iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52604

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 V May 2023- Available at www.ijraset.com

Hard Disk Failure Prediction Using Machine Learning

Prof. R. T. Waghmode[1], Abdulmuiz Shaikh[2], Shreyash Bajhal[3], Harsh Dubey[4], Anuj Bhadoriya[5]
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. RELEVANCE

The project is intended for individual people who are looking for a life prediction of their hard drives.Output from the system will be Fail or Pass. The hard disk failure can cause data loss.Hence predicting failure a way ahead will help people avoid data loss. This can also be brought into the data centres where huge number of hard drives are present.This project can be scaled for data centres.This will avoid data loss and help data centre managers take necessary action before it fails.

III. LITERATURE SURVEY

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 orenterprises. 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 .

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 cor relations among different failures, including batch and repeating failures, as well as the human operators' response to the failures.

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.

©IJRASET: All Rights are Reserved

IV. METHODOLOGY

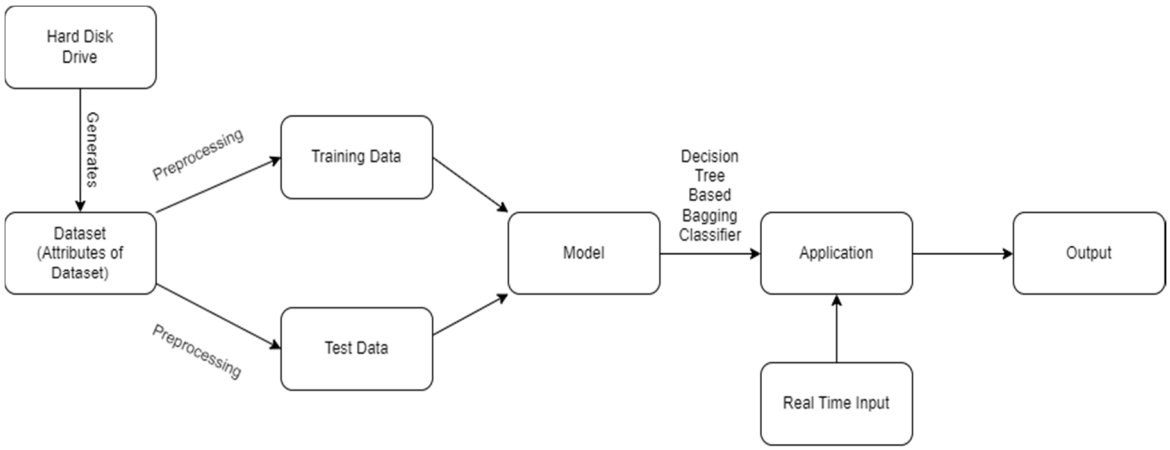


Figure 1: Architecture Diagram

Even if the experimental design process is shared by all study areas in some ways, ML tactics need to be cross-disciplinary. The ML technique's steps that are unique for hard disk failure prediction are as follows: The key five steps are data collection, Preparing the Data, Choosing a Model and Training the Model, Evaluating the Model and System Integration.

A. Data Collection

Collecting data for hard disk failure prediction typically involves monitoring the system performance using various sensors and diagnostic tools. These sensors can collect data on various parameters, such as temperature, humidity, vibration, and noise, that can provide insights into the health of the hard disk.

There are several ways to collect data for hard disk failure prediction, including

1) *Onboard sensors*: Many hard disks come with onboard sensors that can provide information about the disk's performance, such as temperature, read/write errors, and spin-up time.
2) *SMART (Self-Monitoring, Analysis, and Reporting Technology) data*: SMART is a technology that allows hard disks to report various metrics about their performance, such as reallocated sectors, spin retry count, and uncorrectable errors.
3) *External sensors*: In addition to onboard sensors, external sensors can also be used to monitor the environment around the hard disk, such as temperature and humidity.
4) *Log files*: Operating systems and applications may also generate log files that can provide information about the usage patterns and workload of the hard disk.

