess and control the failure upon using an "unbalanced data ratio" of 5000:1 between the healthy and the defective samples within a real-world environment which is sparsely controlled. The contribution of this dataset, as explained by Tomer et al. (2021) to the prediction process typically optimizes the data evaluation with higher accuracy, especially when using cutting-edge technologies in research in recent times. Upon focusing on the importance of this dataset, an actual specification has been enhanced with its origin. Backblaze has been identified as an online "backup & cloud storage provider" that provides necessary information on hard drives using SMART attributes across a period of 2013-2019 (Tomer et al. 2021). Tomer et al. (2021) explained these hard disk drives (HDDs) are integrated from different HDD vendors such as Hitachi, Seagate, Western Digital, and others. Thus, it can be considered that this heterogeneous dataset serves real-time application in predicting hard drive (HDD and SSD) failures with the application of SMART readings.

2.2 Hard Drive failure detection using Machine Learning Algorithms

The prediction of hard disk failure is a vital approach that requires forecasting of data to understand whether there is a significant failure in the "material system." According to the explanation provided by Leukel et al. (2021) task accomplishment attains significant strategies that retain effective industrial maintenance, for example, predictive

maintenance. Therefore, researchers typically solve this prediction task using different algorithms, especially machine learning algorithms, since they allow an adequate prediction of hard disk drive failure (HDD). An effective prediction process discreetly indulges the necessary data extraction process that has increased exponentially because of the application of various sensing technologies. As explained by Carvalho et al. (2019) with a progression toward Industry 4.0, there has been a strong embrace of the digital system in industrial applications to enable a faster and more reliable information exchange. In this regard, extensive use of data in contemporary industrial systems is essential to keep a significant pace in the long-term success. While understanding the necessity of vast data in industrial applications and consumer markets, adequate storage space and components are essential to increase efficiency and reliability. The application of hard drives such as hard disk drives (HDD) and solid-state drives (SSD) has presented comprehensive options with enhanced build-up. But, due to unprecedented incidences related to storage degradation, failure in hard disks is a common consequence.

In a study conducted by Shen et al. (2018), the author specified the necessary approach to detect this failure at an early stage to prevent service interruption in the user's computer system. In this regard, the primary consideration for machine learning models serves as better prediction accuracy algorithms using different prediction classifiers. As the author further stated, SMART attributes have been used by manufacturers and research experts to analyze the condition of hard drives in the interior region and also HDD data of the exterior region through sensors countries. Although the technology has served potential applications for a long time, the detection consistency rate is estimated to be only 3-10%. Shen et al. (2018) implied that different methods are proposed and applied to enhance prediction accuracy; however, specific disadvantages are aligned with the approaches. The above study has introduced an ensemble algorithm - random forest (RF) for classification. The model is a "multitude" or constructed by combining multiple decision trees and utilizes randomized training samples as well as features to obtain accurate prediction results. The model has used the Backblaze dataset containing SMART records of nearly 64,193 drives and provided an experimental result showing improved accuracy compared to existing "state-of-the-art" methods.

In another study conducted by Li et al. (2017), the author applied two classification and prediction models -"decision trees" (DT) and "gradient-boosted regression trees" (GBRT) and compared the prediction accuracy of both models. While using two distinctive models, a real-world hard drive dataset containing approximately 121,698 drives has been used, thus, the study aimed at providing a suitable prediction outcome with real-world testified datasets. As per the experimental outcome, decision trees have provided a prediction accuracy of 93% with a "false alarm" rate of less than 0.01%, while the gradient-boosted regression trees have provided a prediction accuracy result of 90% without any indication of false alarms. Thus, the evaluation explained the potentiality of both models in the hard drive failure prediction process, although GBRT proves to be more reliable with no false alarm rates. Understandably, hard drives such as HDD are an essential storage component in the user's computer system. It has been identified that the speed with which data is suggestively transferred, as well as programs loaded in the system, typically depends on the compatibility of the disk drive. As discussed, excessive reliance on IT infrastructures on this disk drive can reduce its effectiveness and induce challenges in predicting failure, thus impacting the data integrity and availability as well

as the business continuity.

Over the years, experts have recognized the credibility of datasets in enhancing the predictability level through advanced classifiers. Ganesh et al. (2023) explained that a potential challenge with the public dataset such as the Backblaze dataset is recognizing the "class imbalance", which is further addressed through an "oversampling" approach by using an "adaptive Synthetic Algorithm". Apart from this, the study has introduced three different feature selection classifiers - Logistic regression (LR), Decision Tree (DT), and Random Forest (RF) to substantially approach the prediction of disk failure. As per the review of the experimental result, it has been observed that the random forest classifier has precisely predicted the disk failure with an accuracy of 92%, precision rate of 86%, recall rate of 90%, and F1-score of 88%, respectively. Thus, it can be stated that among all the machine learning models, random forest has proven to be more effective in the prediction of HDD failure.

2.3 Hard Drive failure detection using Deep Learning Algorithms

An extensive approach to cloud-based computing systems in recent decades, data centers are aimed at providing high service to users with almost negligent failure occurrence. Gao et al. (2019) explained that with an extensive count of large-scale data centers, cloud providers are facing immense challenges with hardware software failures. Among these hardware, challenges are rigorously perceivable in predicting the failure in hard disk drives, thus resulting in task failure as well as incompetent storage capacity. The system reliability with suitable prediction of the failure is specifically addressed as an underlying issue by hard disk manufacturers therefore, they have instigated the necessity for combining improved methods with SMART attributes for detection purposes. Gao et al. (2019) in their study introduced an advanced deep neural architectural model - a bidirectional long-short-term-memory (Bi-LSTM) model, which suggestively identifies the risk related to hard disk failure. The experimental result shows that the algorithm has outperformed previous machine learning and deep learning methods in the prediction process.

According to De Santo et al. (2022)), hard disk drive (HDD) component failure has been identified as a common problem of service downtime within data centers. The author specified that distinct approaches, such as "predictive maintenance techniques," are introduced to reduce the RUL of HDDs while minimizing service shortage data loss. While recognizing the need for successive prediction of HDD failure, De Santo et al. (2022)) have introduced a proposed deep learning model - LSTM, which combined with SMART attributes to estimate the condition of hard drives and predict the failure rate. The experiment has been performed using two real-world datasets containing 23,395 disks and 29,878 disks, thus establishing a predictive failure 45 days earlier than the meantime. Another study performed by Cahyadi and Forshaw (2021)) introduced a research interest in predicting hard disk failure by using public datasets, which are highly imbalanced. The study showcased the prediction leverage obtained using the LSTM model on the Backblaze Dataset when combined with SMART attributes and thereby achieved a correlation coefficient estimated to be 0.71. The result shows that this model is universally applicable and portable to enhanced operational datasets as well as disk types.

Apart from this understanding of the suitability of the LSTM model combined with SMART attributes when using the Backblaze dataset, studies have further preferred convolutional deep learning models, which combined with LSTM architecture to develop a "Convolution-LSTM" (C-LSTM) model for the accurate prediction process. Shi et al. (2022) introduced a similar model for predicting hard disk drives and SSD failures while achieving successive fault warnings. The experimental conduct performed in the above study has included different parameters and critically evaluated the outcome, showing that the proposed model has performed better than most other algorithms in predicting the failure of mechanical HDD storage components. Another study conducted by Wu et al. (2021) has introduced a similar convolutional LSTM model, although multi-channeled, has achieved a prediction accuracy of 99.8% with a reduced false-alarm rate of 0.2%. The overall information, therefore, explained that advanced deep learning models, particularly enhanced neural architectures, are more effective in predicting mechanical hard disk drive failure with the Backblaze dataset than existing models. However, it is recommended that the area needs further attention from experts to explore as a future research scope while using extensively imbalanced datasets for HDD failure detection.

