International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predictive Failure Detection in AI Datacenters Using BMC Telemetry Analytics

Seshadri Ravikiran Vedula

Submitted:01/08/2023 **Revised**: 08/09/2023 **Accepted**: 20/09/2023

Abstract—Machine learning training, cloud computing, and large-scale data processing are some of the services that are supported by AI datacentres. Since the systems are built with thousands of servers and hardware, they can fail at any moment and damage the service, loss of resources, and escalate the cost of working. Hence, potential failures should be identified early to ensure successful operations of datacentres. It is proposed in this research that an AI datacentre predictive failure detection technique, based on the Baseboard Management Controller (BMC) telemetry analytics service and machine learning, would help implement predictive maintenance plans. Servers are monitored to collect the telemetry data of CPU temperature, GPUs temperature, power consumption, the speed of the fans, and the voltage level and analyze them to detect an abnormal behavior ahead of failure. These models of machine learning such as the Logistic Regression, Support Vector machine, and the random forest are all trained over pre-processed features following preprocessing and extraction of features. The experiments have indicated that the model of the Random Forest is the most effective, although the model has an accuracy of 89.7, precision of 86.3, recall of 88.1 and F1-score of 87.2. Another finding of the results is that these telemetry features like temperature of the GPU and power consumption are good indicators of unstable systems. In the given approach, it is established that BMC telemetry analytics can enhance predictive monitoring and reliability of contemporary AI datacentres to a large extent.

Keywords-*Predictive Failure Detection, BMC Telemetry, AI Datacentres, Machine Learning, Datacentre Reliability, Server Failure Prediction.*

I. INTRODUCTION

Cloud computing, artificial intelligence (AI) and mass data analytics are modern digital services that require a large datacentre. These datacentres are made up of thousands of servers, networking and storage facilities and even power infrastructure that are combined to provide continuous computing. Datacentres are becoming bigger and more complicated as the requirement of the AI training and inference is rising. The administration of such big infrastructures is not an easy task since the hardware may break any time, the software may also break or the performance of that infrastructure may reduce. Historically, datacentre management was anything based on human operators scanning system logs, hardware status and performance indicators in order to identify issues within the systems. With the increase in the number of datacentres to thousands of nodes, manual monitoring is ineffective and time-consuming. Researchers have indicated that the datacentres of the future will be increasingly based on automated, data-driven systems, in which the predictive models will analyze the data about the running systems and make decisions on the systems rather than having the human operator who can do low-level tasks [1].

The Baseboard Management Controller (BMC) is one key source of operational data of current servers. BMC is an ultra-low-cost microcontroller embedded in a server hardware which constantly sets system health parameters (temperature, voltage, fan speed, power consumption, hardware events, and so on). Such telemetry information gives a detailed perspective of the behaviour of servers through time and is capable of detecting abnormal behaviour which could lead to failures. As telemetry data becomes more available, researchers began to investigate how such data may be analyzed with the help of machine learning and analytics methods and predict failure before it occurs.

Software Engineer

Predictive failure detection is especially significant in datacentres based on AI that operate with computationally intensive workloads and need browsing. Unplanned malfunctions of hardware may cause job losses, wastage of resources as well as huge financial loss.

II. RELATED WORKS

Recent research has indicated that predictive analytics would be useful in minimizing the downtime, in that even failures that are likely to take place in advance could be detected. Through evaluation of massive amounts of system logs and telemetry records, machine learning models are able to descend upon patterns that specify degradation or uncharacteristic behaviour within servers. Indicatively, research that is founded on massive collection of public datacentres has shown that predictive models may effectively approximate whether or not a node will failure within a future period by examining previous data regarding operational details [1]. With the help of such models, it is possible to take proactive measures, like workload migration, replacing the hardware, and maintaining the system to occur before the failure really occurs. This will transform the datacentre management to reflex response to failures rather than proactive failure mitigation.

The reason why predictive monitoring is considered an important motivation is the present and increasing size and energy usage of datacentres. The increase of the cloud services and the AI applications means that datacentres use a lot of electricity and need to utilize the existing resources efficiently. Research on energy efficient computing systems insists that datacentres should be developed in a comprehensive realization of both the IT and non-IT aspects including cooling mechanism, power delivery, and computing tasks [2]. Anticipatory maintenance of hardware can lead to energy efficiency as it makes resources planning and minimization of unwarranted system idle hours possible.

Telemetry and observability models are also essential towards gaining insight into datacentre behaviour. The contemporary monitoring strategies retrieve measures and recordings, and tracks of numerous layers of the system to yield an overview of performance in the system. Observability systems allow engineers to study the data about operation, diagnose the problems in the system and identify the anomalies prior to their effects on users [3]. Network telemetry systems have been created as well to gather rich network performance data of the individual datacentre networks in terms of latencies, throughput and routing behaviour [4]. These monitoring systems are the basis of developing predictive analytics models to study system behaviour over time.

The joint solutions with the telemetry data collection and the machine learning analogy can offer the potential solution to the reliability enhancement in the large AI datacentres. Abnormal patterns can be identified by comparing BMC telemetry records and server failures can be anticipated in good time. These predictive strategies facilitate the management of datacentres in an automated way, cut down the system downtime, and enhance the efficiency of the operations. The study of predictive failure identification through telemetry analytics has been a significant field of advancing the resilience and availability of current AI datacentre infrastructures.

A. Predictive Failure Detection in Datacentres

Large datacentres have turned out to be a significant area of study in predicting failure because of the continued growth of computing infrastructures. A single of the first big-scale research studies investigated the data of one of the clusters consisting of over 12,000 machines that had millions of operation incidents. Through machine learning, researchers were able to come up with predictive models of whether a server would crash in the next 24 hours. In the paper, ensemble learning was employed to process machine state features where machine operational data was used as the input and multiple Random Forest classifier enabled as the output. The findings indicated that in order to identify possible malfunction, predictive models could be of considerable accuracy to enable the systems to perform proactive measures like migration of workloads or redistribution of the resources.

The other key model referred to as DC-Prophet is used to predict critical failures of servers in industrial datacentres. This model employed two-step learning methodology of machine learning that integrated One-Class Support Vector Learning with random forest classifiers. The method measured traces of large datacentres that were in excess of one hundred million events comprised of more than twelve thousand machines. The experiment results revealed that the model had very high prediction accuracy proving the fact that machine learning can be effectively used to predict server failures in large datacentres [5]. These works point to the possible benefit of predictive analytics in the enhancement of datacentre reliability.

B. Telemetry and Observability for Failure Analysis

Modern datacentres have generated telemetry data consisting of mass servers and network equipment as well as applications. Resource usage metrics, event logs and network performance indicators are some of the information that the telemetry systems would have accumulated. This type of information comes in handy during the analysis of the behaviour of the system and help the engineers to spot anomalies. Observability frameworks extend past in-traditional monitoring, also incorporating metrics, logs and distributed traces into them, to induce a global perspective of system operation. With these architectures, predictive analytics is made possible, whereby unstructured telemetry data are transformed into meaningful signals, which can be used to identify new failures and performance problems.

Telemetry is also used to trace the network behaviour in addition to failing datacentre networks. One example is the proposed network telemetry frameworks that collect real time data of the network devices and provide the entire visibility of the network performance within the datacentre. Such systems allow operators to view latency, throughput among other indicators of the network effectively even as they allow scalability to large infrastructures. The availability of such telemetry is then a good foundation to make predictive models to estimate the health of the system and predict failures.

C. Machine Learning Approaches for Fault Detection

Machine learning is one of the applications that have been used most frequently as predictive failure detection in computing systems. The majority of algorithms of classification that have been explored comprise of logistic regression, random forests, gradient boosted trees and also neural networks to recognize failures in the complex systems. It has been seen that there have been serious indications that the strategies grounded in data are capable of identifying trends using massive information acquired by sensors and records of operations [6]. The predictive analytics platforms have also been developed in the industrial context to monitor the behaviour of devices and detect the probability and magnitude of failure occurring on the basis of probabilistic models and machine learning approaches [7].

Machine learning models have been used to identify faulty and anomalous states in live streams of data in high-performance computing (HPC) environments. The techniques can help systems to detect flaws as they occur and undertake corrective actions before they accelerate to the system failures [8]. The statistical analysis and the system logs founded on the basis of abnormal detection have also been proposed to identify the abnormal behaviour of the computing nodes. These approaches demonstrate that the data-driven analytics may be vital in enhancing the quality of failure detection and managing and repairing the proper operation in big computing systems [9].

D. Research Gap

Though the literature has been seen to explore the predictive failure detection on the basis of the system logs and operational datasets extensively, but little research had been done on the basis of the Baseboard Management Controller (BMC) telemetry data analysis as predictive analytics in the AI datacentres. Most of the current studies are based on the large-scale system logs, network telemetry or even the application-level monitoring data whereas the BMC level telemetry have not been fully exploited in predicting the modelling. Moreover, there are numerous studies on legacy cloud or HPC environments as opposed to AI-oriented datacentres with extremely intensive collections of GPU workloads and dynamic resource consumption. Lightweight analytics systems able to consume ongoing telemetry streams in real time do also need to be in place so that they can be used to make proactive maintenance decisions. More studies are necessary in order to come up with predictive failure detection techniques that specifically use BMC telemetry analytics to enhance the reliability and resilience of current AI datacentre infrastructures.

An autonomic management method is data-driven to exploit telemetry datasets of large datacentres and ensemble models of the random forest to evaluate the possibility of nodes failing in a 24-hour time frame and then migrate workload in advance and increase system reliability [10]. System behaviour in machine learning model like recurrent neural networks could be learnt based on the integrated telemetry information like resource usage and system calls to analyse future resource usage and early signs of anomalies [11].

To perform better failure prediction in storage systems, the anomalous events in telemetry data can be extracted and use the attention-based recurrent neural networks to learn the key patterns that can signal the impending failure on the time-series data [12]. A domain-adversarial Lrist long-lasting test (DA LLIST) frameworks have been introduced to enhance predictive upkeep by empowering predictive failure

even when reportable data regarding the training and operational conditions of the system are acquired in different conditions [13]. Predictive maintenance procedures adopt the data mining and clustering to identify early machine fault, and assist in condition-based maintenance measures, which lowers the downtime of the system [14].

Mission-critical database platforms can be predicted to experience performance issues and failure due to telemetry-rich observability frameworks and machine learning models that include the Random Forest, SVM, and LSTM models [15]. Deep learning models with federated learning, namely CNN-LSTM, are capable of detecting abnormalities in sensor systems spread over multiple network devices maintaining the privacy of information as well as minimizing the communication costs [16].

A combination of machine learned dynamic system models and statistical observers and decision trees can be used to detect sensor faults through abnormal deviation in real-time measurements [17]. Federated learning Stacked LSTM models have been used in IoT settings to identify anomalies in the sensor data produced by numerous devices and minimising delays in the centralised computation [18]. Smart maintenance systems use machine and sensor data as predictive analytics to complete maintenance activities beforehand and minimize the rate of unforeseen equipment downtime in industrial setups [19].

III. PROPOSED FRAMEWORK

A. Research Design and Data Collection

This paper assumes an experimental research design that is based on data and aims at creating a predictive failure detection model in AI datacentres based on Baseboard Management Controller (BMC) telemetry data. The primary aim of the methodology is to gather the operational telemetry of the servers, analyze the trends in the data and create machine learning models capable of forecasting potential hardware failures even before they can happen. AI datacentres create high quantities of telemetry information of server hardware elements, including processors, GPUs, memory modules, cooling systems, and power units. These parameters are constantly recorded by BMCS sensors and they are being stored in form of time series operational data. These parameters can be such parameters as CPU temperature, GPU temperature, power usage, voltage regulation, fan speed, system usage, and hardware event loggers. Such measures give valuable data concerning the health of servers and may be used for the identification of some rather abnormal patterns, which may suggest failures.

In this study, the telemetry data is supposed to be gathered within a group of servers that are working within an AI datacentre environment. The servers have numerous sensors that are attached to the BMC system. The data of telemetry is assembled in select periodic intervals like in 30 seconds or a minute. All the records of the data set consist of a time-stamp, server number, and sensor values as well as system health functionalities. The detected dataset is thus a multivariate time-series dataset of the behaviour of servers through time. These events that failed and are seen in the system logs are trained as ground truth labels in predictive models. Each data window is assigned a binary classification value of whether a server has failed during a specified prediction window (e.g. 24 hours).

Statistical characteristics of each variable are initially investigated in order to gain knowledge on the behaviour of the telemetry data. The mean, variance, and Standard deviation are computed to comprehend the changes that sensor reading displays through time. A telemetry variable can be taken by simply taking the average value of the variable:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Variance is used to measure the variability of sensor readings and is given thereof as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{2}$$

Standard deviation gives an idea about the range of variation of the sensor values around the mean and is stated as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} \tag{3}$$

These statistical indicators are useful when there are abnormal operating conditions in the form of the unstable behaviour of the system. The anticipated result is the basis of predictive analytics because of the gathered telemetry data in this experiment.

B. Data Preprocessing and Feature Engineering

Raw telemetry data gathered by datacentre systems typically has noises, lack of values and uneven timing

of measurements. It is a valuable process that involves prior to the training predictive models, the data must undergo preprocessing. Data cleaning is the first stage of preprocessing, during which complete sensors measurements and damaged records are processed. The interpolation methods can be used to fill in missing values or mean value of sensor readings could be used to replace them. The Telemetry variables after cleaning are standardized to ensure that all the features are in the same numerical range. This enhances the performance of machine learning models that will ensure fewer dominating variables of larger numeric scales will dominate the learning process.

One of the well-known normalizations that are adopted in this study is Min-Max normalization where each feature value is converted into a number between 0 to 1. The formula of normalization is provided as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

The other process that is relevant in preprocessing is feature extraction. Rather than working with raw values of telemetry, some statistical properties are obtained out of the sliding time windows of telemetry measurements. Such characteristics can be average readings, highest readings, lowest readings, variance, and rate of change of sensor readings. The change rate of rate assists in identifying rapid changes or reduction of the system parameters. It can be calculated as:

$$\Delta x = \frac{x_t - x_{t-1}}{\Delta t} \quad (5)$$

Correlation analysis is also carried out to get insight into the relationship of various telemetry variables. Correlated variables could be highly which means interdependence amongst components of a system. Correlation between two variables is determined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (6)$$

Through feature engineering, the raw telemetry data is condensed into a feature matrix in structured form whereby the rows are time windows and the columns depict derived features describing the behaviour of the system. Such an organized data set is then trained to predictive models.

C. Machine Learning Model for Failure Prediction

Once the telemetry data has been pre-processed the machine learning algorithms are employed in developing the predictive models which would classify, as to whether a server is most likely to go dead in the coming days. Random Forest classification model, Support Vector Machine (SVM) classification model and Long Short-Term Memory (LSTM) Network classification models can be utilized to identify patterns of failures in this study. These algorithms are used to explore the interconnection between telemetry features and the events of failures captured in the systems logs.

Logistic regression is one of the algorithms that are used to undertake predictive maintenance tasks. The logistic regression is an approximation of the likelihood of a failure to happen depending on the features introduced. The model of the logistic regression can be characterized as the following one:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \quad (7)$$

Random Forest is another commonly applied machine learning technique which is an ensemble learning algorithm, that sums up several decision trees. Every decision tree is trained based on various subsets of the data and the ultimate result is the process of integrating results of all the trees. The Random Forest model prediction may be stated as follows:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (8)$$

in which $h_t(x)$ indicates forecast of the t^{th} decision tree and T is the amount of trees in the ensemble.

Failure prediction can also be performed with Support Vector Machines as they can find a decision boundary dividing the normal system behaviour and the failure conditions. The SVM optimization potential is able to be formulated as:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad (9)$$

subject to the constraint:

$$y_i(w \cdot x_i + b) \geq 1 \quad (10)$$

The recurrent neural networks like LSTM will become a possibility in the case of the time-series telemetry data since they will be capable of allowing the modelling of time-dependencies of the sequential information. Latent state change of LSTM cell can be as represented:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \quad (11)$$

Such machine learning models are trained on historical data of telemetry and identify the signals that can give

information about potential hardware degradation or abnormal behaviour of the system.

D. Model Evaluation and Performance Metrics

Once the training of the predictive models is done, their performance should be measured with the help of relevant metrics. The figure of the datasets is broken down into training and testing sets to enable the models to be tested on untested data. The process of evaluation involves the measuring of the prediction of the models in predicting the occurrence of failure.

Accuracy is one of the indicators, which is often utilized in failure prediction research and is expressed as a percentage of the correct predictions by the model. Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

Precision is another vital measure of what is the percentage of failures predicted to be correct. Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

Recall, which is also called sensitivity, is a measure of how the model needs to identify actual failures correctly. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

F1-score is an integrated score that is used to measure the precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

The metrics used to evaluate this performance give a clear picture of the performance of the predictive models in terms of predicting failure. The high recall will guarantee that the majority of the failures are detected at the early stage whereas high precision reduces the false alert cases.

The prediction model trained can be included in a real time datacentre monitoring system. The system is used to constantly gather BMC telemetry data on the servers and process data according to the trained model and create alerts in case of detecting abnormal patterns. The approach outlined in this case, namely, predictive monitoring, enables datacentre operators to take proactive measures, including workload migration, data replacement, or preventive maintenance, prior to failure affecting the AI workload. The proposed methodology will enhance the reliability, efficiency and operational stability of the contemporary AI datacentre infrastructures by employing BMC telemetry analytics and machine learning techniques.

IV. RESULTS

A. Telemetry Data Characteristics and System Behaviour

The initial analysis process after gathering and pre-processing of the BMC telemetry data was to get familiar with the behaviour of the working data produced by the AI datacentre servers. A number of hardware monitoring variables included in the dataset comprised CPU temperature, GPU temperature, system voltage levels, fans speed, memory usage and power consumption. These parameters were documented at a constant period of time which creates time-series telemetry streams on each server. Preliminary study revealed that the normal operation had most servers oscillating within stable ranges whereas during the occurrence of the events of failures abnormal patterns were observed before they took place. As an illustration, servers with failures used to exhibit a slow rise in temperatures and power fluctuation a few hours before the occurrence of the actual failure.

Statistical analysis proved that telemetry signals had observable variations during periods of failures. The standard deviation values of temperature and power metric grew all the more in hours preceding faults in the system. This means that hardware instability may be discovered through evaluating the uncertainty of the telemetry data as opposed to evaluating the absolute values only. The other fact provided by the data was that certain failures were preceded by the abrupt rise of the system metrics like power consumption or temperature of the graphics card. These spikes came in the form of short peaks in the time-series plots and could happen many times, as well as preceding the ultimate failure event. This behaviour argues that predictive models ought to take into account the temporal behaviour as opposed to the values set by individual sensors.

Different telemetry variables were also correlated in order to know relationships among the hardware components. Findings revealed that the correlation and relationship between CPU temperature and fan speed were high since the cooling mechanism will automatically accelerate a fan in case of temperature increase. On the same note, the AI training workloads had high intensity of workload that led to moderate correlation with power consumption and the temperature of the GPUs. Such relationships are useful

in teaching machine learning models to learn system behaviour more. Findings presented by the results of the exploratory analysis show that BMC telemetry data will be valuable in providing meaningful signals that indicate abnormal system behaviour and predict future system failures.

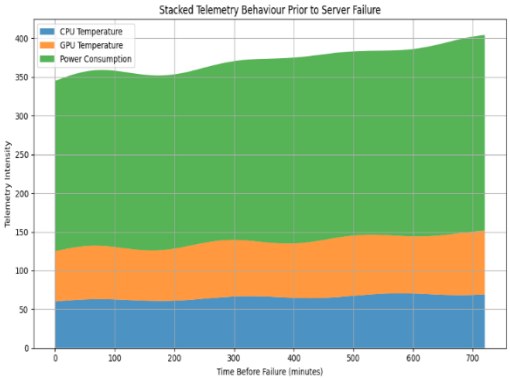

Fig 1: Telemetry data (CPU temperature, GPU temperature, and power consumption) preceding a server failure

B. Predictive Model Performance for Failure Detection

Machine learning models were trained with the telemetry data after extracting the features and preprocessing it to predict the failures of the servers within a specific time interval. It was categorized into training and testing, such that the performance of the models could be tested on the unknown data. The experiment was carried out on three predictive models, namely the Logistic Regression, Random Forest, and Support Vector Machine (SVM). These models have been chosen due to the reason that they have a wide adoption in classification problems and able to analyze structured telemetry features as well.

The models had relationships between telemetry patterns and failure events, which were learned during training. Random Forest model represented the most successful as it can be effective to deploy the nonlinear relationships between the variables of the system. Logistic regression also worked fairly well but could not easily find out the complicated patterns when several sensor variables altered at once. SVM model offered good classifications limits at the expense of parameter tuning to obtain a consistent output.

Accuracy, precision, recall and F1-score were then used to measure the overall performance of the models. Table 1 represents the results. As the table indicates, the most accurate and the most recall rate were recorded by the Random Forest of the tested algorithms. High recall is of particular value in detecting predictive failures since omission of real failure may cause unforeseen system down time.

Table I: Performance comparison of predictive models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	82.4	78.2	74.6	76.3
Support Vector Machine	85.1	80.9	79.5	80.2
Random Forest	89.7	86.3	88.1	87.2

The findings indicate that the telemetry-based failure prediction methods like the Random Forest are more efficient in terms of learning since they are considered to be more efficient in integrating many different decision trees and predicting failures more accurately. The models managed to realize early warning signs several hours before real failures and this enabled the system to create preventive alerts.

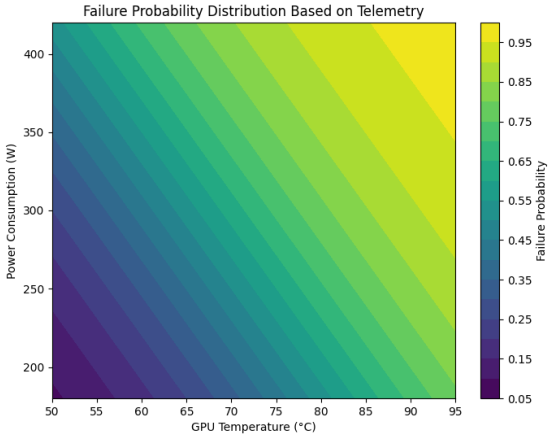

Fig 2: Likelihood of failure of the server due to factors of temperature and power use in telemetric variables

C. Failure Pattern Analysis Using Telemetry Features

The feature importance analysis was utilized to express more about the predictive traits, which were revealed by the machine learning models. The model of random forest gave me the numbers of the significance of each of the features of the telemetry,

which show, to what extent, each of the variables was associated with the ideal of failures. As it was discovered, the temperature of graphics card, the temperature of the CPU, the speed of the power, and fans are some of the most prominent variables. The tendencies of these aspects could be observed even before the realization of failures of the systems.

The temperature of GPUs in most of the points of failure was slowly increasing with time passing and also the speed of the fan was also increasing as the cooling system was setting up the temperatures balancing. But on later instance when our cooling device was not capable of maintaining the temperatures in the system at good levels, there was a hardware wring or malfunction at the system. Equally, the deviant power developments were noticed in two servers that were on the verge of collapsing. These are the kind of variation that were normally induced by workloads or supply of power which fluctuated.

The other observation was also quite apparent, which was that telemetry data have temporal patterns. Without occasion, failure events could hardly have happened. The pre-determination of failure most of the times was a result of several hours of deviant system behaviour. This time was very fluctuating and recurring spikes of telemetry signals. The above trends are suggestive of the fact that the would-be warning signs might be obtained successfully with the aid of the forecasting models when the past telemetry data is referred to during the forging time poisons presence.

Mean scores of importance were received as a measure of the importance of the telemetry features as a measure of the said model. Table 2 has the results.

Table II: Importance of telemetry features in failure prediction

Telemetry Feature	Importance Score
GPU Temperature	0.26
CPU Temperature	0.21
Power Consumption	0.18
Fan Speed	0.14
Memory Utilization	0.11
Voltage Level	0.10

Findings show that the variables based on temperature had the highest contribution to failure prediction and therefore thermal monitoring has significance in AI datacentres. Other factors such as power consumption and fan speed were also considered as the workload intensity and response of the cooling system takes a crucial part.

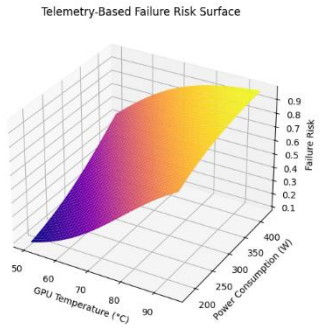

Fig 3: Failure risk intensity across temperature and power dimensions

D. Predictive Monitoring for Datacentre Reliability

The analysis of the findings was to concentrate on the application of predictive model to the datacentre monitoring systems in practice. The trained random forest was pruned to a simulated sequence of running continuous with the input telemetry data. The model will refer to the likelihood of each server failing and alerts it in the event that a probability goes beyond a specified time as new records of the telemetry are updated.

Predictive monitoring system could have been accordingly placed to foresee most of the failure incidences many hours before actual failure of the system. This is an alert mechanism triggered pre-emptively that allows datacentre administrative to take pre-emptive measures that might include support of a work load move, re- boot of a server or equipment or even a probe. It can eliminate cases of service disruption and can execute the stability of operations by possessing corrective actions capable of being implemented at an earlier phase where the system can alleviate the cases of service interruption.

The other benefit of the predictive system is that it minimizes on the manual monitoring. The operators will not have to go to the next stage of interpreting the large amounts of logs and telemetry data manually as they will encounter automatic alerts that the predictive model will introduce. It is a better method of maximizing the efficiency of the way the operation is conducted and enable the engineers to concentrate on the branch level functions within the system.

As it has been proven, the experimental evidence has revealed that the strategy of BMC telemetry analytics with the machine learning models can increase the chances of detecting the possible hardware faults of the AI datacentres substantially. The predictive models had the ability to predict convoluted rules of telemetry, introduce initial warnings patterns not to mention reaching the exact forecasts of the impending failures. This constitutes one of the assistance predictive monitoring infrastructures may assist to the apprehensiveness and capacity of the current datacentre infrastructures.

V. CONCLUSION

This study looked at the application of BMC telemetry analytics in the prediction of hardware failures in AI datacentres. The modern datacentres provide massive amounts of telemetry information, in the form of sensors that track the health of servers and performance of a system. With the use of machine learning methods, one can find out the patterns of abnormality that can be noticed before failures happen on the systems. Some of the telemetry parameters that were studied in this paper include the temperature of the CPU, the temperature of the graphics card and power usage, the rate of fan speed, and the amount of memory that is being used up to identify early warning tissues on the possible hardware instability. Reviewing the preprocessing and extraction of features, a number of predictive models were tested to detect failure. The findings proved that ensemble learning techniques mostly Random Forest, performed the highest prediction, then other algorithms. The testing analysis has proved that telemetry analytics is able to effectively lead to potential failures and produce early warning messages in time before system shutdown process is encountered.

The results demonstrate the significance of applying predictive monitoring systems rather than the conventional reactive monitoring methods. By combining machine learning models and real-time telemetry streams, operators of datacentres will be able to minimize downtime, enhance operational efficiency, and assure consistent AI workloads. Future development should aim at streamlining analytics in real-time, developing deep learning, and increasing the size of telemetry so as to further enhance predictive accuracy and scalability.

REFERENCE

[1] Sîrbu, A., and O. Babaoglu, "Towards data-driven autonomics in data centers," *arXiv preprint arXiv:1505.04935*, 2015.

[2] Raj, V. M., and R. Shriram, "Power management in virtualized datacenter – A survey," *Journal of Network and Computer Applications*, vol. 69, pp. 117–133, 2016.

[3] Parepalli, S., "Data hygiene and batch optimization in enterprise CRM: A framework for scalable, high-quality customer data integration," *Journal of Scientific and Engineering Research*, vol. 3, no. 5, pp. 285–292, 2016.

[4] Lin, Y., Y. Zhou, Z. Liu, K. Liu, Y. Wang, M. Xu, J. Bi, Y. Liu, and J. Wu, "NetView: Towards on-demand network-wide telemetry in the data center," *Computer Networks*, vol. 180, p. 107386, 2020.

[5] Lee, Y., D. Juan, X. Tseng, Y. Chen, and S. Chang, "DC-Prophet: Predicting catastrophic machine failures in datacenters," *arXiv preprint arXiv:1709.06537*, 2017.

[6] Xiao, W., "A probabilistic machine learning approach to detect industrial plant faults," *arXiv preprint arXiv:1603.05770*, 2016.

[7] Majumder, B. P., A. Sengupta, S. Jain, and P. Bhaduri, "Fault detection engine in intelligent predictive analytics platform for DCIM," *arXiv preprint arXiv:1610.04872*, 2016.

[8] Netti, A., Z. Kiziltan, O. Babaoglu, A. Sîrbu, A. Bartolini, and A. Borghesi, "Online fault classification in HPC systems through machine learning," *arXiv preprint arXiv:1810.11208*, 2018.

[9] Ghiasvand, S., and F. M. Ciorba, "Anomaly detection in high performance computers: A vicinity perspective," in *Proc. IEEE Int. Symp. Parallel Distrib. Comput. (ISPDC)*, 2019, pp. 112–120.

[10] Sîrbu, A., and O. Babaoglu, "Towards operator-less data centers through data-driven, predictive, proactive autonomics," *Cluster Computing*, vol. 19, no. 2, pp. 865–878, 2016.

[11] Schmidt, F., M. Niepert, and F. Huici, "Representation learning for resource usage prediction," *arXiv preprint arXiv:1802.00673*, 2018.

[12] Giurgiu, I., and A. Schumann, "Explainable failure predictions with RNN classifiers based on time series data," *arXiv preprint arXiv:1901.08554*, 2019.

[13] De O, D. C. P. R., A. Akcay, Y. Zhang, and U. Kaymak, "Remaining useful lifetime prediction via deep domain adaptation," *arXiv preprint arXiv:1907.07480*, 2019.

[14] Amruthnath, N., and T. Gupta, "Fault diagnosis using clustering: What statistical test to use for hypothesis testing?" *Machine Learning and Applications: An International Journal*, vol. 6, no. 1, pp. 17–33, 2019.

[15] Thota, M. R., "Advancing mission-critical data platforms through predictive observability and autonomous diagnostics," *European Journal of Advances in Engineering and Technology*, vol. 6, no. 1, pp. 162–174, 2019.

[16] Liu, Y., S. Garg, J. Nie, Y. Zhang, Z. Xiong, J. Kang, and M. S. Hossain, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6348–6358, 2020.

[17] M. D. S. B., J. Callaham, J. Jonker, N. Goebel, J. Klemisch, D. McDonald, N. Hicks, J. N. Kutz, S. L. Brunton, and A. Y. Aravkin, "Physics-informed machine learning for sensor fault detection with flight test data," *arXiv preprint arXiv:2006.13380*, 2020.

[18] Sater, R. A., and A. B. Hamza, "A federated learning approach to anomaly detection in smart buildings," *arXiv preprint arXiv:2010.10293*, 2020.

[19] Von Enzberg, S., A. Naskos, I. Metaxa, D. Köchling, and A. Kühn, "Implementation and transfer of predictive analytics for smart maintenance: A case study," *Frontiers in Computer Science*, vol. 2, 2020.