iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49802

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

A Survey Paper on Hard Disk Failure Prediction Using Machine Learning

Abdulmuiz Shaikh[1], Harsh Dubey[2], Shreyash Bajhal[3], Anuj Bhadoriya[4]
Department of Computer Engineering, Sinhgad Institute of Technology and Sciences, Narhe, Pune, Mahararshtra, India

Abstract: Failure of Hard Disk is a term most companies and people, fear about. People get concerned regarding data loss. Therefore, predicting the failure of the HDD is an important and to ensure the storage security of the data center. There exist a system named, S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) in hard disk tools or bios tools which stands for Self-Monitoring, Analysis and Reporting Technology. Our project will be predicting the failure of hard drive whether it will fail or not. This prediction will be based on Machine Learning algorithm. S.M.A.R.T. values of hard disk will be extracted from external tool.

Keywords: S.M.A.R.T., Hard Disk Failure, LSTM, XG Boost

I. INTRODUCTION

HDD failure will not only cause the loss of data, but also may cause the entire storage and computing system to crash, resulting in immeasurable property loss to individuals or enterprises. Being able to detect in advance, an HDD failure may both prevent data losses from happening and reduce service downtime.

There exist a system named, S.M.A.R.T. in hard disk tools or bios tools which stands for Self-Monitoring, Analysis and Reporting Technology. This method returns unlabelled data overtime, and the healthy and faulty data are highly mixed. This returned data will be fed to ML algorithm to predict the hard drive failure.

II. LITERATURE SURVEY

In the paper [2], Use of decision trees, The fault prediction model can handle the failed hard disk in advance data backup and migration timely, so as to avoid failure and data loss, to protect the data security in the data center. But it also says decision trees are largely unstable compared to other decision predictors, also, they are less effective in predicting the outcome of a continuous variable.

In [3],the author proposed use of Deep Recurrent Neural Networks (DRNN). DRNN was chosen because of its remarkable performance in many applications including HDDs failure prediction. The limitation of [3] was the computation of this neural network is slow, training the model can be difficult task, also it faces issues like exploding or Gradient vanishing.

In the paper [4] author uses XGBoost, LSTM and ensemble learning algorithm to effectively predict disk faults. Also in this paper here,LSTM takes longer to train, requires more memory. XGBoost does not perform so well on sparse and unstructured data.

In , [5] the dataset statistically used to discover failure characteristics along the temporal, spatial, product line and component dimensions. And specifically focus on the correlations among different failures, including batch and repeating failures, as well as the human operators' response to the failures.

It focusses more on RAID architecture and have high project requirements.

The study [6] is based on empirical observation that reallocated sector count, a metric recorded by the disk drive, increases prior to failure. But it's limitation is in empirical observations, calculations can be very expensive, and also shows lack of reliability. [7] presented that this work proposes a method for fault detection in HDD based on Gaussian Mixture Model (GMM). One drawback of GMM is that there is lot of parameters to fit, and usually requires lots of data and many iterations to get good results.

In this paper[8],it propose a failure prediction method using a Bayesian Network. The method uses the deterioration over time of a HDD, calculated via SMART for predicting eventual failures. This study is less accessible to many investigators because the analyses are usually run using software that is unfamiliar or less familiar to many of them.

The [9],shows hard drive failure prediction models based on Classification and Regression Trees, which perform better in prediction performance as well as stability and interpretability compared with the Backpropagation artificial neural network model. But a small change in the dataset can make the tree structure unstable which can cause variance.

Paper [10] introduce a novel method based on Recurrent Neural Networks (RNN) to assess the health statuses of hard drives based on the gradually changing sequential SMART attributes. But limitation is training an RNN is a very difficult task. It cannot process very long sequences if using tanh or ReLU as an activation function.

Citation [11] says it used, Mahalanobis distance for aggregating all the monitored variables into one index, which was then transformed into Gaussian variables by Box-Cox transformation. A sliding window based generalized likelihood ratio test was proposed to track the anomaly progression. Significant progression indicates the HDD is approaching failure. But in [11] the Box-Cox transformation cannot cover for a poor model, and it might obscure the fact that the model is poor fit. The Box-Cox transformation can be hard to interpret.