2.4 Hard Drive failure detection using Hybrid Models

In the above sections, the specification of models in accurate prediction of HDD failures has been explored based on the understanding of the contribution of machine learning and deep learning classifiers. Apart from deploying these models, researchers have also focused on other algorithms and improved models. In the study conducted by Wang et al. (2022), the author has introduced a novel algorithm that combines "Generative Adversarial Network" (GAN) and "Long-Short-Term-Memory." The hybrid model has shown its capability to alleviate the issue of "data imbalance" while expanding the dataset of failed disks. The application of this trained model has shown improved accuracy in the failure detection process, Approximately 300 originally obtained failed disk data have been found to induce a significant impact on improving the fault detection of hard disks.

Apart from the above understanding of the model used, the preference for an improved algorithm has distinctly explored the research paradigm of novel HDD failure detection models. In Xu and Xu (2023) their study, tested the health status of hard disk drives from the selected dataset, which contains highly unbalanced data. The approach has been aimed at understanding the application of the proposed "convolutional transformer model" (ConvTrans-TPS) in the prediction of disk failures. The accuracy of the model has been determined using a large-scale dataset, typically the Backblaze dataset, thus specifying a comprehensive approach with an accuracy level of 96% and a correlation coefficient estimated to be 0.92. As per the understanding of the suitability of the model, it can be stated that the detection accuracy of HDD failures has shown an improvement compared to the CNN-LSTM model.

2.5 Research Gap

Existing literature in the domain of predictive maintenance in cloud computing often addresses various aspects of hardware failure prediction, a comprehensive integration of novel approaches remains scarce. The current body of work lacks in-depth exploration of the correlation metrics for smart metrics, which are crucial for understanding the dynamics of predictive maintenance. The extraction and analysis of the top features from smart parameters are not extensively covered in the existing research. The combination of deep learning models, specifically the integration of CNN, GRU, and their hybrid forms, presents an unexplored avenue in optimizing predictive maintenance models for cloud environments. Addressing these gaps would not only enhance our understanding of predictive maintenance but also contribute to the development of more robust and efficient strategies for hardware failure mitigation in cloud computing and Datacenter management.

2.6 Research question

How can machine learning methods leveraging SMART parameters can be effectively utilized to predict Drive failures and enhance the quality of cloud storage services ?.

2.7 Research objectives

Below are the research objective for the research
1) Investigate and analyze the correlation metrics for smart metrics in the context of predictive maintenance for hardware in cloud computing environments.
2) Identify and extract the top 15 features from smart parameters to enhance the understanding of their significance in predicting hardware failures in cloud systems.
3) Evaluate the individual and combined performance of deep learning models, including CNN, GRU, and hybrid architectures, for optimizing predictive maintenance in cloud computing environments.

2.8 Document Structure

There are multiple sections in this research report. Related work of different machine learning approaches on failure detection and their findings are provided in Section 2. The methodology followed in this research along with the Dataset description and Data Processing is explained in Section 3, Descriptive Analysis of the deep learning model that has been applied is presented in Section 4, The Implementation of the described models is explained in Section 5, Results obtained from the model is presented in Section 6.

3 Methodology

The reliability of data centers is significantly impacted by the performance and health of storage devices, encompassing both hard disk drives (HDDs) and solid-state drives (SSDs). Precise anticipation of potential failures in storage devices enhances the overall dependability of data centers by facilitating preemptive measures to mitigate data loss

risks. This research flows into an in-depth investigation of HDD and SSD failures, conducting a thorough analysis of disk logs sourced from real-world data centers spanning a comprehensive a-year duration. The analytical focus encompasses the examination of Self-Monitoring, Analysis, and Reporting Technology (SMART) traces extracted from HDDs specifically obtained from the real-world data center. To implement an effective and efficient method that can automatically predict the failure of such drives, a rigorous methodology is followed which is shown in Fig1. Various steps are carefully considered to make a real-world solution starting with the collection of data from a valid source, followed by in-depth data preprocessing and exploratory data analysis which can derive a better path for feature extraction and feature engineering which are subsequent steps followed in this methodology. Since different deep learning models are deployed in this research, a critical evaluation step is accounted for by employing different evaluation metrics to select the best-performing algorithm. Each step performed in this methodology is detailed in further subsections.

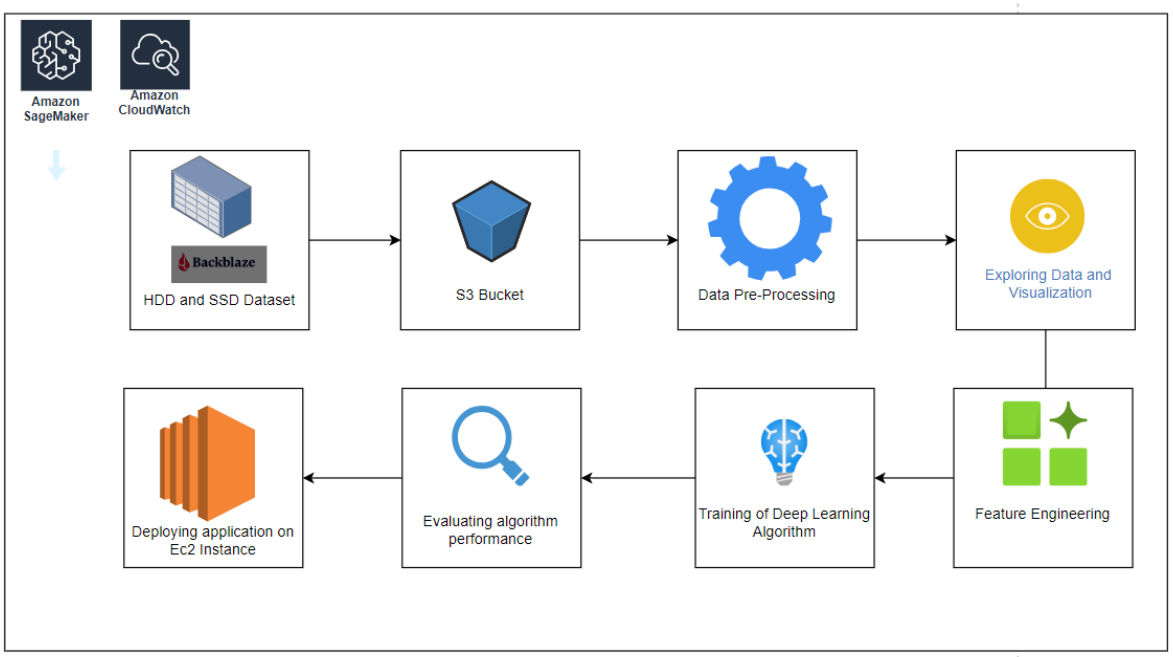

Figure 1: Methodology

3.1 Dataset Description:

In the pursuit of developing a robust predictive model for SSD/HDD failure, the initial phase involved comprehensive data collection from the Backblaze site. The dataset, extracted from the year 2022, contained detailed information about various hard drives, each characterized by distinct attributes. These attributes include essential parameters such as model specifications, capacity in bytes, and SMART (Self-Monitoring, Analysis, and Reporting Technology) metrics, which are indicative of the drive's health and performance. The dataset exhibits a binary classification task, with the target variable 'failure' indicating the occurrence of failure events. Notably, the dataset offers a substantial volume of non-failure instances, with a class distribution of 206,951 instances classified as non-failures and only three instances marked as failures. Blackblze in their website has

declared that their dataset can be used for Research purpose. The wealth of information embedded in the dataset serves as a foundation for training a predictive model capable of discriminating patterns indicative of potential failures in storage devices. The diverse set of features, ranging from raw SMART metrics to normalized values, facilitates a refined analysis. The dataset, thus, forms a pivotal component in the overarching objective of constructing an effective and reliable SSD/HDD failure prediction model.

