Predictive Analysis and Screening Diagnosis of HDD Storage Deterioration Using Smart MI Strategies

Dr. V. S. Prakash
Assistant professor, Department of Computer science, Kristu Jayanti College, Bengaluru
vsprakash@kristujayanti.com

U. Udayakumar
Assistant Professor Department of Computer Science, RM Institute of Science and Technology, Ramapuram
phd.udayakumar1993@gmail.com

Dr. G. Rohini
Professor, Department of ECE,St. Joseph's Institute of Technology, Chennai
rohini.manoharan@gmail.com

Dr. Vishal Ratansing Patil
Assistant professor,
Computer Science and Engineering department,
Sandip University, Nashik.
vp0106@gmail.com

Sudhir Shenai
Associate Professor , Department of Information Science Engineering, Nitte Meenakshi Institute of Technology , Bangalore
shenai.sudhir@gmail.com

Dr. R. Thiagarajan
Professor, Dept of IT,Prathyusha Engineering College
rthiyagarajantpt@gmail.com

Article Info
Page Number:890 - 898
Publication Issue:
Vol 70 No. 2 (2021)

Abstract
Data centres are predominantly increasing in complexity as the hard discs get older. Due to the continuous monitoring in computers, it can reduce the efficiency of the control over security. In this proposed approach, artificial intelligence with machine learning is used to predict the potential features in the disk. Statistical tests over compatibility determine the detection of hard disc features. Due to those improper ways of handling the system, proactive failure occurs. To analyse those different signs of hard disc failure, different techniques are utilized. These warning signs are recognised and detected to identify the early detection of the failure in the hard disk. Potential data losses destroy the hard disc files using the malware or virus-infected files. These data crashes fail up until the boot process, where they get corrupted. Firmware is a type of corruption that can make data unreadable and damage its integrity. The system halts the process due to electronic failure and a power surge. The AI is used to predict the failure over the hard drive model for identifying the exact

precision. The HDD will overheat due to the high consumption of energy and overheating due to the maximum range. To avoid such disruption, failure is analysed using AI and machine learning. The comparative results are analysed and recognised to predict HDD failure using AI and MI methodologies. An external hard drive has to be checked and monitored based on the failure statistics report. Using the SVM, random forest, and nave Bayes classifier, we analyse the test parameters with accuracy and obtain approximate results.

Article History
***Article Received**: 05 September 2021*
***Revised**: 09 October 2021*
***Accepted**: 22 November 2021*
***Publication**: 26 December 2021*

Keywords: HDD, AI, Physical Failure, Firmware, Statistical Tests, External Hard Drive, ML

1. Introduction

Physical failure can cause physical damage where the stored data gets completely damaged and cannot be recovered properly. Different failures occur due to logical failures, where the data can be recovered using software. Unwanted noise needs to be analysed to identify the logical failure. Malfunctioning in those drives or corruption of the data can cause hard disc failure. System devices are interconnected to hard drives to store large amounts of data. Using intelligent artificial intelligence can determine the reliability of the drive and the failure rate. Using supervised learning, distinctive characteristics are analogized to deconstruct the accuracy of performance based on the results. The hard drive stores the operating system and application data in the software it contains. It checks and monitors varied hard drive properties. Intelligent AI technology validates the data on the hard disk. They form the dataset, and the attribution is extracted. Data gets pre-processed, from which progress is extracted. These pre-processing steps train and validate those exact results. Classified results give those resultant data along with training data. The use of predictive flags analyses the smart way of predicting failure over hard discs. Malware on the disc can impact complicated memory by lowering payload and accelerating traffic utilisation algorithms. Through physical degradation that inhibits stored information from being adequately cleaned and retrieved, hard drives might physically fail. Artificial intelligence can reduce data complexity by intelligently detecting hard disc failures. Inappropriate system management is the root cause of proactive hard disc failures. Internal and external failures in the computer system are real-time issues in the recognition of the computer system. Potential causes of losses occur mainly due to external hard discs. Such issues are characterised by the need to analyse the damages within those data. HDD self-monitoring is regarded as a technique for early fault diagnosis since it has the ability to examine failures. Significant positive patterns are one type of mistake that can occur as a result of read defects, monitoring data requests, and achieving error situations. Those enable users to monitor their data and back it up continuously. Using signals with predicted flags is an intelligent method of analysing full disc failures. Specific constraints are analysed using machine learning, which derivates the failure detection. Using the trained and tested model, the correlation yields the prediction of data.

Improper handling of the system on the hard disc causes proactive failure. An abnormal set of parameters analyses the behaviour of system loading performance and network traffic mechanisms with CPU utilization. This failure can be analysed to understand the recovery actions. Failures cannot be avoided due to system service failures. By using the runtime to train the data, the data are systematically related. Machine learning approaches are improvised to improve quality and interpretation over data labelling and predictive data classification. Use machine learning

techniques to analyse data labelling and prediction objectives. Statistical assumptions are used to predict which discs will fail along the way. Improve performance by checking the stability and reliability of your hard drive. Based on the training data and sample optimization, the data are parameterized using an optimization procedure. Benchmarking approaches are used to analyse the evaluation of various machine learning techniques. These attributes are extended by model specialisation through drive property differentiation. These trained samples are used to classify the data by classifying approaches when using data randomization. Improve performance by checking the stability and reliability of the hard drive. Attributes are validated for analysing the implied type of failure. The threshold range easily identifies the predefined value. Data reliability must be maintained to avoid data clutter. Traditional approaches do not provide adequate methods for analysing privacy and confidentiality.

Some hybrid systems experience data errors while communicating with each other. This study enhances efficiency using machine learning and AI methods. These different parameters are used to identify good drives and failure-type drives. Imbalanced data are inferred using the predictive analysis. Overhead data computes those data that enhance the efficiency of different failure systems. Memory is kept on a hard disc, which is a secondary storage device. Hard discs are used in data centres to store data for quick access and to boost the dependability of storage systems. Through the creation of cutting-edge models, machine learning offers a high level of reliability in disc failure prediction. Data may be stored on hard drives because they can hold both data and information. These spherical discs are magnetic and are part of computers. Certain devices only need a small amount of computing space to carry out the commands. This data is stored with random access and is non-volatile. The file system is then improved by error correction through redundancy and recovery. Hard drive storage is a type of non-volatile memory stored internally in computers and data centers. Modern hard discs are checked for progress using the following block sizes: They use low-drive instructions to harden the data. Using modern technology, the block size is specified when there is a manufacturer's specification for the product. External hard drives can expand the storage capacity of attached devices. The hard drive stores the operating system data and the software applications it contains. They generally transfer data slower than internal hard drives. These traces are cyclically enhanced according to their distance from the center. These use slower data transfers due to their slow transfer speed.

Computer code gets easily damaged which self-replicates virus. These virus replicates itself and corrupts those files within the computer memory. These viruses can infect both the application and software which acknowledges the user and injects without user knowledge. These have the ability to steal private information from infected files. Viruses have the ability to reproduce and infect a victim's software and system. By engaging in malevolent conduct, they cause victim actions. Data complexity may be decreased by intelligently detecting hard disc failures using artificial intelligence. One of the assaults that might jeopardise the confidentiality and privacy safeguards on your system is boot sector malware. Some boot viruses insert their victims where they already exist in the program. Malware can control the entire host system, where it injects the system memory to reduce the payload and also increases traffic consumption. Different viruses can easily halt the process where these stored viruses can infect the operating system. Such viruses can reform or clean up the environment by removing those data. These can easily breach data confidentiality and privacy. Boot sector viruses can potentially cause damage to the computer system.

2. Literature Review

Author G. Wang [1] studies that the most serious problem in data centre reliability is hard disc drive failure. We treat long-term HDD information as a scenario and divide it into multiple cases. Therefore, predicting HDD failures is an important step in ensuring information centre storage protection. A study of HDD data from a telecom company and his Back blaze database server shows that using the proposed strategy yields much better results than other methods. Such self-monitoring, analysis, and monitoring techniques in recovered failed hard discs contain large series containing many unsupervised learning opportunities, even with a high degree of mixture of good and failed data.

Salkhordeh [2] introduced a paper on the storage subsystem is characterised as the computing system's efficiency constraint. SSD cache is used to reduce the drawbacks of SSDs while utilising their great performance. Cache strategies haven't addressed these factors in recommended systems and have exclusively concentrated on one evaluation metric. To ensure a successful exchange among efficiency and sustainability, the suggested architecture adjusts optimum rules based mostly on workload parameters. Researchers had put the suggested design into practise on a system that has commercial hard discs and SSDs. These empirical findings demonstrate that the suggested architecture enhances efficiency while also lowering power utilization.

According to Li [3], most HDDs degrade with time, but traditional technologies are unable to accurately capture this loss. This method may address warnings produced by the estimation method in accordance with the severity of each alert. Additionally, we provide a health level system based on regression trees that may provide the user with a health evaluation as opposed to just a categorization result. To create an external hard predictive model utilising the SMART properties, certain statistical and artificial intelligence approaches were presented. These techniques have shown high levels of predictability. Massive storage devices' dependability may be greatly increased while their maintenance and design costs can be decreased thanks to the categorization approach.

Wang [4] proposed, based on this research, that multiple parameterized techniques were created to use a collection of predictive methods to forecast when HDDs will fail. When a particular interval anomaly incidence is determined to be statistically significant, it is an indication that the HDD is about to fail. After that, a synthesised set of data was used to demonstrate the usefulness of the established technique for forecasting errors. To avoid losing important data, it is critical to foresee when a hard disc may fail. It's crucial to strike a balance between the incidence of failure recognition and false alarms, in addition to giving users prior notice of HDD failures so they can back up sensitive data on a timely basis.

Takashima [5] developed research where it shows how non-volatile FeRAM caching can increase hard disc drive efficiency. An idea of non-volatile FeRAM caching is offered in order to fully use memory space while disobeying Windows operating system flush cache directives. To fulfil HDD criteria, a 128-MB chain FeRAM matrix structure and information route structure, as well as a total control power grid, are given. This enhanced HDD efficiency is proven using simulations and measurements. According to Y. Shiroishi [6], the existing design of volatile RAM may be replaced by a future computer architecture employing elevated non-volatile memory space and higher capacity HDDs. To circumvent the small track header readability and super-paramagnetic limitation of medium concerns, various rapid technology possibilities for HDD storage were presented and are currently being explored. The HDD sector is at a crucial juncture in technology,

and thus it is crucial that we proceed ahead with thorough plans for pushing over the superparamagnetic limitation.

3. Methodology

Data Pre-processing

Organize all those data, check for erroneous attributes to create valuable information, and use nulls to fill in missing values. Detect malfunctions by recognising attributes and predictively detecting comparisons across datasets using machine learning algorithms and AI. The presence of unfiltered attributes reduces the quality of overall data performance. Statistical data parameters are trained in data processing. In general, the data are inconsistent and incomplete, which is said to be "raw data." Raw data needs to be transformed into an understandable format. Data parsing produces meaningful data where the encoded data gets interpreted. An intelligent way to report data is to check data monitoring software that automatically detects hard disc malfunctions.

Data Training

Training a dataset allows you to make more accurate predictions based on the model data. By labelling the data, the credibility of the processed information can be increased specifically for further processing. An input set validates inputs and produces results. an initial dataset to train the programme to understand and obtain refined results. Trustworthy information deals with identifying data. Reports are generated sequentially based on test results. Untagged data is typically expanded when tagged data is tagged according to data sampling. This test model evaluates performance metrics based on observations.

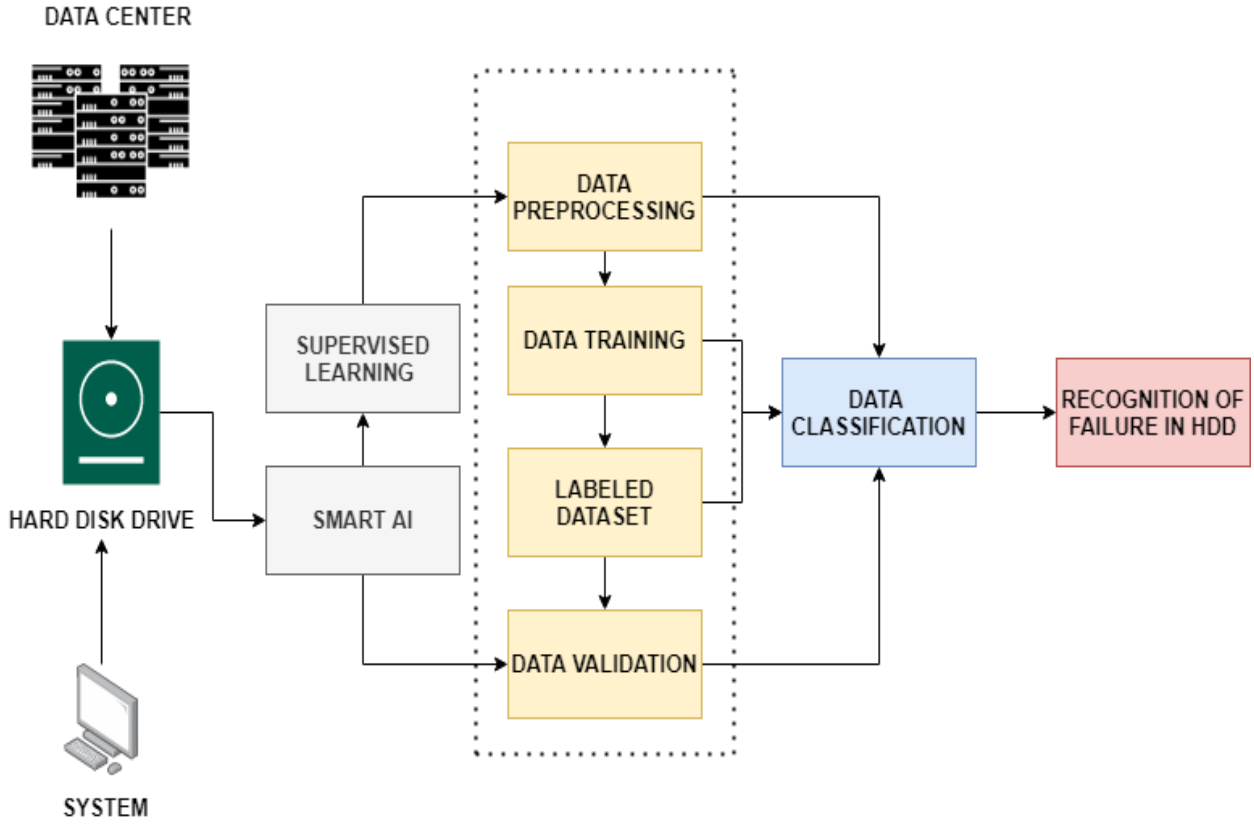

Figure 1: Systematic Representation of Analyzing Malfunction in HDD

Data Validation

Data validation checks the data for errors and inaccurately records them. Typically occurring data is split into smaller sorts of fraction depending on iteration. Data validation is carried out in accordance with limitations like objectives. This is implemented using a number of techniques before moving on to data integrity techniques. analyse and combine data to check for mistakes and assess the system's correctness. Such data is verified to make sure that it is appropriate and to allow for a data cleaning procedure that creates uniformity in data usage.

Supervised learning approach

In order to generate data and outcomes for forecasting models, algorithms for machine learning employ a supervised method. One form of classification algorithm that samples a hyperplane into several or more category streams is the support vector machine approach. Data points are classified using class labels based on a collection of characteristics. Using anticipated and actual grades as a comparison, an additional matrix prediction technique is used. The various decision tree algorithms that finally disperse the random variables are predicted by a forest random classifier. The overall accuracy of this method is determined using the following formula: The representations of TP, TN, FP, and FN are:

$$\text{False Positive Rate (FP)} = \frac{FP}{(TN+FP)}$$

$$\text{True Positive Rate (TP)} = \frac{TP}{(TP+FN)}$$

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

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

Whenever cross-validation is used for efficiency evaluation, it employs normalised data. The likelihood of categorising the issue is determined by sensitivity, while the likelihood of generating inaccurate predictions is determined by specificity. To verify efficiency using test suites, the aforementioned equations are generated. To prevent any form of possible bias, it has been established that they are overfit.

Recognition of Failure

Drive Size: Shows the size of the total size of the HDD.
Drive Days Count: total time the hard drive has been in use.
Drive number - Range of drives
Drive Error: Hard Drive Status is Failed or Down
Error rate: percentage of error conditions
Defect level: low, high, or medium status
Unsupported File: Yes/No

4. Experimental Results

Drive Supported Criterion

Drive Size	***Drive Days Count***	***Drive Count***	***Drive Failure***	***Failure Rate***	***Defective Level***	***Unsupported File***
4 Tb	80,689	3,953	9	0.28%	Medium	No
4 Tb	60,638	1,500	0	0%	-	-
8 Tb	1,60,009	9,575	16	0.60%	High	Yes
8 Tb	1,36,313	13,363	13	0.30%	Medium	No