III. LITERATURE REVIEW

TABLE I

S.no	Paper Title	Author	Year	Algorithm/ Methodologies used	Disadvantages/limitations
1.	"Disk Failure Early Warning Based on the Characteristics of Customized SMART"	Jian Zhao, Yongzhan He, Hongmei Liu, Jiajun Zhang, Bin Liu	2020	Use of decision trees, The fault prediction model can handle the failed hard disk in advance data backup and migration timely, so as to avoid failure and data loss, to protect the data security in the data center.	Decision trees are largely unstable compared to other decision predictors, also, they are less effective in predicting the outcome of a continuous variable
2.	"Predicting the Health Degree of Hard Disk Drives with Asymmetric and Ordinal Deep Neural Models"	Fernando D. S. Lima, Francisco Lucas F. Pereia, Lago C. Chaves	2020	Use of Deep Recurrent Neural Networks (DRNN). DRNN was chosen because of its remarkable performance in many applications including HDDs failure prediction.	The computation of this neural network is slow, training the model can be difficult task, also it faces issues like exploding or Gradient vanishing.
3.	"Prediction of HDD Failures by Ensemble Learning"	Qiang Li and Hui Li, Kai Zhang	2019	Based on hard disk's SMART data, this paper uses XGBoost, LSTM and ensemble learning algorithm to effectively predict disk faults	LSTM takes longer to train, requires more memory. XGBoost does not perform so well on sparse and unstructured data.
4.	"What Can We Learn from Four Years of Data Center Hardware Failures?".	Guosai Wang, Wei Xu, Lifei Zhang	2017	The dataset statistically to discover failure characteristics along the temporal, spatial, product line and component dimensions. And specifically focus on the correlations among different failures, including batch and repeating failures, as well as the human operators' response to the failures.	It focusses more on RAID architecture and have high project requirements

5.	"Predicting Disk Drive Failure Using Condition Based Monitoring"	Paul H. Franklin, Primus software	2017	This model is based on empirical observation that reallocated sector count, a metric recorded by the disk drive, increases prior to failure.	In empirical observations, calculations can be very expensive, and also shows lack of reliability.
6.	"A Fault Detection Method for Hard Disk Drives Based on Mixture of Gaussian and Non-Parametric Statistics"	Lucas P. Queiroz, Francisco Caio M. Rodrigues, Joao Paulo P. Gomes, Felip T. Brito	2016	Previous works on failure prediction based on parametric approaches, however that might always hold true. The following work proposes a method for fault detection in HDD based on Gaussian Mixture Model (GMM).	One drawback of GMM is that there is lot of parameters to fit, and usually requires lots of data and many iterations to get good results.
7.	"BaNHFaP: A Bayesian Network based Failure Prediction Approach for Hard Disk Drives".	Iago C. Chaves, Manoel Rui P. de Paula, Lucas G. M. Leite, Lucas P. Queiroz, Joao Paulo P. Gomes, Javam C. Machado	2016	In this paper,it propose a failure prediction method using a Bayesian Network. The method uses the deterioration over time of a HDD, calculated via SMART for predicting eventual failures.	Less accessible to many investigators because the analyses are usually run using software that is unfamiliar or less familiar to many of them
8.	"Hard Drive Failure Prediction Using Classification and Regression Trees".	Jing Li, Xinpu Ji, Yuhan Jia, Bingpeng Zhu, Gang Wang	2014	Paper proposes new hard drive failure prediction models based on Classification and Regression Trees, which perform better in prediction performance as well as stability and interpretability compared with the Backpropagation artificial neural network model.	A small change in the dataset can make the tree structure unstable which can cause variance.
9.	"Health Status Assessment and Failure Prediction for Hard Drives with Recurrent Neural Networks".	Chang Xu, Gang Wang, Xiaoguang Liu, Dongdong Guo, and Tie-Yan Liu	2014	Introduce a novel method based on Recurrent Neural Networks (RNN) to assess the health statuses of hard drives based on the gradually changing sequential SMART attributes	Training an RNN is a very difficult task. It cannot process very long sequences if using tanh or ReLU as an activation function
10.	"A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives"	Yu wang, Eden W. M. Ma, Tommy W. S. Chow and Kwog-Leung Tsui	2014	Mahalanobis distance was used for aggregating all the monitored variables into one index, which was then transformed into Gaussian variables by Box-Cox transformation. A sliding window based generalized	The Box-Cox transformation cannot cover for a poor model, and it might obscure the fact that the model is poor fit. The Box-Cox transformation can be