3.2 Data Preprocessing:

A solid basis of data preprocessing is essential to building reliable and accurate prediction models. The first stage in this preparation method was dealing with the dataset's intrinsic class imbalance problem. The frequency of non-failure cases much exceeded the frequency of failures, which might introduce bias and impair the generalized ability of the model. A well-planned strategy was used to reduce this disparity. The dataset's quarterly interval organization and temporal nature offered a special chance to use temporal data to address the class imbalance. A more representative dataset was created by methodically removing failure occurrences from all files that were accessible and balancing them with a sample of non-failure examples. The goal of this methodological enhancement was to increase the model's exposure to failure cases thereby enhancing its skill in detecting subtle indications pointing towards possible failures in storage devices. Subsequent data preparation methods ensured the dataset's relevance and purity, going beyond correcting the class imbalance. The 'date' column was arranged chronologically to help with temporal comprehension of the data (an important feature in failure prediction because patterns may show temporal relationships). Duplicate entries were also removed. Since null values are frequently present in real-world datasets, they were handled with care. Excessively high null rate columns were removed to ensure data integrity and strike a balance between the volume of data and informativeness.

The preference for normalized SMART characteristics over raw data highlighted the significance of feature engineering in achieving the best possible model performance. The dataset was simplified by placing a strong emphasis on normalized values, giving priority to characteristics that significantly aid in failure prediction and eliminating unnecessary features. Given that the dataset includes certain columns with null values, inputting '0' to these columns eliminates the null values from the data. To ensure that the remaining features conveyed a variety of useful information, potential sources of noise were addressed by excluding columns that had a single unique value. These procedures in data preparation provide a dataset that is more balanced, informative, and temporally aware, making it a powerful tool for later model training and assessment in the difficult task of SSD/HDD failure prediction.

3.3 Exploratory Data Analysis:

The exploratory data analysis (EDA) deals into various facets of the SSD/HDD failure prediction dataset, offering valuable insights into its characteristics and potential patterns. Initial explorations centered around understanding the distribution of different HDD models visualized through a scatter plot showcasing the count of each model

The distribution of failure and non-failure instances across different HDD models was examined through a grouped bar plot as shown in Fig2. This visualization explains the

varying counts of failures and non-failures for each model, contributing to a comprehensive understanding of model-specific performance in failure prediction.

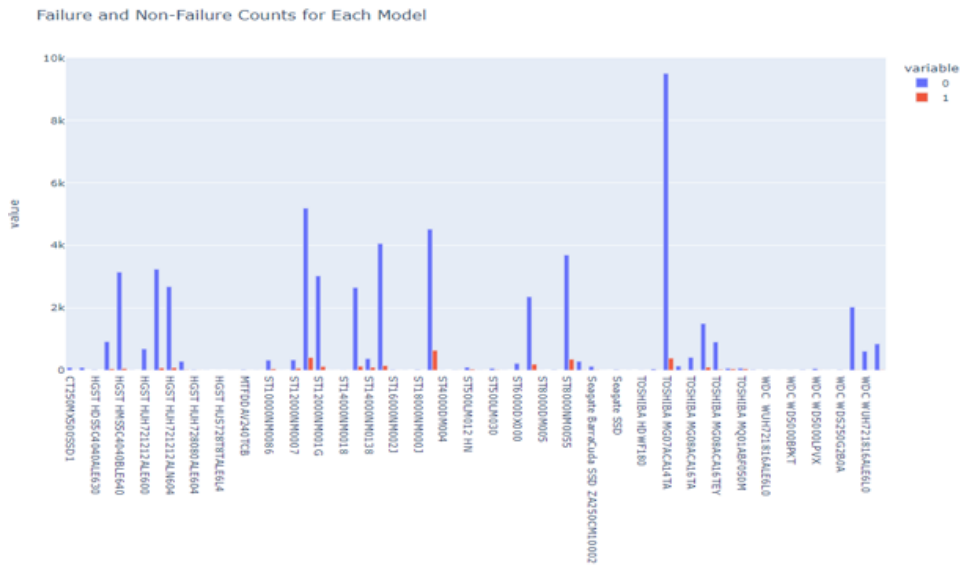

Figure 2: Failure Vs Non-Failure Counts of Each Model

An analysis is exhibited in Fig3 represents the counts of HDD failure in each month. From the Figure below it is evident that the maximum number of failures accounted for in August followed by October and September.

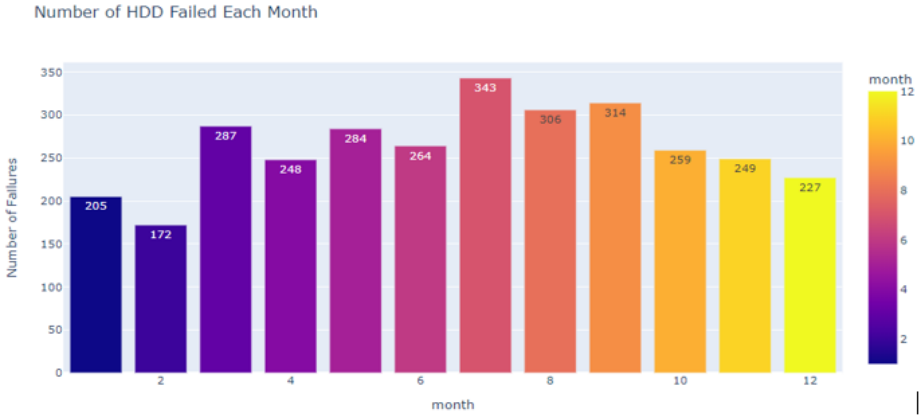

Figure 3: Counts of HDD Failure in a Month

Co-relation metrics was plotted to relate the relationship between the smart metrics, As represented in Fig4. Each cell in the diagram represents co-relation co-efficient. Some of the main observations is that smart 196 which is reallocation event count is strongly associated with smart-187(Reported uncorrected error and smart-188(command time out). Similarly, Smart-184(End-to-End error) is strongly associated with smart-9(Power-On Hours) parameter.

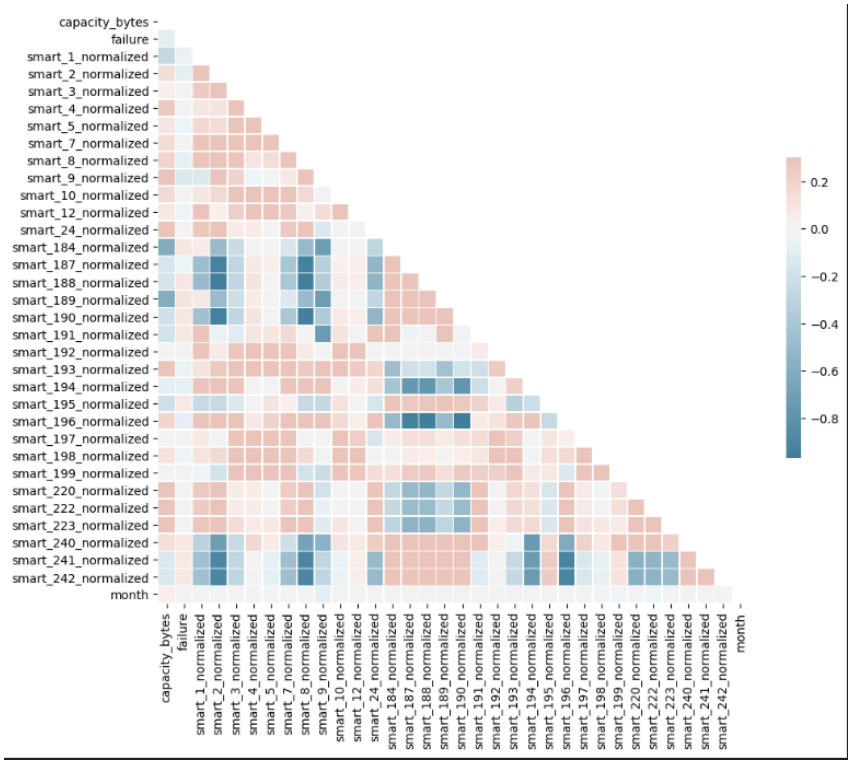

Figure 4: Corelation-metrics of SMART parameter

3.4 Feature Engineering

:

Feature engineering was a crucial step in refining the dataset for optimal model performance. Initially, the target column, 'failure,' was extracted as it represents the outcome to be predicted. Following this, categorical columns such as 'model' and 'serialnumber' were encoded using label encoding, converting them to a format suitable for modeling. To address the class imbalance as presented in Fig5, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, effectively balancing the representation of failure and non-failure instances in the dataset as depicted in Fig6.