4 Tb	1,73,954	19,572	12	0.12%	Medium	No
6 Tb	80,657	896	14	0.35%	Medium	No
10 Tb	1,75,388	1400	82	0.90%	High	Yes
12 Tb	3,19,001	15,655	26	0.76%	High	Yes
14 Tb	6,65,477	8,343	20	1.15%	High	Yes

Table 1: Various attributes for the recognition of failure in HDD dataset

In this paper, a large and drive-supported dataset is used to predict the HDD drive. The graph depicts the benefits and drawbacks of cross-validation when utilising various hard disc drive identification techniques. A hard drive determines if it is a good drive or a bad drive based on certain criteria. Users may find out when a drive needs to be replaced through drive deployment.
We may restore the drive as previously configured and check the drive state after the drive comparative evaluation. Next-step algorithms are carried out using intelligent monitoring approaches such as hard disc drive technology, self-monitoring, analysis, and reporting. It employs a machine-learning approach, where the decision trees are followed by a threshold range D that represents the percentage of all samples. Wh denotes size, for example, WH.

Different Algorithm	***Accuracy***	***Precision***	***Recall***	***F-measure***
Support Vector Machine	0.868	0.985	0.006	0.015
Decision Tree Algorithm	0.676	0.787	0.665	0.756
Random Forest (Multiple Decision Trees)	0.987	0.961	0.946	0.967

Table 2: Cross-Validation Ranges of Different Algorithm

(1) Initiate a sample using the sampling algorithm.
(2) Use each tree T after the sampling process.
(3) Compute the classification step using the tree classification and the sample S.
(4) Check whether the abnormal samples are relatively large.
(5) Check if the hard disc is normal or if any parts are damaged.
(6) Prediction results are generated.
(7) Terminate the process.

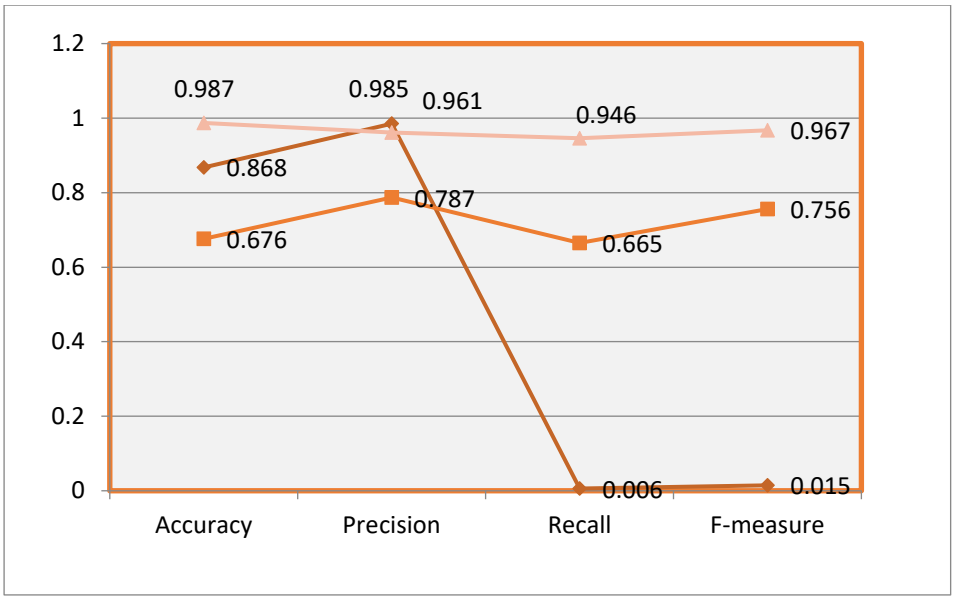

Figure 2: Predicted Analysis

Hard disc self-monitoring can analyse failures and is considered a method of early failure detection. These allow users to be aware of their data and automatically back up their data. Self-monitoring, analysis, and reporting methods check and monitor various attributes of hard disc drives. Intellectual use of artificial intelligence can check drive reliability and determine error rates. Check various attributes to determine the error rate. Hard drives can typically hold up to a terabyte of data, which is stored in sequential order. Physical damage to a hard drive might result in a physical failure when the stored data is fixed improperly. Through lowering the payload and boosting bandwidth usage, hard drive viruses can infect the memory management. Data complexity may be reduced by employing AI technology to discover hard disc issues proactively.

5. Conclusion

Some of the configured computers are malfunctioning because of hard disc failure. Due mainly to some natural calamities and external factors, the external hard disc gets impacted and failure happens. Storage system reliability can be handled using the prediction of hard disc failure. The smart way of detecting that malfunction in HDD with those detailed reports, such as malware and viruses, is determined using the AI. The lack of efficiency leads to time complexity with continuous monitoring. In this proposed approach, the AI is used to predict the failure over the hard drive model for identifying the exact precision. Due to those improper ways of handling the system, proactive failure occurs. With the enhanced accuracy, the data outline values are easily attained. Using those supervised and unsupervised approaches, each set of data is computationally handled in trained and tested sets. These HDD can reduce speed, which increases the runtime complexity with threshold range. One of the easy ways of handling failure is by analysing and identifying those failures in advance.

References

1. G. Wang, Y. Wang and X. Sun, "Multi-Instance Deep Learning Based on Attention Mechanism for Failure Prediction of Unlabeled Hard Disk Drives," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-9, 2021, Art no. 3513509, doi: 10.1109/TIM.2021.3068180.

2. R. Salkhordeh, M. Hadizadeh and H. Asadi, "An Efficient Hybrid I/O Caching Architecture Using Heterogeneous SSDs," in IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 6, pp. 1238-1250, 1 June 2019, doi: 10.1109/TPDS.2018.2883745.
3. Li, Jing & Ji, Xinpu & Jia, Yuhan & Zhu, Bingpeng & Wang, Gang & Li, Zhongwei & Liu, Xiaoguang. (2014). Hard Drive Failure Prediction Using Classification and Regression Trees. Proceedings of the International Conference on Dependable Systems and Networks. 383-394. 10.1109/DSN.2014.44.
4. Y. Wang, E. W. M. Ma, T. W. S. Chow and K. -L. Tsui, "A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives," in IEEE Transactions on Industrial Informatics, vol. 10, no. 1, pp. 419-430, Feb. 2014, doi: 10.1109/TII.2013.2264060.
5. D. Takashima, Y. Nagadomi, K. Hatsuda, Y. Watanabe and S. Fujii, "A 128 Mb Chain FeRAM and System Design for HDD Application and Enhanced HDD Performance," in IEEE Journal of Solid-State Circuits, vol. 46, no. 2, pp. 530-536, Feb. 2011, doi: 10.1109/JSSC.2010.2091324.
6. Y. Shiroishi et al., "Future Options for HDD Storage," in IEEE Transactions on Magnetics, vol. 45, no. 10, pp. 3816-3822, Oct. 2009, doi: 10.1109/TMAG.2009.2024879.
7. Thiagarajan.R & Moorthi.M “Quality of Service Based Adhoc Ondemand Multipath Distance Vector Routing Protocol in Mobile AD HOC Network”, Journal of Ambient Intelligence & Humanized Computing , Springer, vol.12, no. 5, April 2020.
8. B. D. Strom, S. Lee, G. W. Tyndall and A. Khurshudov, "Hard Disk Drive Reliability Modeling and Failure Prediction," in IEEE Transactions on Magnetics, vol. 43, no. 9, pp. 3676-3684, Sept. 2007, doi: 10.1109/TMAG.2007.902969.
9. Thiagarajan.R, Moorthi.M A Billing Scheme of Tollbooth in Service Oriented Vehicular Network, UGC Journal-JNCET, vol.8, issue 3 May 2018.
10. Murray, Joseph & Hughes, Gordon & Kreutz-Delgado, Ken. (2005). Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application. Journal of Machine Learning Research. 6. 783-816.
11. Haeng-Soo Lee, Young-Hoon Kim, Tae-Yeon Hwang and Cheol-Soon Kim, "VCM design to improve dynamic performance of an actuator in a disk drive," in IEEE Transactions on Magnetics, vol. 41, no. 2, pp. 774-778, Feb. 2005, doi: 10.1109/TMAG.2004.840312.
12. Thiagarajan.R , Moorthi. M “Energy consumption and network connectivity based on Novel-LEACH-POS protocol networks”,Computer Communications, Elsevier, (0140-3664), vol.149,pp. 90-98, November 2019.
13. Ki Myung Lee and A. A. Polycarpou, "Transient and steady state dynamic vibration measurements of HDD," Digest of the Asia-Pacific Magnetic Recording Conference, 2002, pp. TU-TU, doi: 10.1109/APMRC.2002.1037674.