The Journal of Supercomputing (2021) 77:13494–13513
https://doi.org/10.1007/s11227-021-03811-7

Smart predictive maintenance for high-performance computing systems: a literature review

André Luis da Cunha Dantas Lima[1] · Vitor Moraes Aranha[1] · Caio Jordão de Lima Carvalho[1] · Erick Giovani Sperandio Nascimento[1]

Accepted: 12 April 2021 / Published online: 27 April 2021

Abstract
Predictive maintenance is an invaluable tool to preserve the health of mission critical assets while minimizing the operational costs of scheduled intervention. Artificial intelligence techniques have been shown to be effective at treating large volumes of data, such as the ones collected by the sensors typically present in equipment. In this work, we aim to identify and summarize existing publications in the field of predictive maintenance that explore machine learning and deep learning algorithms to improve the performance of failure classification and detection. We show a significant upward trend in the use of deep learning methods of sensor data collected by mission critical assets for early failure detection to assist predictive maintenance schedules. We also identify aspects that require further investigation in future works, regarding exploration of life support systems for supercomputing assets and standardization of performance metrics.

Keywords Predictive maintenance · High-performance computing · HPC · Artificial intelligence · Deep learning

1 Introduction

As cloud services grew in popularity, a great part of data processing formerly performed on desktops or local servers moved toward large data centers, with thousands of interconnected computers. There is a steady growth in the demand for computing nodes of high-performance computing (HPC) environments, especially for computer modeling and simulations. Such tasks require high computational power to obtain results quickly enough to support companies in strategic decision-making.

✉ Erick Giovani Sperandio Nascimento
ericksperandio@gmail.com

[1] Faculdade de Tecnologia SENAI CIMATEC Salvador, SENAI CIMATEC Manufacturing and Technology Integrated Campus, Salvador, BA, Brazil

In such complex large-scale environments, multiple servers are interconnected by network layers and many assets rely on different configuration parameters. It is evident that failures will happen.

According to [15], data centers and cloud-based computing systems are very dynamic in nature, because of the modularity of their components, which allow addition, removal and repair of nodes without interrupting the service. To support system managers, proactive failure management techniques are paramount to characterize behaviors, detect anomalies and foresee failure dynamics, therefore becoming an effective approach to improve the reliability of the system. To preserve the availability of these complex environments, a promptly available redundant infrastructure is required for incidents or emergency situations to prevent the fault and failure of these devices. As an immediate consequence, both energy consumption and operation costs grow, especially when the environment in question is mission critical, composed of redundant powered devices available 24 h a day, 7 days a week.

However, for a supercomputing environment with electronic devices, cooling equipment such as water pumps, heat exchangers, chillers and energy supply units such as power generators, the maintenance management must carry regular performance and health checks with a systematically structured maintenance program. It is possible to say that the maintenance of a mission critical supercomputing environment is analogous to the one of an industrial facility, where the continuous productive cycles requires devices to work reliably and uninterruptedly.

As seen in [29], in industry 4.0, the use of intelligent sensors offers a reliable solution for the real-time monitoring of systems, especially when the data collected are applied to predictive maintenance. This way, maintenance tasks may be planned more efficiently to minimize operation downtime while preserving the health of the entire system. In the context of supercomputing, many factors degrade the performance and health of a system, from literal wear and tears of parts to operational failures from high demand of nodes. Applications may even enter in unwanted states because of incorrect configurations. For this reason the large numbers of sensors available in a high-performance computing system may be used to monitor and infer the health status of both main and support devices. However, manipulating this large volume of data in order to identify problems is a difficult task for system managers and operators [2].

Thus, predictive maintenance becomes an interesting approach to be applied in mission critical environments, in order to predict real working conditions based on historical data. This maintenance can identify possible problems early on when those are still just potential defects, which allows the team to fix them before growing in severity and eventually becoming a failure.

Very large amounts of data and logs are obtained with constant monitoring in a high-performance supercomputing environment. For this reason, it is not possible for a single operator to identify, sort and categorize the entirety of the data to obtain any meaningful information. A human operator, overwhelmed by multiple logs from thousands of sensors is very likely to carelessly discard data that would otherwise be helpful in detecting critical behavior. Events like this may interfere negatively in the prediction of the system status.