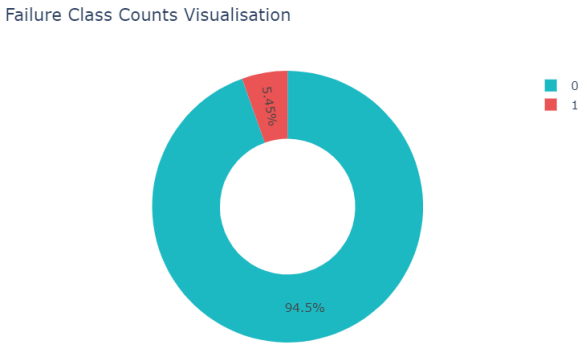

Figure 5: Imbalance Class Counts in Target Variable

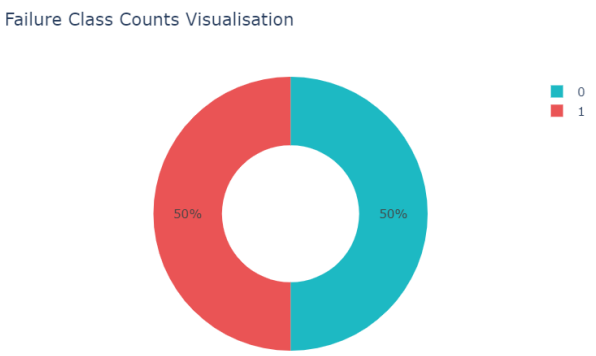

Figure 6: Balanced Classes After SMOTE Oversampling

Next, scaling down the features to a range of [0, 1] using MinMaxScaler was undertaken to ensure uniformity in feature magnitudes. The importance of this transformation lies in mitigating the curse of dimensionality associated with datasets containing numerous columns.

A feature importance analysis was conducted using the XGBoost classifier to enhance model interpret ability and efficiency. This involved training the model on the dataset and extracting feature importances as depicted in Fig7. The resulting feature importance scores were then used to identify the top 15 features contributing most significantly to the predictive capacity of the model.

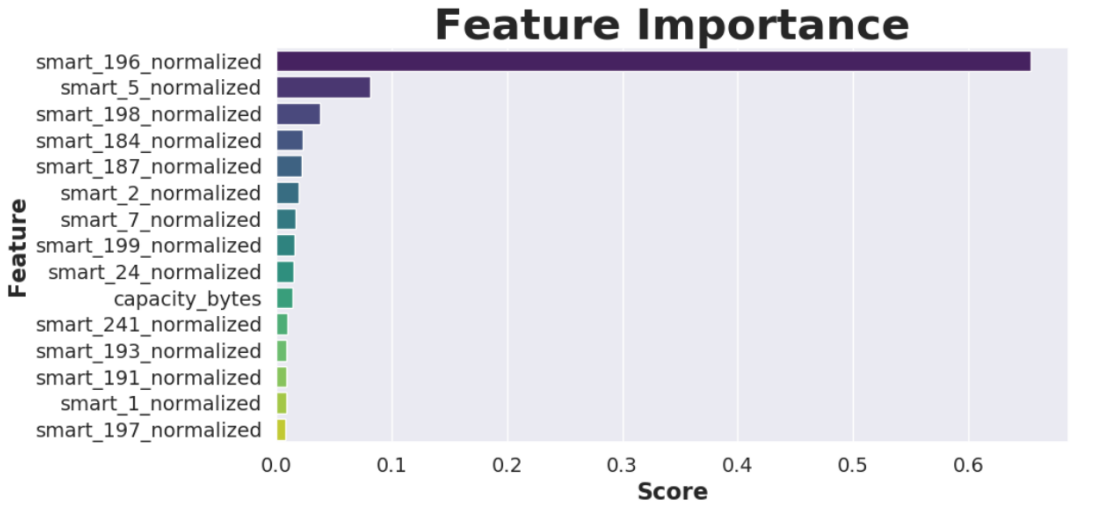

Figure 7: Feature Importance of Variables

The top features, including 'smart_196_normalized,' 'smart_5_normalized,' and 'smart_198_normalized,' were selected based on their importance scores. Subsequently, these features were retained in the final dataset, ensuring a efficient set of predictors for improved model efficiency. The entire feature engineering process, encompassing target extraction, label encoding, class balancing, feature scaling, and feature importance analysis, collectively aimed to optimize the dataset for subsequent machine learning model

training, ensuring a more effective and interpretable predictive model for HDD failure prediction.

3.5 Model Training

:

The model training process involved the utilization of a Convolutional Neural Network (CNN), a Gated Recurrent Unit (GRU), and a combination of both, known as Convolutional-GRU (Conv-GRU). Following the data split into training and test sets, the data was reshaped to meet the input requirements of the models. For this purpose, the training and test data were reshaped into a three-dimensional format, to accommodate the input shape expected by the models. Subsequently, a TensorFlow session was cleared to ensure a clean slate for model training. For each model, a customized architecture was defined. In this work, each model underwent compilation using appropriate loss functions (binary_crossentorpy), optimizers, and evaluation metrics. The training process involved fitting the models to the training data with 15 epochs, allowing the models to learn patterns and relationships within the data. The performance of each model was evaluated on the test set to measure its generalization capabilities.

3.6 Model Evaluation

: The evaluation of each model was conducted using the reserved test data, highlighting the significance of assessing model performance on unseen samples to gauge its generalization capabilities. The chosen evaluation metrics, including Accuracy, Precision, Recall, and F1-Score, collectively provided a comprehensive assessment of the models' predictive ability. Accuracy served as a fundamental measure of overall correctness, capturing the proportion of correctly predicted instances. Precision and Recall root deeper into the model's ability to minimize false positives and false negatives, respectively. The F1-Score, being a harmonic mean of Precision and Recall, offered a balanced metric that considered both aspects of classification performance. Employing multiple evaluation metrics was justified as it allowed for an understanding of the model's strengths and weaknesses, offering insights into their capacity to correctly identify positive and negative instances. The Integrated evaluation strategy ensured a robust and insightful assessment of the predictive models on the task of HDD failure prediction.

4 Design Specification

In this exploration of deep learning, three distinct models—Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and Convolutional-Gated Recurrent Unit (Conv-GRU)—a combination of two models to exhibit properties of both algorithms. A clear technical blueprint of these models is specified in the below subsections.

4.1 Convolutional Neural network (CNN)

CNNs are a class of deep neural networks specifically designed to handle data with a grid-like structure, such as images and time-series data. CNNs are primarily used for image processing but can be adapted for time-series data, which is relevant in monitoring HDD and SSD performance and can be used for predicting Hard disk and SSD Failures.

CNN can extract spatial features from the raw input data, such as patterns in disk usage or temperature fluctuations. This feature extraction is critical for identifying potential precursors to hardware failure. The key Component of CNN includes the Convolution layers, pooling layers, and fully connected layers. Where each layer plays an important role in performing accurate prediction. CNN is considered as most commonly used, yet effective architecture in the deep learning paradigm. The architecture of CNN is shown in Fig8. Reference for FigurePhung and Rhee (2019)

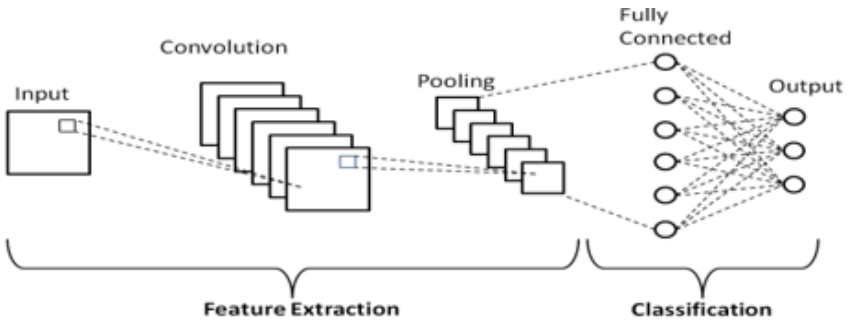

Figure 8: Architecture of Convolutional Neural Network

4.2 Gated recurrent units (GRUs)

Gated recurrent units (GRUs) are a powerful type of neural network designed to handle sequential data, like text, speech, and sensor readings. Unlike traditional RNNs that struggle with long-term dependencies, GRUs excel at remembering relevant information over extended periods while efficiently discarding irrelevant details, which is crucial for predicting future events like failures. As per analysis, GRU can offer significant value in predective maintenance and reducing failure rate. GRU uses input layers, one or more GRU layers and dense layers for output. The gated Recurrent Units architecture is shown in Fig9.Reference for Figure Wu et al. (2020)

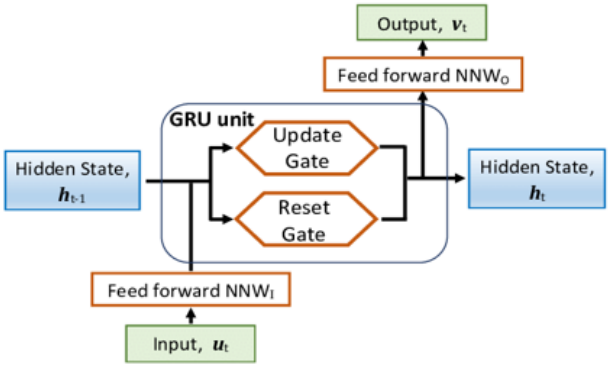

Figure 9: Architecture of Gated recurrent units (GRUs)

4.3 Conv-GRU (Convolutional Gated Recurrent Unit)

ConvGRUs combine the strengths of CNNs and GRUs, making them ideal for tasks like HDD/SSD failure prediction. It leverages CNNs' ability to extract features from raw

data and GRUs' ability for understanding temporal relationships. Convolutional GRU can simultaneously analyze spatial and temporal features of the disk drive data, such as detecting anomalies in data write/read patterns while considering the temporal context of these events which helps GRU in understanding the disk's health, leading to more accurate predictions of potential failures. As compared to CNN and GRU, ConvGRU is more complex to train and optimize and requires more data for accurate prediction. The architecture of Conv GRU is shown in Fig10. Reference for figure Wang et al. (2019)

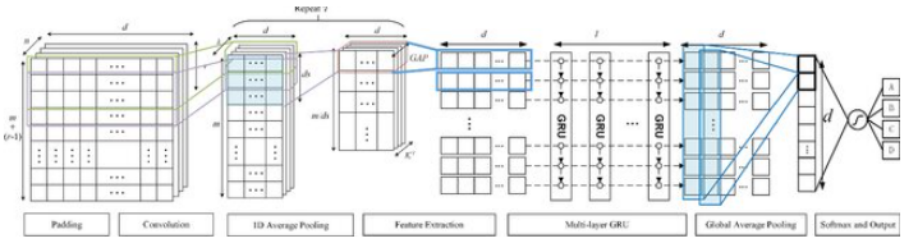

Figure 10: Architecture of Convolutional Gated Recurrent Unit

5 Implementation

The implementation of this Research commenced with data preprocessing phase aimed at addressing challenges inherent in the raw dataset. Recognizing the issue of imbalanced data, a systematic approach was adopted, involving the extraction and concatenation of failure instances from multiple quarters to form a more balanced dataset. The subsequent steps involved random sampling of non-failure instances, concatenating these datasets, and undertaking necessary data cleaning operations. The preprocessing steps were conducted using essential libraries such as Pandas for data manipulation and AWS Boto3 for interacting with the S3 storage. Further steps included handling missing values, dropping columns with excessive null rates, and eliminating features with raw data, all of which were crucial for refining the dataset.

Following data preprocessing, a comprehensive Exploratory Data Analysis (EDA) was performed to gain insights into the dataset's characteristics. Various visualizations, including scatter plots, bar charts, and pie charts, were generated using the Plotly and Seaborn libraries. Feature engineering was a critical phase in enhancing the dataset for model training. Removing redundant columns, encoding categorical variables using LabelEncoder and employing the Synthetic Minority Over-sampling Technique (SMOTE) for addressing imbalanced labels were pivotal steps.

Scaling features to a range of [0, 1] using MinMaxScaler was undertaken to ensure uniformity in feature magnitudes. Feature importance analysis using XGBoost aided in identifying and selecting the top features crucial for model performance. The chosen features were then scaled down for improved model efficiency. Subsequently, the thesis delved into model training using a Convolutional Neural Network (CNN), a Gated Recurrent Unit (GRU), and a hybrid Conv-GRU architecture. The data was split into training

and testing sets. Each model was implemented from scratch using TensorFlow and Keras, incorporating relevant layers, activations, and dropout mechanisms to enhance generalization. The model evaluation involved assessing accuracy, precision, recall, and F1-Score on the test data, This highlighted the importance of evaluating model performance on unseen samples. The XGBoost library facilitated feature importance analysis, providing insights into the crucial contributors to the models' predictions. This multifaceted implementation strategy ensured a thorough exploration of the dataset, robust model training, and comprehensive evaluation, collectively contributing to the successful realization of the thesis objectives.

5.1 Cloud services

Amazon S3: Amazon S3 provides scalable object storage in the cloud, allowing efficient and secure storage of Blackblaze datasets for all four quarters of 2022, supporting seamless data access and retrieval for machine learning tasks.

Amazon SageMaker: Utilizing Amazon SageMaker, I deployed and trained convolutional neural networks (CNN), gated recurrent units (GRU), and an ensemble model combining both (Conv-GRU) for HDD failure prediction, benefiting from a managed environment for machine learning model development and deployment.

Amazon EC2: Amazon EC2 instances host a web application dedicated to HDD failure prediction based on smart parameters, providing a scalable and reliable environment for serving predictions to end-users, with the capability to deploy and scale applications as needed.

Amazon CloudWatch: Amazon CloudWatch, integrated with EC2, monitors and ensures the health and performance of the deployed web application, offering real-time insights and alerts for proactive management, optimization, and troubleshooting.

5.1.1 Web Application Development

The web application is designed to predict disk failures based on a set of input parameters. It uses a pre-trained machine learning model (CONV_GRU.h5) and is built using Flask, a Python web framework. The application features an interface that allows users to input specific data points, which are then used to predict the likelihood of disk failure. Technologies Used to develop the web application are as follows.

- **Flask:** For creating the web server and handling requests.
- **HTML/CSS:** For frontend development and styling.
- **TensorFlow/Keras:** For loading and utilizing the pre-trained Deep learning model.
- **Python:** As the primary programming language for backend development.

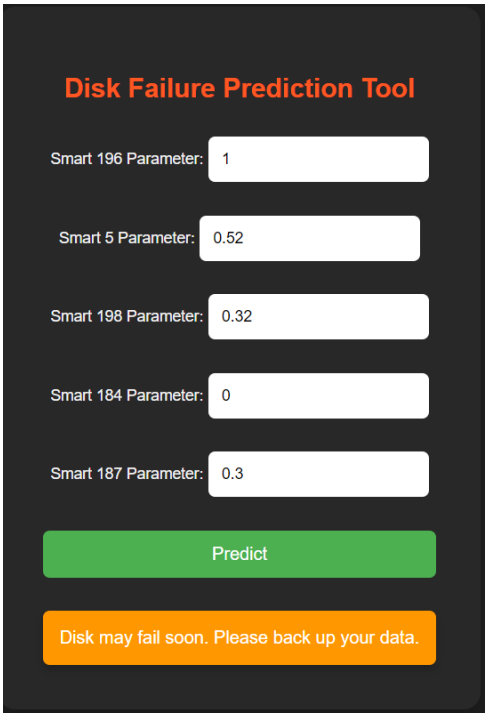

Figure 11: Web Application For Predicting Disk Failure based on SMART Parameters

Web URL For the application: http://18.206.229.126:5000/ (http://ec2-18-206-229-126.compute-1.amazonaws.com:5000)

6 Evaluation

This research task composes of three distinctive algorithms which are Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) and Convolutional-Gated Recurrent Unit (Conv-GRU). Analysis of obtained results and evaluation of executed models on various metrics on the test data is highly recommendable in opting out the most effective and best-performing model among other models. In this work, four different metrics accuracy, precision, recall, and F1-score are considered for the evaluation purpose therefore the results obtained by these models on all these metrics are discussed in the below subsections. The plotted graphs in this section reflects the model's performance on the validation set during the training phase.

6.1 Evaluation Based on Accuracy:

In evaluating the models based on the accuracy metric epoch-wise, the CNN model started with an initial accuracy of 0.7524 in the first epoch, gradually decreasing to 0.7059 by the fifteenth epoch. The GRU model exhibited a almost steady trend with little varriation, initiating at 0.7766 and achieving 0.7740 accuracy at the end of the training which is lower than it's initial phase value. On the other hand, the Conv-GRU model started at high note of 0.7640 and reaching an accuracy of 0.7879 in the final epoch. Notably, the Conv-GRU model outperformed both the CNN and GRU models in accuracy at the end of training phase. The CNN model exhibited fluctuating accuracy values, suggesting a varying degree of performance across different epochs. In contrast, the GRU and

CNN-GRU model demonstrated more stablity comparited to CNN model. The Conv-GRU model showed superior accuracy, making it the best-performing algorithm among the three. The accuracy values of ensemble model achieved by the Conv-GRU model underscores its effectiveness in accurately classifying the test data, showcasing its robust performance in comparison to the CNN and GRU models. The comparison of models based on Accuracy is displayed in Fig12.

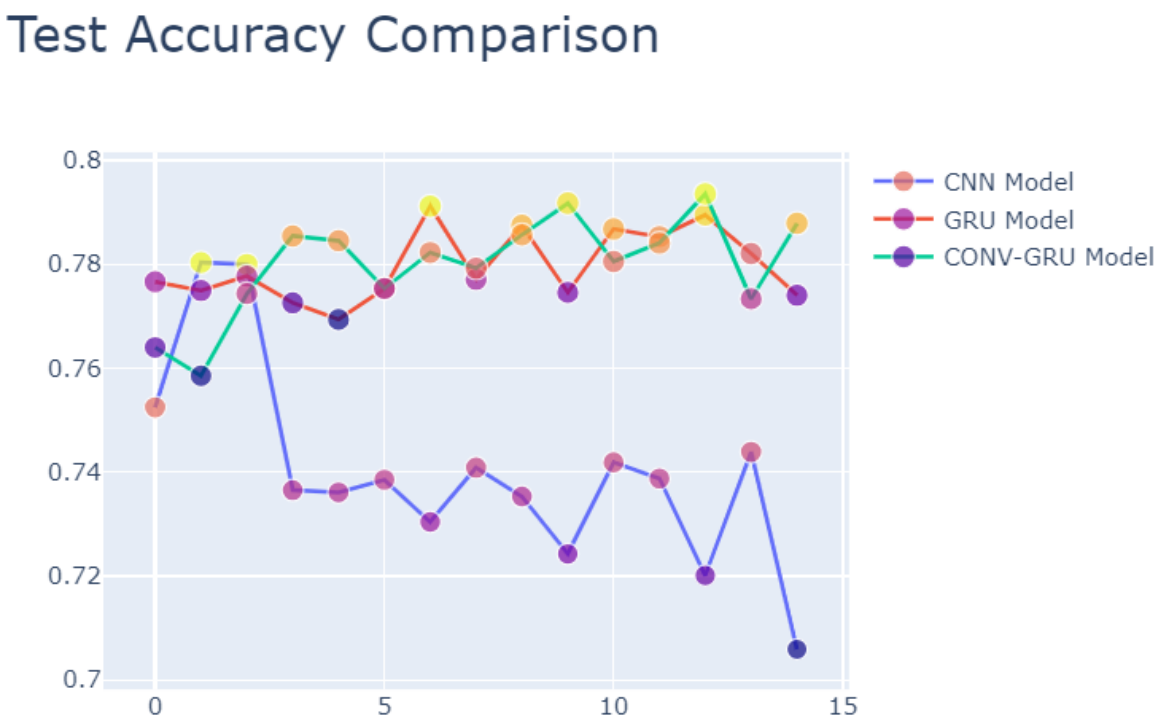

Figure 12: Test Accuracy Comparison of Models

6.2 Evaluation Based on Precision:

In the precision evaluation across epochs, the CNN model began with an initial precision of 0.8836 in the first epoch, experiencing a decline to a minimum of 0.6885 by the fifteenth epoch. This model exhibited decline in precision, indicating varying precision levels across different epochs. A similar trend was observed in the GRU model, starting at 0.9753 and ending at a precision of 0.9818 in the final epoch. In contrast, the Conv-GRU model consistently demonstrated precision enhancements, commencing impressively at 0.9607 and achieving a noteworthy precision of 0.9780 in the concluding epoch. Notably, while the Conv-GRU model consistently outperformed the CNN model in precision throughout the entire training period, the GRU model showcased a more consistent and stable ascent in precision over the epochs. The Conv-GRU model, while generally surpassing the CNN model, demonstrated a slightly lower precision compared to the GRU model, Analyzing the training validation value It can be observed that both GRU model and Ensemble model almost displayed same precision value during each Epoch. The comparison of models based on precision is displayed in Fig13.

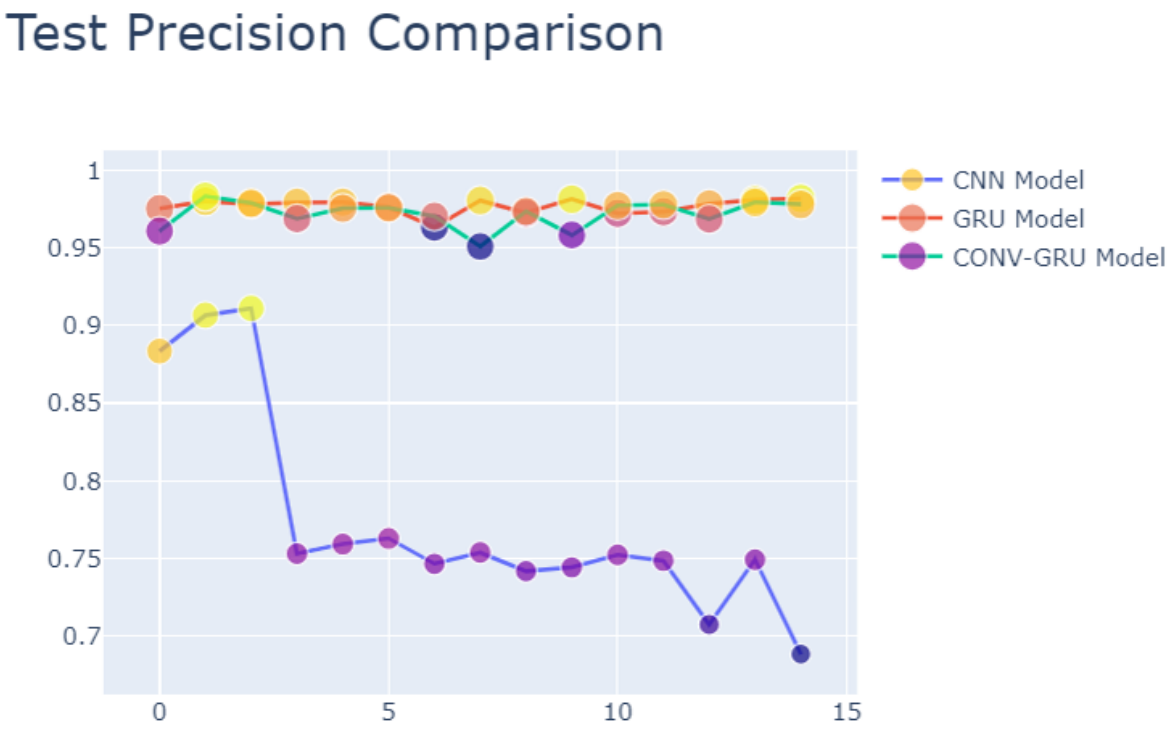

Figure 13: Test Precision Comparison of Models

6.3 Evaluation Based on Recall:

In the assessment of models based on the recall metric throughout the epochs, distinctive patterns emerge for each algorithm. The CNN model initiated with a recall of 0.5813, reaching a peak of 0.7517 in the fifteenth epoch. Meanwhile, the GRU model started with a recall of 0.5674, achieving a final recall of 0.5582. The Conv-GRU model started with value 0.5502 also showed different variations reaching the lowest point comparatively at epoch-1 but gradually gained and fluctuating again, reaching the final recall of 0.5888. Examining the recall values epoch-wise, the CNN model exhibited fluctuations, indicative of varying sensitivity to true positive instances across epochs and displayed better results than the other 2 models. Both GRU model and CNN-GRU model displayed different variations but the ensemble model at the end of Training phase ended having Recall value higher than GRU model. The comparison of models based on recall is displayed in Fig14.

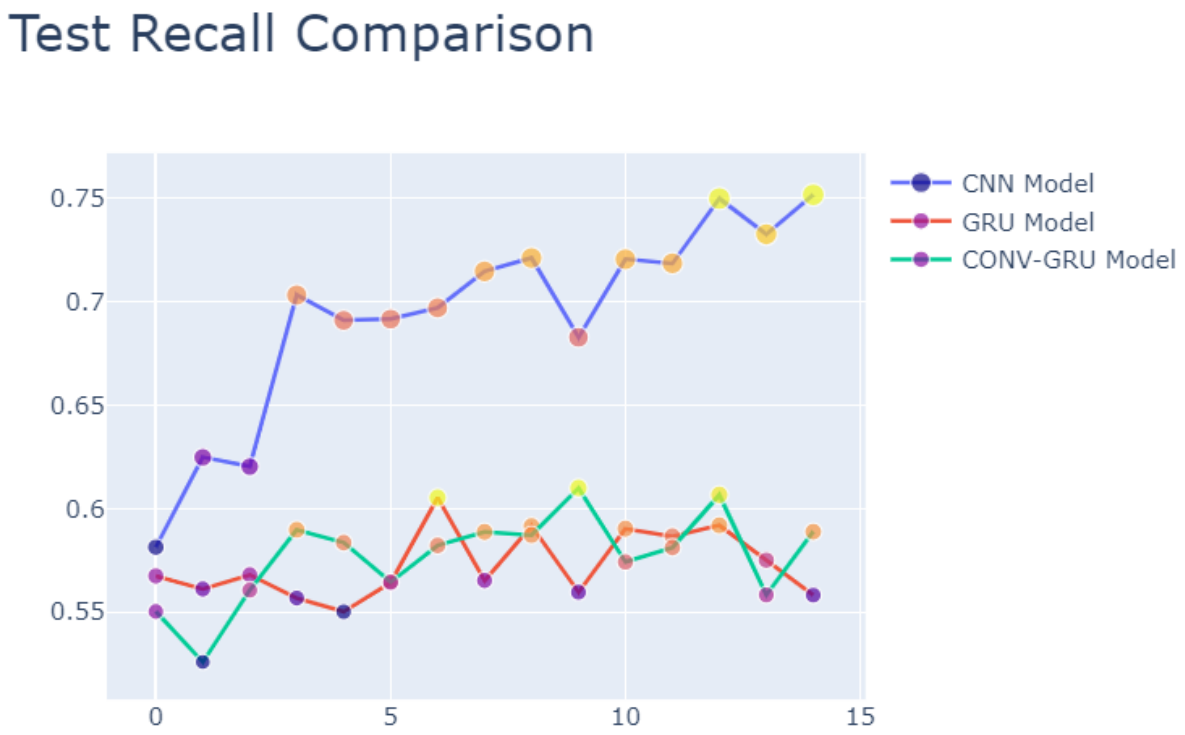

Figure 14: Test Recall Comparison of Models

6.4 Evaluation Based on F1-Score:

Evaluating the models based on the F1-score metric reveals fine insights into their performance throughout the epochs. The CNN model commenced with an F1-score of 0.6464, showcasing fluctuations in its ability to balance precision and recall. The GRU model demonstrated an increasing trend and achieved F1-score of 0.7269 Notably, the Conv-GRU model exhibited consistent improvement, with an F1-score of 0.7263. Analyzing the F1-score values epoch-wise, the CNN model displayed variability, indicating the delicate balance between precision and recall across training iterations. Overall it can be said that in terms of F1-Score CNN model outperforms as compared to GRU and CONV-GRU algorithms over the test data. The comparison of models based on F1-Score is displayed in Fig15.

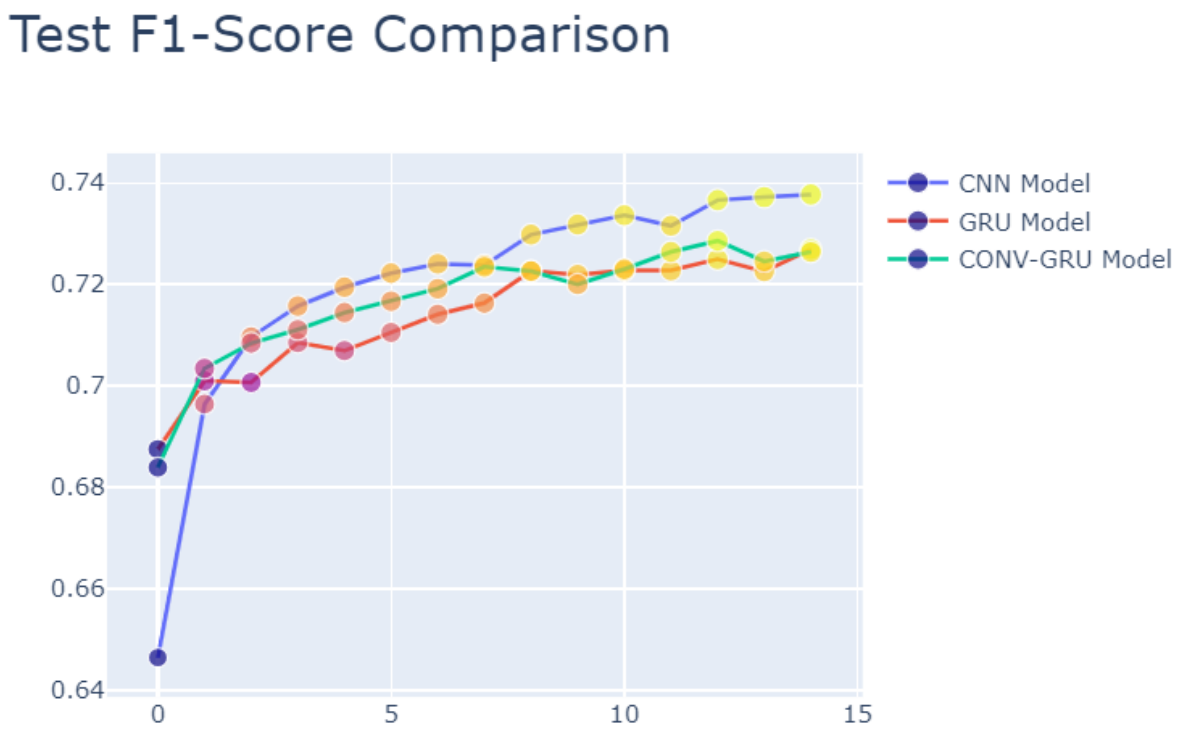

Figure 15: F1 score Comparison of Models

6.5 Evaluation Discussion

The below Table 1 represents the results obtained of the trained model on test data