B. Preparing the Data

Preparing the data for hard disk failure prediction involves several important steps to ensure the data is in a suitable format for training and evaluating the prediction model. The following steps are commonly involved in preparing the data:

1) *Data Cleaning*: The first step is to clean the data by identifying and handling any missing values, outliers, or erroneous entries. Missing values can be imputed using techniques such as mean, median, or interpolation. Outliers can be detected and either removed or adjusted based on domain knowledge or statistical methods.
2) *Feature Selection*: Next, feature selection techniques can be applied to identify the most relevant and informative features for predicting hard disk failure. This involves analyzing the correlation between different features and the target variable and selecting the features that contribute the most to the predictive power of the model. Techniques such as correlation analysis, feature importance ranking, or dimensionality reduction

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

3) *Data Splitting*: After preparing the features, the data needs to be split into training, validation, and testing sets. The training set is used to train the model, the validation set is used for hyperparameter tuning and model evaluation during development, and the testing set is used to assess the final model's performance.
4) *Data Balancing*: In cases where the dataset is imbalanced, where the number of instances of one class (e.g., failure) is significantly smaller than the other class (e.g., non-failure), techniques such as oversampling the minority class or undersampling the majority class can be applied to balance the data and prevent bias in the model.

C. *Choosing a Model and Training the Model*

Choosing a suitable model and training it are crucial steps in hard disk failure prediction.Some important steps are:

1) *Model Selection*: There are various machine learning models that can be used for hard disk failure prediction, and the choice depends on the specific requirements of the problem. Some common models include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. The selection can be based on factors such as the complexity of the problem, interpretability of the model, and the availability of data.
2) *Model Training:* Once a model is selected, the next step is to train it using the prepared dataset. This involves feeding the labeled training data into the model and adjusting its parameters to minimize the prediction error. The training process typically involves an optimization algorithm, such as gradient descent, that iteratively updates the model's parameters based on the training data.
3) *Model Evaluation*: After training the model, it is essential to evaluate its performance on unseen data. This is typically done using a validation set or through cross-validation techniques. Performance metrics such as accuracy, precision, recall, F1-score, or area under the receiver operating characteristic (ROC) curve can be used to assess the model's predictive ability. It is important to choose appropriate evaluation metrics that align with the problem's specific requirements and considerations.
4) *Iterative Refinement*: If the model's performance is not satisfactory, further iterations of model selection, hyperparameter tuning, and training may be required. This iterative refinement process helps improve the model's accuracy and generalization ability.

D. *Evaluating the Model*

Evaluating the model in hard disk failure prediction is a critical step to assess its performance and reliability:
Cross-Validation: Employ cross-validation techniques to assess the model's generalization capability and robustness. K-fold cross-validation divides the dataset into k equally sized folds, with each fold serving as a testing set while the remaining folds are used for training. This process is repeated k times, and the average performance across all folds provides a more reliable estimate of the model's performance.

E. *System Integration*

System integration in hard disk failure prediction using Python and Tkinter involves incorporating Tkinter functionality to create a desktop application with a graphical user interface (GUI) for interacting with the prediction system.

1) *Tkinter Integration and GUI Design*: Tkinter is a standard Python library for creating GUI applications. Integrate Tkinter into the system to design and implement the frontend interface for the hard disk failure prediction application
2) Utilize Tkinter's widget toolkit to design the graphical user interface. Create windows, frames, labels, buttons, entry fields, and other GUI elements to build an intuitive and interactive interface for users.
3) *User Input*: Implement input mechanisms in the GUI for users to provide relevant data for hard disk failure prediction. This may include options for entering SMART data manually, importing data from files, or connecting to external data sources. Tkinter's entry fields, file dialogues, and data import functions can be utilized for this purpose.
4) *Data Processing*: Integrate the necessary data processing functionalities into the backend logic of the system. Use Python's data manipulation libraries such as NumPy and pandas to preprocess and transform the input data before passing it to the prediction model. Implement data cleaning, normalization, feature extraction, and any other required data processing steps.
5) *Model Integration*: Incorporate the trained hard disk failure prediction model into the system's backend logic. Connect the model's functionality to the GUI to receive input data, process it using the trained model, and generate predictions or risk scores. Ensure the integration is seamless, allowing for efficient and accurate predictions.
6) *Real-time Monitoring*: Enable real-time monitoring of hard disk health and failure prediction within the GUI. Continuously collect the SMART data or other relevant data, process it using the integrated model, and update the monitoring results dynamically. Display the real-time prediction outcomes or risk scores to provide users with up-todate information.



Fig: Home page of application

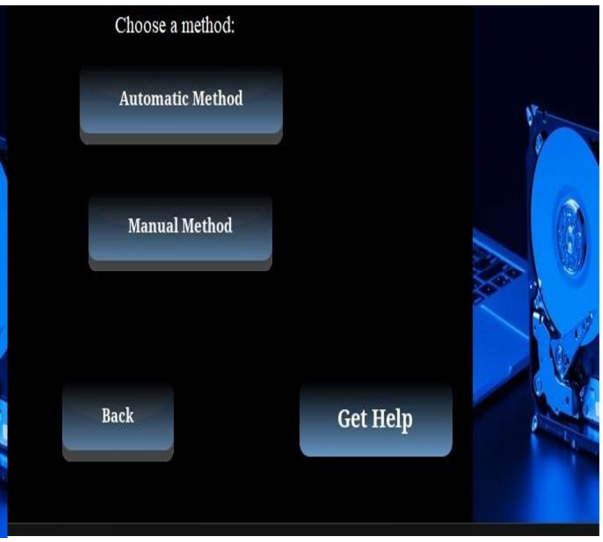


Fig: Choosing a method for result

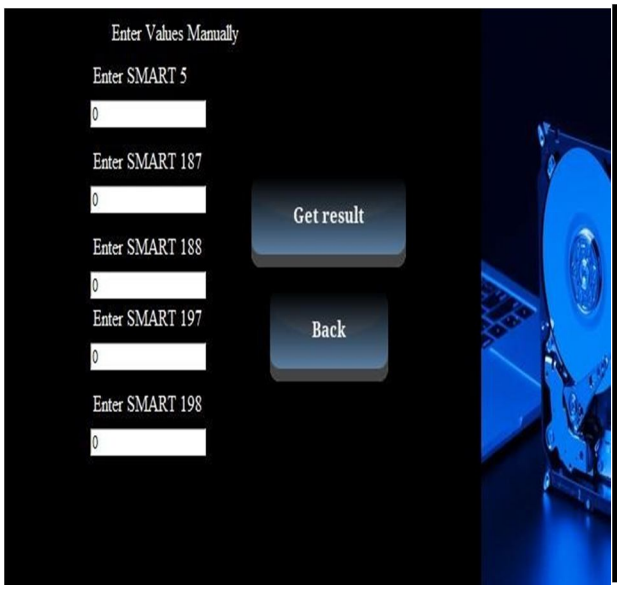


Fig: Manual checking

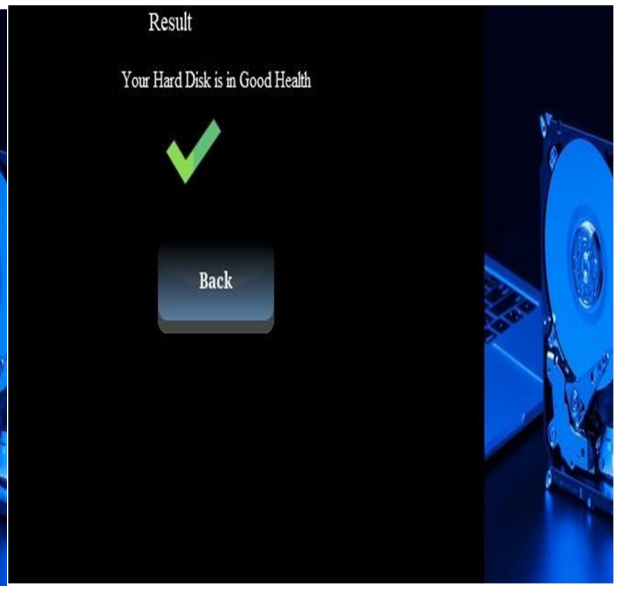


Fig: Final output

V. ALGORITHMS

1) *Bernoulli Naive Bayes:* The algorithm calculates the likelihood of each feature occurring given each class label and estimates the prior probability of each class label. It then applies Bayes' theorem to calculate the posterior probability of each class given the observed features. The class with the highest posterior probability is assigned as the predicted class label.
2) *Bagging With Decision Tree:* Bagging with Decision Trees is a powerful ensemble learning technique used for classification and regression tasks. Bagging, short for Bootstrap Aggregating, aims to improve the stability and accuracy of individual models by combining multiple models trained on different subsets of the dataset.

In the case of Bagging with Decision Trees, multiple decision tree models are trained using bootstrap sampling. Bootstrap sampling involves randomly selecting subsets of the original dataset with replacement, creating new training sets for each decision tree. This process allows each tree to be trained on slightly different data, introducing diversity in the models.

During training, each decision tree is grown by recursively splitting the data based on different features and thresholds. The trees can be deep or shallow, depending on the complexity of the problem and the desired level of generalization. Each tree independently makes predictions based on its internal structure and the majority vote (in classification) or average (in regression) of the predictions from all trees is used as the final prediction.

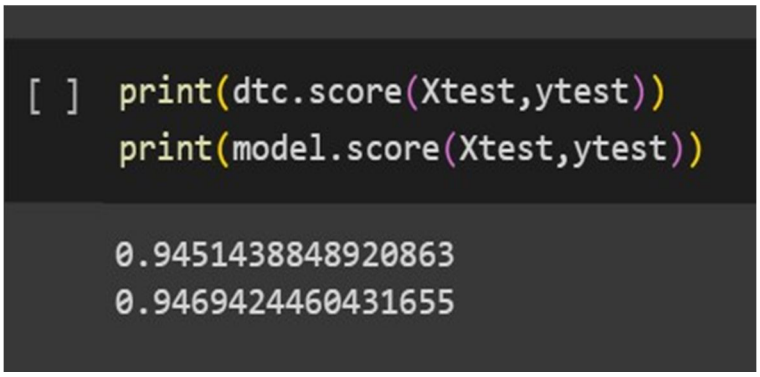


Fig: Accuracy of the model

A. Challenges

1) Feature selection and engineering are critical challenges in hard disk failure prediction. While hard disk SMART attributes provide valuable information about the health and performance of the drives, determining the most informative features and engineering new ones that capture underlying failure patterns can be complex. Domain expertise and feature selection algorithms can assist in identifying the most relevant attributes, but it remains a non-trivial task to extract the most discriminative features from the available data.

2) Moreover, hard disk failures often exhibit complex patterns that may be influenced by various factors such as environmental conditions, usage patterns, or manufacturing defects. Incorporating these factors into the prediction models and accurately capturing the underlying dynamics of failure occurrences pose additional challenges. It requires exploring advanced machine learning techniques, such as ensemble methods, deep learning, or hybrid models that can handle complex patterns and dependencies.

3) Additionally, the dynamic nature of hard disk failure prediction poses a challenge. Hard disks are subject to evolving conditions and usage patterns over time, and prediction models should be capable of adapting to these changes. Developing models that can handle real-time or near-real-time data updates, incorporate feedback mechanisms, and adjust predictions based on the latest information is a challenge that requires careful design and implementation.

VI. FUTURE SCOPE

With the proliferation of cloud computing and remote monitoring capabilities, hard disk failure prediction systems can be integrated into cloud-based platforms. This enables centralized monitoring and analysis of large-scale disk arrays, providing insights into the health and failure risks of multiple hard disks simultaneously.

Developing real-time monitoring systems that continuously collect and analyze hard disk data can provide timely alerts and notifications when the risk of failure increases. This allows for immediate action to be taken, such as data backup, migration, or replacement, to minimize the impact of potential failures.

Instead of relying solely on SMART data, the fusion of multiple data sources and modalities can provide a comprehensive view of the hard disk's health. Incorporating additional sensor data, such as temperature, vibration, or acoustic signals, along with SMART attributes, can improve the predictive capabilities and enable more robust failure detection.