Traditional approaches in mission critical environments may rely on alarm thresholds to handle potential failures within a reasonable time frame. However, during the lifetime of a process, critical thresholds may be exceeded several times without necessarily incurring in failures. Also, some failures may evolve too quickly, leaving virtually no response time available for human operators to fix the problem. The ability to capture failures in an initial stage is often preferable, and for this task, automated algorithms perform better than humans.

Prediction of failures using collected historical data of system status becomes then a valuable approach to plan for resource allocation, system reconfiguration and equipment maintenance. This helps minimizing the costs of operation and to maximize service availability.

In this context, predictive maintenance allied to artificial intelligence techniques may offer a solution for the problem of large databases of sensor data. Recent literature features a wide range of algorithms specialized in detection and prediction of patterns over large databases (also known as Big Data). These techniques could be successfully applied to infer and forecast behavior of such mission critical supercomputing devices with much greater performance than a human could obtain. These results may offer support in decision making to guide corrective actions for potential problems, improving system reliability, availability and productivity as well.

In this study, we verify the contributions of the research literature and summarize evidences of artificial intelligence applied to predictive maintenance. Our goal is to identify the newest techniques in published research works that cover predictive maintenance with either machine learning (ML) or deep learning (DL) techniques. Through a systematic literature review, this work offers a useful basis on ML and DL techniques, their performance, their results and it offers support for future investigations in the field of predictive maintenance applied to mission critical supercomputing environments. The next sections of this work are organized as following: Sect. 2 describes the types of maintenance and the advantages of predictive maintenance; Sect. 3 briefly describes artificial intelligence and deep learning; Sect. 4 provides the planning and execution of the systematic literature review; Sect. 5 presents the results and discussions about the findings; finally, Sect. 6, concludes the document, stresses the contributions of this paper and lists possible future works.

2 Types of maintenance

According to [33], there are four main methods of industrial maintenance: corrective, preventive, proactive and predictive.

Corrective maintenance (run-to-failure) is applied only in the circumstance of a failure. This is the most common and the simplest approach; however, it is the least effective, as the costs of intervening and the inactive period of the systems far outweigh the costs of preventive measures [32, 36].

Preventive maintenance, on the other hand, is based on a planned course of action that considers the time iterations of devices in order to prevent failures. However,

this approach is not yet ideal because of unneeded interventions being carried out periodically causing inefficient resource utilization and raise operational costs [32].

According to [36], as these two traditional approaches become less capable of answering the growing demands of efficiency, reliability and safety, more and more intelligent techniques have been receiving attention. As a result of this scenario, predictive maintenance has become another important method to provide early failure detection.

Predictive maintenance is an efficient and promising solution when compared to corrective and preventive maintenance, because it allows the evaluation of the health status of an equipment from its previously collected field registries with the goal of predicting the best moment to intervene in a system before failure happens, thus avoiding system failures and non-planned stops.

In a mission critical environment, predictive maintenance allied to periodic monitoring of the critical asset can reduce uncertainties and downtime while providing a cost reduction for management teams, especially in the following topics:

- Increased useful life of assets by preventing damage.
- Optimized operation process and technical support.
- Increased availability and reliability of the asset.
- Reduced number of corrective and preventive maintenances.

3 Artificial intelligence (AI)

Artificial intelligence goals are focused on the execution of cognitive tasks, especially those which humans are used to execute well, using paradigms/algorithms learned by machines. An AI system is capable of representing, reasoning and learning through experience [16]. Deep learning is a specific sub-field from the machine learning area which handles with stacked learning layers using increasingly significant representational data. Therefore, in the processing and correlation of a great volume of collected data, this technique enables the learning of complex concepts through some simpler ones using many steps, i.e., the human operator can formally specify all the desired knowledge that the machine needs, storing a concept hierarchy [14].

Deep learning methods are becoming one of the most popular topics in diagnosis and prognosis oriented techniques for machinery. Mainly due to their ability for allowing the extraction and automated building of useful information from complex data pre-processing and knowledge [29].

Succinctly, this process consists in a sequence of representational layers of the input data, which are processed by a neural network architecture. The depth of the model can be understood as the number of layers that were used. Usually, the output of a layer is used as an input for the next one and each iteration of the model will help with the adjustment of the parameters for the next iteration, executing feedback process [11].

According to [10], DL offers the following advantages for applications of management and monitoring equipment's health:

– Automatic process large amounts of monitoring data;
– Automatic extraction of useful features from heterogeneous and high-dimensional data;
– Learning of functional and temporal relationships between and within the time series of conditional monitoring signals;
– Transfer of knowledge between different operational conditions and units.

In the recent literature, a large number of publications have adopted deep learning techniques focused on the identification and prediction of failures in industrial devices and machines. The usage of long short-term memory (LSTM) networks stands out for the treatment of time series [7, 35, 37]. These networks have the ability to reproduce and predict data with precision because of their architecture, and they are able to store changes in sequential states, as the ones that usually happen in time series [17].

The use of convolution neural networks for regression and classification problems in the context of system health prognosis has also intensified. These networks are based on convolution operations through the input signal and for that reason they are particularly efficient in the detection of features that are present in sub-regions of the input signal. Being initially popular in the image pattern recognition area, some authors have used these architectures in time series and sensor data as well [40].

Among the advantages of the deep learning techniques in the context of failure detection, we highlight the ability to quickly process large amounts of data quickly and to build relationships on it so as to predict or classify the input signal behavior with precision greater than that of a human operator.

4 Methodology

This study was conducted as a systematic review of the literature based on original directives as mentioned by [19, 23]. In this case, the objectives of this review are: to identify works that offer solutions developed with artificial intelligence algorithms to predict failures in mission critical environments for supercomputing and that use deep learning techniques; to identify research methodologies used in equivalent contexts and their results, in addition to map the essential requirements to better define the machine learning technique to predict multivariate time series applied to computing equipment. The stages of the systematic review of the literature are documented below.

4.1 Research process

First, the following questions were defined for the survey of relevant publications:

Q 1 Are there studies on predictive maintenance applied to mission critical supercomputing environment using deep learning techniques?

Q 2 Are there deep learning algorithms to predict failures on supercomputing environments?
Q 3 Which requirements are essential to better define the correlation technique of the collected data?
Q 4 Which requirements are essential to better define the machine learning technique to predict failures in time series applied to mission critical environments?
Q 5 Can the proposed techniques be used in a supercomputing environment?

After this definition, we began the process of automatic search through the creation of filters in our bibliographic research tools using the different combination of the descriptors or keywords: *high-performance computing*, *predictive maintenance*, *artificial intelligence* and *deep learning*.

4.2 Inclusion and exclusion criteria

Studies available in journals or conferences between 2010 and 2021 and written in English were defined as the inclusion criteria. Publications that did not fit these categories, without a focus on industrial environments, HPC, cluster, or datacenter, as well as, reviews, bibliographies, editorials, and reports, were not considered in this survey.

The quality evaluation for the inclusion of the works was also based on the following questions:

Q 1 Was the deep learning technique proposed for correlating the collected data tested in a high-performance computing environment, clusters or data centers?
Q 2 Was the study evaluation/method presented?
Q 3 Were the form and the frequency of the data presented? Are they available on the same time basis?
Q 4 Are the collected data labeled with alarm/failure events?
Q 5 Was the deep learning technique used during the process of training of the neural network?
Q 6 Are the conclusions of the study made clear through the use of statistical metrics?

5 Results and discussion

The research was carried out on 2020 July, seventh, and reevaluated in 2021 April, fifth using the descriptors defined in Sect. 4.1 and applying the criteria for inclusion, exclusion, and quality, defined in Sect 4.2. At the end of the search, 32 papers were selected because of their proximity to the context of the questions that guide this review.

Figure 1 presents the distribution of publications by database; of those, 47% (15 papers) were published in IEEE Xplore, 22% (seven papers) in Science Direct and 16% (four papers) in Springer, the remaining 15% (five papers) were distributed as

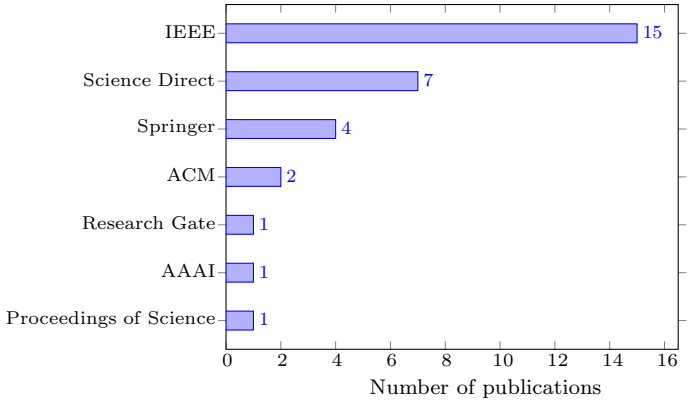

Fig. 1 Number of publications selected by research base.

two papers from ACM, and 1 paper from each of the respective publishers Research Gate, AA AI and Proceedings of Science.