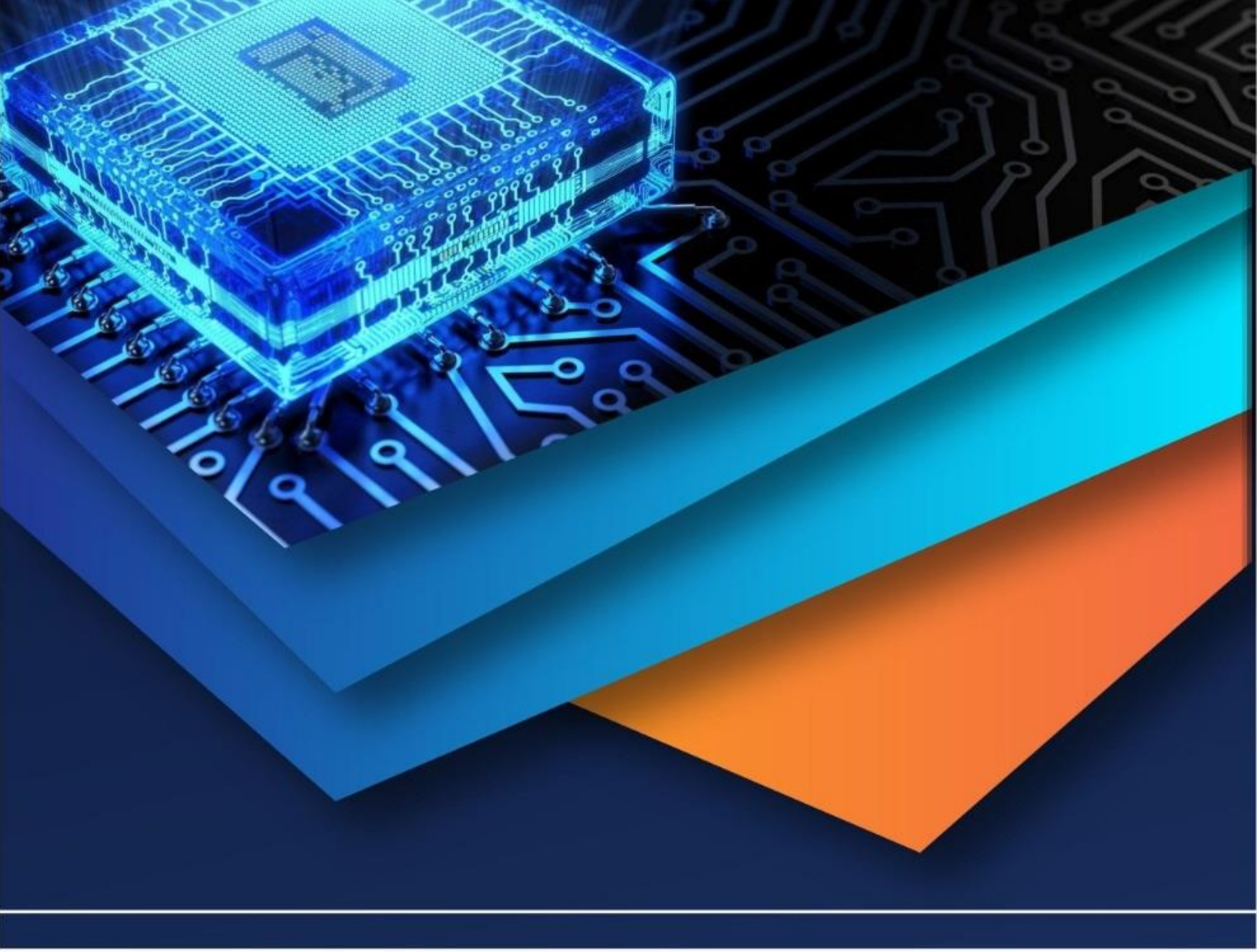
				likelihood ratio test was proposed to track the anomaly progression. Significant progression indicates the HDD is approaching failure.	hard to interpret

IV. CONCLUSIONS

The literature survey summarizes previous works, most of the work was based on neural network strategies. These were expensive methods with all their respective limitation. Some work was lacking accuracy, where some were using out dated software tools. Some work were using much time and memory resources. Hence this summarizes the literature survey.

REFERENCES

[1] Jian Zhao, Yongzhan He, Hongmei Liu, Jiajun Zhang, Bin Liu “Disk Failure Early Warning Based on the Characteristics of Customized SMART”, In 2020
[2] Fernando D. S. Lima, Francisco Lucas F. Pereia, Lago C. Chaves “Predicting the Health Degree of Hard Disk Drives with Asymmetric and Ordinal Deep Neural Models” , In 2020
[3] Qiang Li and Hui Li, Kai Zhang “Prediction of HDD Failures by Ensemble Learning” In 2019
[4] Guosai Wang, Wei Xu, Lifei Zhang “What Can We Learn from Four Years of Data Center Hardware Failures?”, In 2017
[5] Paul H. Franklin, Primus software “Predicting Disk Drive Failure Using Condition Based Monitoring” ,In 2017
[6] Lucas P. Queiroz, Francisco Caio M. Rodrigues, Joao Paulo P. Gomes, Felip T. Brito “A Fault Detection Method for Hard Disk Drives Based on Mixture of Gaussian and Non-Parametric Statistics” ,In 2016
[7] Iago C. Chaves, Manoel Rui P. de Paula, Lucas G. M. Leite, Lucas P. Queiroz, Joao Paulo P. Gomes, Javam C. Machado “BaNHFaP: Hard Disk Failure Prediction Using Machine learning A Bayesian Network based Failure Prediction Approach for Hard Disk Drives”, In 2016
[8] Jing Li, Xinpu Ji, Yuhan Jia, Bingpeng Zhu, Gang Wang “Hard Drive Failure Prediction Using Classification and Regression Trees” , In 2014
[9] Chang Xu, Gang Wang, Xiaoguang Liu, Dongdong Guo, and Tie-Yan Liu “Health Status Assessment and Failure Prediction for Hard Drives with Recurrent Neural Networks”, In 2014
[10] Yu wang, Eden W. M. Ma, Tommy W. S. Chow and Kwog-Leung Tsui “A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives” , In 2014



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY