How to Cite:
Naresh Kumar, M., & Prasika, L. (2022). Fuzzy based sequential learning for HDD health status prediction in data center. *International Journal of Health Sciences*, *6*(S2), 3261–3266. https://doi.org/10.53730/ijhs.v6nS2.5812

Fuzzy based sequential learning for HDD health status prediction in data center

Naresh Kumar M
Final year MCA students at MEPCO Schlenk Engineering College, Sivakasi, Tamilnadu, India
Email: nareshkumar230599@gmail.com

Prasika L.
Assistant Professor in the MCA department of MEPCO Schlenk Engineering College, Sivakasi, Tamilnadu, India
Email: lprasika@mepcoeng.ac.in

***Abstract*---**Data center is the essential part in maintaining data stores and IT systems in an enterprise. HDD (Hard Disk Drive) plays a crucial role in datacenter.The reliability of HDD is to be considered as an important factor.Prevention of Failure in HDD is significant for enabling security over data.In this paper, an Ensemble based sequential learning model was proposed to predict the failure of HDD at earliest.Sequential model learned the history of data and predicts the failure.It learned the pattern and predict the failures.In our approach, LSTM and GRU learning models are used for sequential learning. Ensemble learning was combined with these sequential model for predicting disk failure.The continous data from HDD was observed as SMART data and is used for constructing learning models to predict the results with an accuracy of 89.3%.

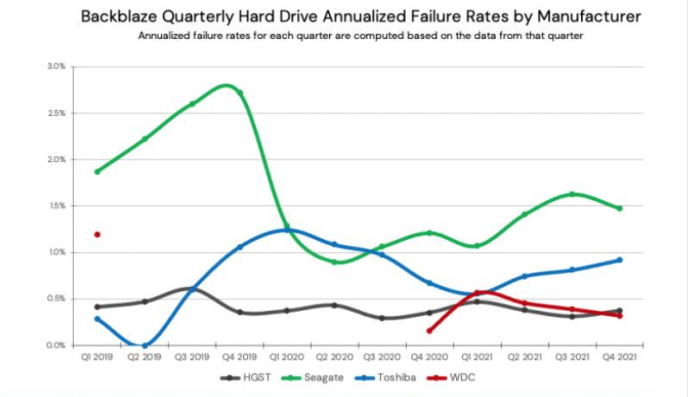
***Keywords*---**sequential learning, HDD failure, SMART, fuzzy ensemble learning.

Introduction

HDD and SDD are the important components of Data center which will enable availability and Reliability of Data.In some places,SDD is preferred over HDD where cost is not a concern.HDD is preferred for providing efficient feasible solution to ensure data availablilty.For a large enterprise, failure of HDD leads to reduction of customer's service and affects the revenue of the company.Sudden failure of HDD had a great negative impact for an enterprise.To avoid this, early

prediction of HDD failure has to be identified to reduce the unsatisfaction of customers.Figure 1 shows the stats of HDD failure by BackBlaze.

The dataset for predicting HDD failure, SMART (Self-Monitoring, Analysis and Reporting Technology) data was collected.It is a system for monitoring internal information of the drive.If the failure is predictable then that can be mined from S.M.A.R.T data.It has 50 features which includes attributes related to predictable failures and non predictable failures. Predictable failure can be forecasted by Machine learning and Deep learning prediction algorithms.Reallocated Sector count (SMART 5), Current pending Sector Count (SMART 197), Reported Uncorrectable Errors (SMART 187), Erase Fail Count, Wear Leveling Count, Disk Temperature.For monitoring Power Cycles , SMART 12 was considered. SMART data is an efficient way to forecasting the health status of HDD. Since the forecasting can be done from the observance of continous data,Sequence learning yields to good results.From the combination of all results given from different learning models. The proposed approach make use of the Sequence learning ,Ensemble learning and Fuzzy inference system for forecasting the health status of HDD.



	Q1 2019	Q2 2019	Q3 2019	Q4 2019	Q1 2020	Q2 2020	Q3 2020	Q4 2020	Q1 2021	Q2 2021	Q3 2021	Q4 2021
HGST	0.41%	0.47%	0.61%	0.36%	0.37%	0.43%	0.30%	0.35%	0.47%	0.38%	0.31%	0.37%
Seagate	1.87%	2.22%	2.60%	2.72%	1.29%	0.90%	1.06%	1.21%	1.07%	1.41%	1.63%	1.48%
Toshiba	0.29%	0.00%	0.60%	1.06%	1.24%	1.09%	0.98%	0.67%	0.56%	0.75%	0.81%	0.92%
WDC	1.20%							0.16%	0.57%	0.46%	0.39%	0.32%

Fig. 1. Reports by BlackBlaze

Literature Survey

Ample research works were done for the prediction of Health status of HDD using different algorithms.Using SMART attribute, HDD failure prediction was analyzed[1]. Vikas Tomer et al.,[2] uses efficient feature selection mechanism for predicting HDD failure by means of SMART attributes.Feature selection is important since SMART deals with 50 attributes.Finding the relevant features can be done for improving the efficiency of the forecasting system. [3]Qiang Li et al., used XGBoost and LSTM for prediction of failure in Hard disk.The Authors used feature selection in combination with Ensemble learning. Tek Raj Chhetri et al., created a knowledge Graph for learning the behaviour of Hard disk and that was used for indicating different value changes in specific attributes. Jing Shen et al., used LSTM based edge computing for dealing with SMART attributes[4]. Joseph F

Murray et al., explored the different Machine Learning models for the failure prediction of HDD[5].

Wasim Ahmad et al., used evaluation algorithms for feature selection which leads to the efficient features for forecasting the health status[6]. S Geetha et al., used LSTM model for predicting the smog level[9]. Jing Li et al., used decision trees for efficient feature selection and then failure prediction[10]. The existing research work focused on simple rule based and Machine learning model based prediction.Our proposed approach used Fuzzy Inference System which use Takagi Sugeno Model which appropriately classify the predicted output into membership classes (Good, Moderate, Below Average, Critical) for representing Health status.

Methodology

Our Proposed approach forecast the health status of hard drive using sequential learning , Ensemble learning and fuzzy inference System.The system had the following modules:

1. Data Preprocessing
2. Sequence Learning
3. Ensembling
4. Display Health status using Fuzzy Inference System

Training and Testing samples were taken in the ration of 70:30 using cross validation. The overall System design which includes Training and Testing process was shown in Fig.2.Fuzzification and defuzzification process is included in FIS.We named our model as SEF since it involves Sequence Learning ,Ensemble learning and Fuzzy mechanism.

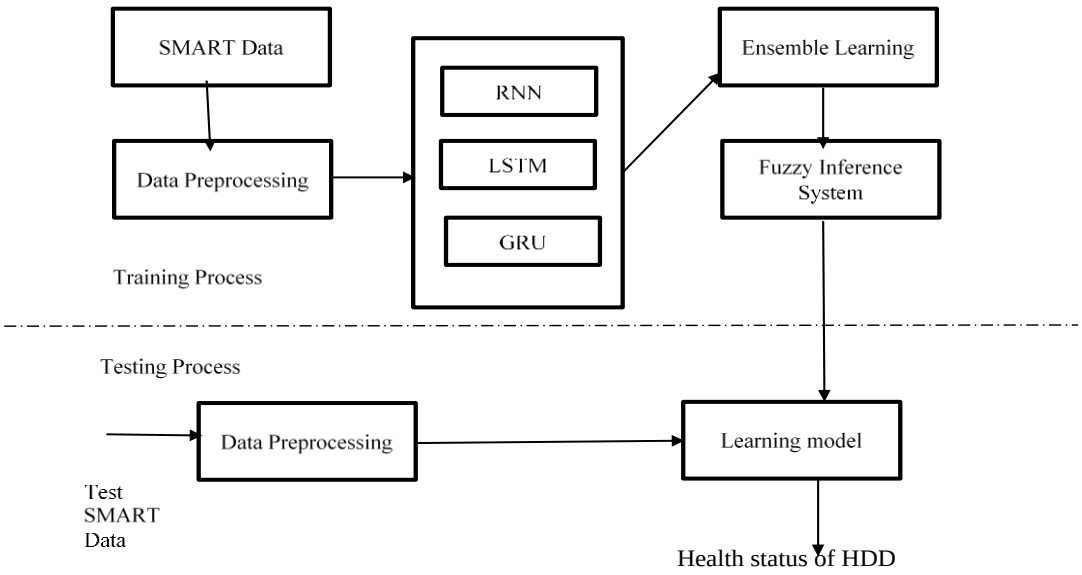


Fig. 2. Overall System Design

Data Pre-Processing

SMART data collected had some missing values.Data pre-processing was essential.Data Normalization is needed for further analysis of SMART data.MinMax normalization is used for normalizing all attributes which had the exact same scale but it will not handle outliers.The formula for min max normalization is given below:

$$\text{Normalized Value} = \frac{(val - n1)}{(n2 - n1)}\,(m2 - m1) + m1$$

For handling missing values,Data impuation by mean or median was used.After preprocessing,the predicatable SMART attributes were ready for further creation of learning models.

Sequence Learning

The forecasting needs continous monitoring of SMART data from which the increase of values of the specified predicatable attributes can be identified.For this prediction,sequence to sequence model was used since it performed well for data which was continously monitored data.The sequence models used are RNN , LSTM, GRU.Simple RNN or vannila RNN has the drawback of vanishing and exploding gradients.By means of forwarding the feedback,gradient were vanished or it may go to infinity.To avoid the vanishing gradient problem, LSTM was used for prediction.In compare with LSTM, GRU work faster and it use less memory.In our proposed approach,the decisions from RNN, LSTM and GRU was taken and it will be given to ensemble learning for better forecasting.

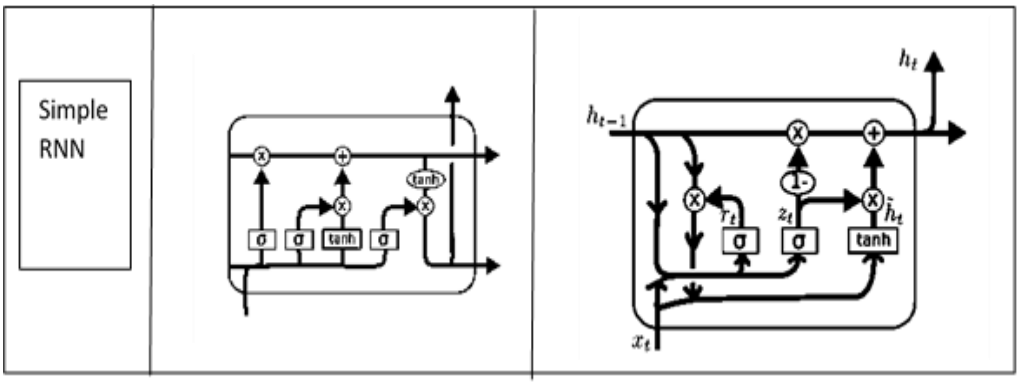


Fig 3. Sequential learning - Process

X : Scaling of information
+ : Adding information
σ : Sigmoid layer tanh: tanh layer
h(t-1) : Output of last LSTM unit
c(t-1) : Memory from last LSTM unit
X(t) : Current input
c(t) : updated memory
h(t) : Current output

Ensemble learning

Decision taken from multiple predictions can be used for better prediction.There were lot of ensemble learning methodologies available in fine tuning the performance. Our proposed approach make use of Boosting method to improve the prediction results.

Fuzzy Inference System

The result of ensembling was a numerical value which has to be mapped with 3 status 'Good', 'Moderate', 'Below Average' and 'Critical'. This was achieved by using Takagi – Sugeno model with Gaussian membership function. FIS (Fuzzy Inference System) had the following process : Fuzzification and Defuzzification which used the components Fuzzy Rule base and knowledge base. Table 1. Shows the health status of HDD based on the predicted value from ensemble learning.

Table 1. Health status of HDD based on precise values

Status	**Value**
Good	0<y<0.5
Moderate	0.5<y<1
Below Average	1<y<1.5
Critical	<1.5

Expremental Result

Performance of the proposed approach was evaluated using measures like Accuracy, F1-score, Precision and recall.

Precision = TP/(TP+FP) Recall = TP/(TP+FN)
Accuracy = (TP+TN)/(TP+FP+TN+FN)

We have compared our approach with Random Forest , Naïve Bayes and Decision tree algorithms and our models (SEF – Sequence learning ,Ensembling and Fuzzy) outperforms the others.Table 2 shows the experimental results in comparison with other approaches.

Table 2. Experimental Results

Learning Model	Precision(%)	Recall(%)	Accuracy(%)
Decision Tree	85	80.2	82
Naïve Bayes	81.4	81.2	82.4
Random Forest	89.3	88	87
SEF model	89.3	88.2	88.3

Conclusion

Health status prediction of Hard drive is a significant research work in the field of Big data center.The approach used seq2seq models and ensembling techniques

for the prediction of Health status.The final output can be retrieved from a fuzzy inference system which handled the fuzzy values perfectly.This fuzzy based mechanism further enhanced the accuracy. When dealing with fuzzified values, it was difficult for mapping to a final class.Fuzzy inference mechanism made this work as an efficient one.Ensembling technique used in this approach was helped the model to learn better.

Future Work

The proposed approach was not focused on feature selection.To improve this,Genetic based feature selection or wrapper based feature selection can be included.This approach can be further enhanced by transfer learning based approaches.Since the SMART data was received from continous monitoring, handling missing values was challenging.To improve this effective imputation methods can be used.

References

1. Vikas Tomer , Vedna Sharma , Sonali Gupta , Devesh Pratap Singh,“Hard disk drive failure prediction using SMART attribute”, Materials Today: Proceedings,2021
2. Qiang Li , Hui Li, Kai Zhang ,“Prediction of HDD Failures by Ensemble Learning”,IEEE
3. Tek Raj Chhetri , Anelia Kurteva , Jubril Gbolahan Adigun and Anna Fensel ,“Knowledge Graph Based Hard Drive Failure Prediction”,Sensors 2022,Volume:22
4. Jing Shen , Yongjian Ren, Jian Wan , and Yunlong Lan ,“Hard Disk Drive Failure Prediction for Mobile Edge Computing Based on an LSTM Recurrent Neural Network”, Mobile Information Systems
5. Joseph F Murray, Gordon Hughes, Ken Kreutz-Delgado,“Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application”,Journal of Machine Learning Research
6. Wasim Ahmad, Sheraz Ali Khan , Cheol Hong Kim , and Jong-Myon Kim ,“Feature Selection for Improving Failure Detection in Hard Disk Drives Using a Genetic Algorithm and Significance Scores”,Applied Sciences 2020,Volume: 10
7. S. SujaPriyadharsini,S. EdwardRajan ,“An efficient soft-computing technique for extraction of EEG signal from tainted EEG signal”, Applied Soft Computing,2012,Volume 12,Issue 3
8. P. Radha, G. Chandrasekaran, N. Selvakumar ,“Deep Learning based Supervised and Unsupervised Neural networks for analyzing the characteristics of powder composite preforms”,Modelling and Simulation,2020
9. S Geetha, L Prasika, ,“Smog Prediction Model using Time Series with Long-Short Term Memory”, International Journal of Mechanical Engineering and Technology,2019
10. Jing Li,Rebecca J. Stones,Gang Wang,"Hard Drive Failure Prediction using Decision Trees”, Reliability Engineering and System Safety,2017.