Figure 2 describes the period when the papers were published between 2011 and 2020. Notice how there is a growing interest in the field of predictive maintenance with artificial intelligence from 2017, the average number of publications between 2011 and 2016 was one and one tenth, and between 2017 and 2019 it rose to eight.

According to [5], the growing demands for data generation and storage in industrial equipment and the recent advances in machine learning algorithms seem to correlate with the number of publications in this field.

In Fig. 3, we present the number of publications and their distribution by conference from multiple fields, such as reliability engineering, intelligent systems, Big Data, artificial intelligence, computer networks, information technology,

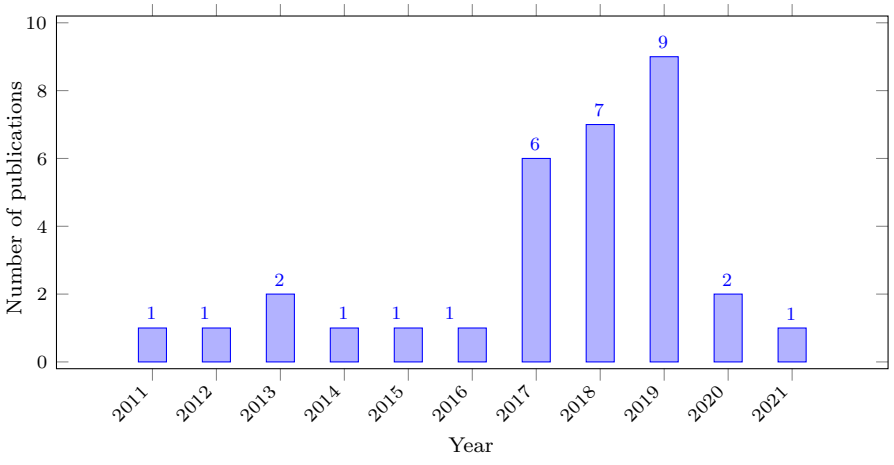

Fig. 2 Number of publications by year

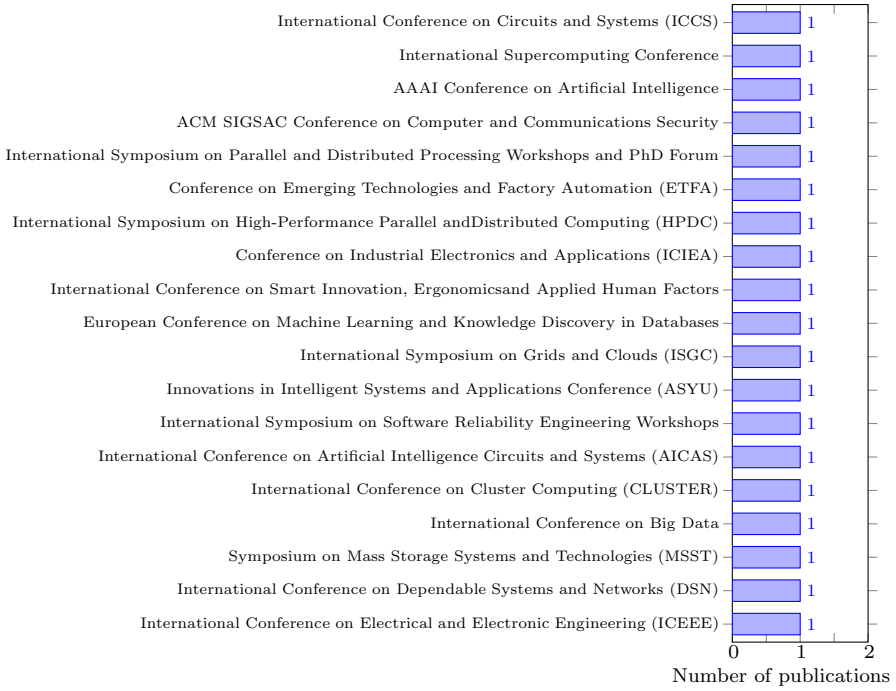

Fig. 3 Number of publications by conference

high-performance computing, cloud computing, parallel computing, software engineering, electronics, electrical engineering, distributed computing and automation.