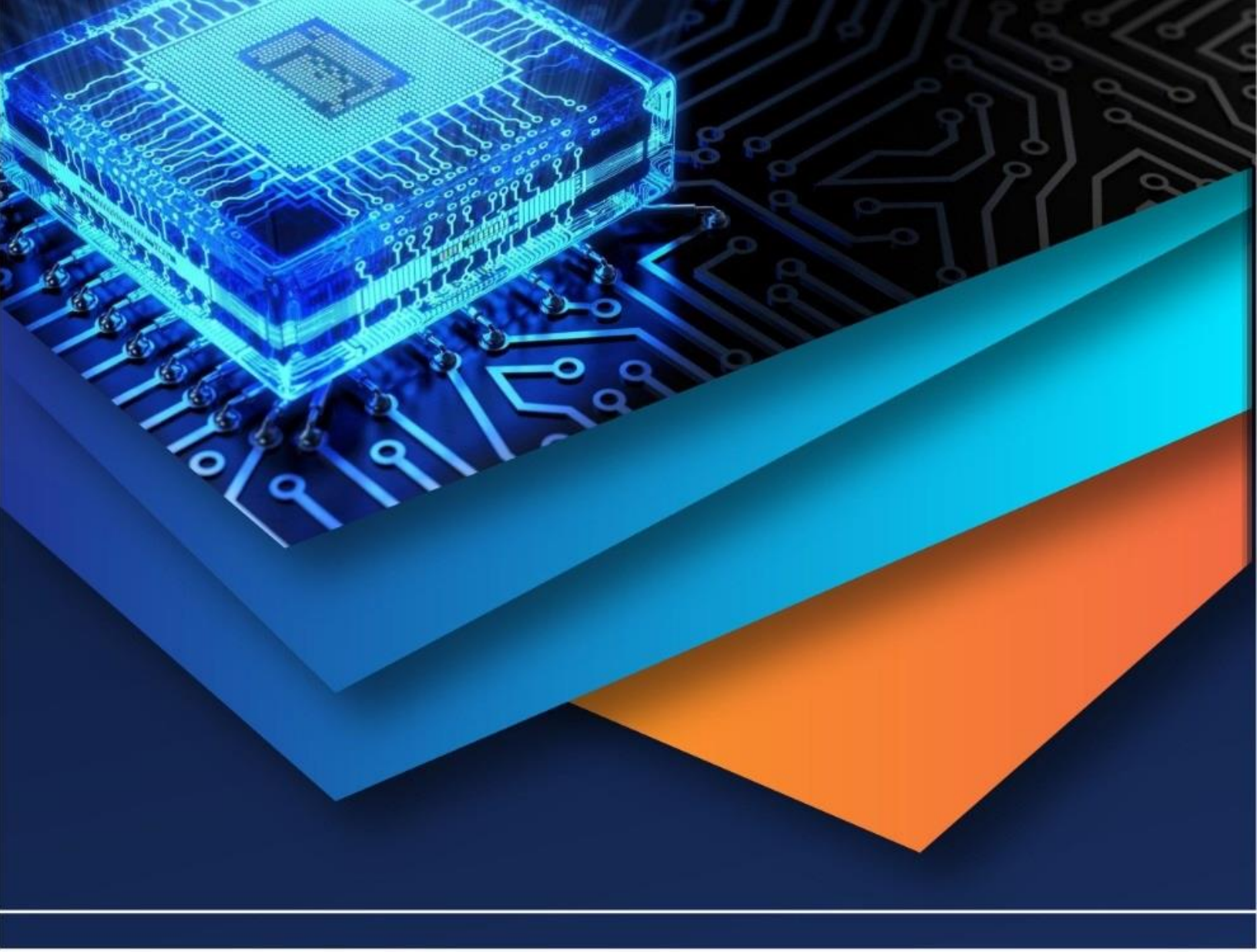
VII. CONCLUSIONS

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

Call : 08813907089 (24*7 Support on Whatsapp)