Model	Accuracy	Precision	Recall	F1 score
CNN	70.59%	0.7076	0.7059	0.7053
GRU	77.40%	0.8366	0.7740	0.7630
CNN-GRU	78.79%	0.8419	0.7879	0.7791

Table 1: Performance metrics of trained model on Test Data

The Summary of the of three distinct algorithms – CNN, GRU, and Conv-GRU – applied to the task at hand. Among these algorithms, the Conv-GRU model emerged as the most robust performer across multiple evaluation metrics, including accuracy, precision, recall, and F1-score. The Conv-GRU model achieved an accuracy of 78.79%, precision of 0.8419, recall of 0.7879, and an F1-score of 0.7791. The superior performance of the Conv-GRU model can be attributed to its inherent architecture, which combines the strengths of both convolutional and gated recurrent units. This combination allows the model to capture intricate spatial features through convolutional layers while retaining the ability to comprehend temporal dependencies crucial in sequential data, as facilitated by the GRU component. The novel combination of these architectural elements enables the Conv-GRU model to effectively discern patterns and dependencies within the dataset, contributing to its predictive capacity. In contrast, while both the CNN and GRU

models demonstrated commendable performance, they exhibited limitations in achieving the same level of balanced accuracy, precision, recall, and F1-score as the Conv-GRU model. The CNN model showed competitive accuracy but faced challenges in maintaining a balanced trade-off between precision and recall. The GRU model exhibited consistent improvement but did not surpass the Conv-GRU model across all metrics. In conclusion, the Conv-GRU model stands out as the algorithm of choice for this particular task due to its ability to synthesize spatial and temporal features effectively. This discussion underscores the significance of model architecture in influencing performance outcomes and highlights the Conv-GRU model as a potent tool for predictive analytics in the context of the presented research.

7 Conclusion and Future Work

The implemented Research presents an exploration of diverse deep learning algorithms, namely CNN, GRU and Conv-GRU, in the context of the specified task. The noble approach of integrating convolutional and recurrent architectures has been a hallmark of this work, fostering a novel understanding of spatial and temporal intricacies within the dataset. In the pursuit of optimal predictive performance, the Conv-GRU model emerged as the standout algorithm, exhibiting superior accuracy, precision, recall, and F1-score metrics. The model's adeptness in synthesizing spatial and temporal features contributed significantly to its elevated performance, surpassing both CNN and GRU counterparts. The significance of this research lies in its contribution to advancing predictive analytics methodologies. By evaluating the strengths and limitations of each model, The research answers the research question on how deep learning algorithm fares in the study of drive failure prediction. The research also showcased co-relation metrics and extracted top 15 features impacting drive failures which fulfilled research objective. It also provides valuable insights for practitioners seeking effective solutions in similar domains. The Conv-GRU model, with its balanced proficiency in handling spatial and temporal aspects, stands as a robust tool for predictive modeling. The future scope of this research involves exploring more complex architectures and incorporating advanced techniques for feature engineering. Considering larger datasets and addressing domain-specific challenges could further enhance the models' generalizability. The comprehensive evaluation undertaken in this research lays the foundation for future endeavors in refining deep learning models for similar predictive tasks, promoting on-going advancements in the field and contributing to lower the downtime of on-premises datacenter and cloud storage environments.

References

Aussel, N., Jaulin, S., Gandon, G., Petetin, Y., Fazli, E. and Chabridon, S. (2017). Predictive models of hard drive failures based on operational data, *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 619–625.

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

Carvalho, T., Soares, F., Vita, R., Francisco, R., Basto, J. and G. Soares Alcalá, S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance, *Computers Industrial Engineering* **137**: 106024.

De Santo, A., Galli, A., Gravina, M., Moscato, V. and Sperlì, G. (2022). Deep learning for hdd health assessment: An application based on lstm, *IEEE Transactions on Computers* **71**(1): 69–80.

Djordjevic, B. (2021). Anomaly detection model for predicting hard disk drive failures, *Applied Artificial Intelligence* **35**: 1–18.

Ganesh, I. G., Sugan, A. S., Hariharan, S., Ramkumar, M. P., Mahalakshmi, M. and Selvan, G. S. R. E. (2023). Hdd failure detection using machine learning, *in* P. Singh, D. Singh, V. Tiwari and S. Misra (eds), *Machine Learning and Computational Intelligence Techniques for Data Engineering*, Springer Nature Singapore, Singapore, pp. 721–731.

Gao, J., Wang, H. and Shen, H. (2019). Task failure prediction in cloud data centers using deep learning, *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1111–1116.

Gu, J., Wang, Y. and Wang, G. (2023). Multi-instance adversarial learning domain adaptation network for failure prediction of unlabeled solid-state drives, *IEEE Transactions on Instrumentation and Measurement* **72**: 1–11.

Leukel, J., Gonzalez, J. and Riekert, M. (2021). Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review, *Journal of Manufacturing Systems* **61**: 87–96.

Li, J., Stones, R., Wang, G., Liu, X., Li, Z. and Xu, M. (2017). Hard drive failure prediction using decision trees, *Reliability Engineering System Safety* **164**.

Phung and Rhee (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets, *Applied Sciences* **9**: 4500.

Shen, J., Wan, J., Lim, S.-J. and Yu, L. (2018). Random-forest-based failure prediction for hard disk drives, *International Journal of Distributed Sensor Networks* **14**: 155014771880648.

Shi, J., Du, J., Ren, Y., Li, B., Zou, J. and Zhang, A. (2022). Convolution-lstm-based mechanical hard disk failure prediction by sensoring s.m.a.r.t. indicators.

Tomer, V., Sharma, V., Gupta, S. and Singh, D. P. (2021). Hard disk drive failure prediction using smart attribute, *Materials Today: Proceedings* **46**: 11258–11262. International Conference on Technological Advancements in Materials Science and Manufacturing.
URL: *https://www.sciencedirect.com/science/article/pii/S2214785321022586*

Wang, X., Wu, P., Liu, G., Huang, Q., Hu, X. and Xu, H. (2019). Learning performance prediction via convolutional gru and explainable neural networks in e-learning environments, *Computing* **101**.

Wang, Y., Dong, X., Wang, L., Chen, W. and Zhang, X. (2022). Optimizing small-sample disk fault detection based on lstm-gan model, *ACM Trans. Archit. Code Optim.* **19**(1). **URL:** *https://doi.org/10.1145/3500917*

Wu, J., Yu, H., Yang, Z. and Yin, R. (2021). Disk failure prediction with multiple channel convolutional neural network, *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Wu, L., Nguyen, V. D., Nanda Gopala, K. and Noels, L. (2020). A recurrent neural network-accelerated multi-scale model for elasto-plastic heterogeneous materials subjected to random cyclic and non-proportional loading paths, *Computer Methods in Applied Mechanics and Engineering* **369**: 113234.

Xu, S. and Xu, X. (2023). Convtrans-tps: A convolutional transformer model for disk failure prediction in large-scale network storage systems, *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1318–1323.

Zhang, X., Tan, Z., Feng, D., He, Q., Ju, W., Hao, J., Zhang, J., Yang, L. and Qi, W. (2023). Multidimensional features helping predict failures in production ssd-based consumer storage systems, *2023 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1–6.

Zhou, H., Niu, Z., Wang, G., Liu, X., Liu, D., Kang, B., Hu, Z. and Zhang, Y. (2023). Proactive drive failure prediction for cloud storage system through semi-supervised learning, *IEEE Transactions on Dependable and Secure Computing* pp. 1–16.

Zhou, H., Niu, Z., Wang, G., Liu, X., Liu, D., Kang, B., Zheng, H. and Zhang, Y. (2021). A proactive failure tolerant mechanism for ssds storage systems based on unsupervised learning, *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pp. 1–10.