Figure 4 presents the number of publications by specialized journal with a focus on Transactions on Industrial Informatics and Reliability Engineering & System Safety, both with two publications. The former is held by IEEE directed toward industrial automation, and the latter is published by Elsevier in association with the European Safety and Reliability Association and the Safety Engineering and Risk Analysis Division, directed toward the development and application of methods for increased safety and reliability of complex technological systems.

Thus, it is possible to observe the growing interest in artificial intelligence methods applied to failure prediction in critical equipment where uninterrupted availability is required.

In the study carried out in [18], a two-stage method for failure prediction was executed in a semiconductor plant. The authors used a common log based data collection. A backpropagation neural network was then trained on the data. In the following step, the authors employed a genetic algorithm (GA) in circumstances where backpropagation was unable to escape from local minima. The authors showed that the evolutive approach performs better than traditional backpropagation techniques at finding global minimum zones.

Genetic algorithm is used in [25] as well. In this work, a neural network architecture is developed with hyper-parameters optimized with the help of a genetic

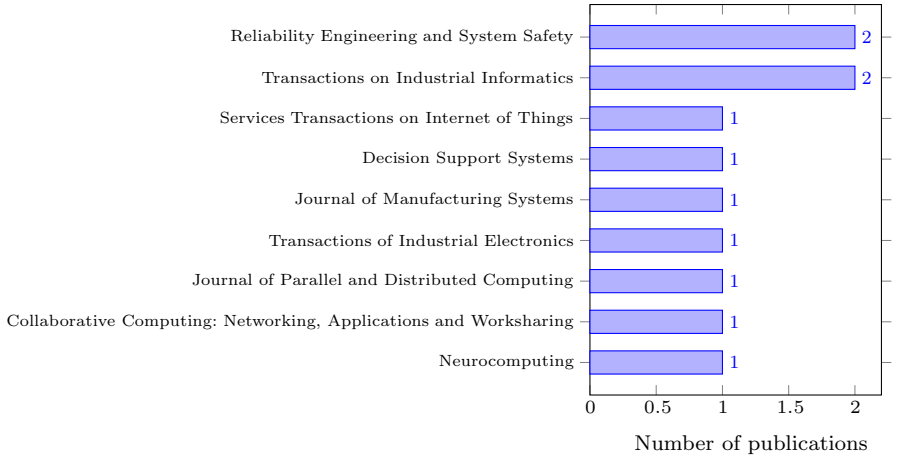

Fig. 4 Number of publications by journal.

algorithm aiming at a data regression for an aircraft rotor dataset. The authors report two main contributions from this approach in the context of predictive maintenance: the developed model helps the migration from current condition-based preventive maintenance to predictive maintenance; And the overall accuracy of the GA optimized network achieved a value of 95%.

A method for predictive maintenance using multiple classifiers is adopted in [33] and tested in a benchmark dataset of a semiconductor plant. The technique was able to handle multidimensional data for multiple prediction horizons. The algorithm was tested together with two popular methods in the literature, support vector machines (SVM) and k-nearest neighbors (kNN). Both methods provided an accuracy of approximately 98.5%, but the authors suggest that even better results could be obtained with the use of relevance vector machines to treat the high dimensionality of the used data.

The studies presented in [21] and [39] used recurrent networks based on low short-term memory (LSTM) to train data and improve the performance of predictions.

Study [21] proposed a compromise solution between model interpretativity and prediction of the useful life of a machine component using variational Bayesian inferences. The technique was tested against the turbofan dataset of aircraft rotors. The result achieved a superior performance to traditional machine learning approaches, especially when the Bayesian inferences were applied together with an LSTM network, achieving an improvement of 37.12% in error scores when compared to random forest. This method also achieved reasonable degrees of interpretativity, which allowed for empirical tests of the model, the authors observe that dropout models offered no particular advantage when used with this approach, and future investigations can be carried out on how to benefit from this type of regularization in this circumstance.

To improve the precision and the performance of its prediction model, study [39] proposed a technique for feature extraction for time-lagged correlations. This approach has been shown to be capable of reducing the training time of a network while preserving the precision of predicted values. The model was tested against three machine learning models (rule-based, d-LSTM and dp-LSTM); however, the best results were achieved by the vanilla LSTM network with an average recall of 85.6% and precision of 78%.

A deep learning 2D convolutional-LSTM (ConvLSTM) autoencoder model was proposed by [9] to predict the velocity of a smart manufacturing machine in an intelligent manufactory process. Sensor data were gathered from an industrial plant of metal packing products. The 2D ConvLSTM autoencoder model surpassed the benchmark statistic models (persistence model and autoregressive integrated moving average), as well as the models RSNet, deep LSTM encoder–decoder and CNN-LSTM encoder–decoder. When compared to a CNN-LSTM model, it obtained 26% better scores on average for the Root-Mean-Squared-Error(RMSE), Mean-Absolute-Error(MAE), and Symmetric Mean Absolute Percentage Error(sMAPE) metrics. To this end, an input sequence was restructured and univariate projection was applied with a multistep time series. The use of the autoencoder technique for dimensionality reduction also improved the time required to train the model in approximately 23% compared to the hybrid CNN-LSTM.

A comparative study on the performance of ten machine leaning algorithms was carried out by [26] with the objective of predicting the remaining useful life (RUL) of an aircraft motor from the turbofan datasets using a supervised learning model. Each dataset contains data from 100 to 250 motors with 21 different sensors each. The results were compared using the RMSE between the actual and the predicted RUL values. The random forest (RF) algorithm presented the best results for average RMSE of the datasets (29.73) capturing the variance from many input variables and at the same time allowing a great number of observations to join the prediction. The resulting algorithms that closely followed the lead of RF were gradient boost (32.97), ada boost (33.18) and decision tree (34.17).

Study [30], used machine learning-based models on graphics processing units (GPU) input data for software error prediction in a large-scale high-performance computing center. The large datasets consisted of relationships between power, temperature, type of workload, and single bit error (SBE) distributions. A two-stage prediction method was provided, in which in the first stage the samples are checked to identify in which nodes the SBE occur. The samples are passed to a second stage in the positive case; this way the size of the actual failure data is reduced. Compared against this approach, many machine learning models were tested, including logistic regression (LR), gradient boost decision tree (GBDT), support vector machines (SVM) and a neural network (NN). The LR technique required the least amount of time to train over the data (4.81 s) compared to GBDT (49.53 s); however, F1 score of GBDT (0.81) was superior to LR's (0.67), achieving a better prediction performance.

The *DESH* framework [7] presents a solution for failure prediction in computing nodes on a supercomputing environment. Based on LSTM networks and using system logs as data, the technique is implemented in three stages: first, the

authors perform process of the log texts to provide work patterns for recognition, and then the word strings are then grouped and sent to an LSTM classifier network. The next stage consists in predicting the events based on the failure patterns obtained from word strings; here, field annotations are added to the preprocessed word strings and sent to a new LSTM. Finally, the data are validated in a subset of failure strings reserved for this stage. The results provide accuracy of 83% and recall of 85% up to 3 min of lead time.

The study carried out by [28] presents the development of classification schemes for data mining to predict compute node failures in a high-performance computing system by collecting and analyzing system logs. The methodology combines system usage and system failure logs of a computing node. Later, a predictive model is applied to predict whether the failure will occur within 1 hour from the current time in the HPC system. Finally, the performance of a binary classification is evaluated using evolution metrics. The experiments were tested against 11 classification methods, and among them random forests presented the best metrics: 73.9% precision, 81.3% recall, 77.4% F-measure and 0.95 ROC area.

The study in [29] provided a framework for predictive maintenance based on sensor data. The authors carry out a system prognosis by using a recurrent neural network oriented to the needs of the system operator. For the probability distribution of the remaining useful life, the first LSTM layer classifies the data intervals. The authors argue that this method offers advantages over other approaches that exclusively perform regression, as these techniques degrade with longer prediction horizons. The confusion matrix for these classification metrics is presented in the study and average accuracy by label is over 90%.

Convolutional networks are applied in [22] to obtain system prognosis and health management of systems using a data oriented approach. The network architecture is composed of 4 convolutional layers to identify data features. A fifth convolutional layer is applied with a single filter to combine the feature map. Finally, the output of these layers is forwarded to a single multilayer perceptron for the final result. The C-MAPSS dataset is used to compute the benchmark for the proposed technique, which achieved average RMSE of 12.61 and standard deviation of 0.19.

The concept of convolutional networks for failure diagnosis in a pump system is also carried out in work [40]. A previous classification of equipment work profiles is carried out with sensor data and dynamometers. The dataset consists of real data obtained from Chinese oil wells provided especially to this research. Next, the data are tested against two CNN architectures developed by the authors. The first one uses 2D convolutions treating the sensor data as images, and the second considers 1D convolutions over time series data. The results are tested against KNN and random forest. The metrics for performance evaluation are precision, recall and F1-score. The authors point to the superior performance of the 2D Convolution network, in relation to the other methods.

A model based on automatized logs for IT systems is adopted in [38]. To group and parse the logs, traditional text mining techniques are used (topic modeling, bag of words). Next, clustering and tokenization (log clustering tree) are carried out in the pattern recognition step. The final step of failure prediction is carried out by an

LSTM network where each input string generated a failure probability for the next 70 min. The model obtained recall of 90% against the benchmark with test datasets.

The model proposed by [13] offers a study of different machine learning models such as multilayer perceptron, XGBoost, KNN and random forest, to process data belonging to Storage Resource Managers from the Italian Research Center (IFNF-CNAF). The authors point to a few challenges during the study; for example, constant and frequent updates of this large database, data organization is also severely lacking. Feature filtering was performed to predict system health with a frequency of 15 min. The authors do not specify the performance metrics for each method.

The experiment developed in [1] presents a model based on LSTM networks, capable of verifying the state of each component in a high data processing machine. This verification evaluates each component separately in order to obtain an individual alert for any component requiring a replacement. The model was trained and tested on an open-source engine dataset provided by NASA. Accuracy of 85% was obtained with the model, thus confirming the expectations of recurrent networks for this use case.

Work [24] points to how features that represent failures tend to weaken with perturbations from noisy data. For this reason, the authors observed the need to identify and filter these external dynamic properties affecting the data. To this end, a neural model based on sensor signal processing is employed together with a dynamic system. The neural model consisted in a set of autoencoders organized in a processing stack combined with a backpropagation neural network. With this approach, the authors obtained the rate at which the equipment health degraded. The best accuracy performance was 98.1% with a set of 5 autoencoders.

With the goal of solving imbalanced classification during the training process caused by large amounts of data generated by machinery and overcoming challenges such as reducing unwanted noise from oversampling, [41] proposes a new algorithm for failure prediction using generative adversarial networks (GAN). The method was developed in three modules; the first one used infoGAN to generate synthetic samples for failure and non-failure status, the second module trained a network for inference by sharing weights of the first layer with the GAN discriminator, and lastly the third module trained a second GAN to reinforce the consistency of the first module and the data labels generated by inference. This algorithm was experimented with industrial datasets used as benchmarks for predictive maintenance: 1 dataset belonging to an air pressure system for truck braking systems and 4 TURBOFAN datasets NASA's CMAPSS aircraft propellers. The GAN for failure prediction was compared against 4 classifiers, deep neural network, support vector machine, random forest and Decision Tree in 4 different sampling configurations (undersampling, weighted loss, SMOTE oversampling and ADASYN oversampling). The GAN for failure prediction obtained better results than all the machine learning techniques tested against for the AUC (area under curve) and F1 score metrics.

Few studies in the field of computer vision aimed at predictive maintenance provide good accuracy results when the data are evaluated as images, as done in [4]. In this experiment, it was possible to build a model based on convolutional networks capable of analyzing the axis of rotating machines through the rotor orbit shape. This model was capable of classifying the state of rotors and predicting their useful

life. An accuracy of 99.9% was obtained, with early error detection to minimize maintenance costs and predicting accidents.

Mission critical supercomputing environments are directly related to an infrastructure of uninterruptedly powered equipment to reduce the risks of service unavailability and by consequence to reduce the operation costs [10]. Papers published recently handle such environments, aiming to devise solutions for automated pattern detection and failure prediction to support decision-making and an effective system management [23].

Anomaly detection in high-performance computers brings significative advantages for system administrators, mainly because these systems tend to be very large, with many integrated components, and prone to unwanted behaviors, failure conditions and faults. Thus, the use of failure detection mechanisms as early as possible reduces the costs of corrective measures and service interruptions of HPC systems [12]. For system administrators, system logs help understanding the state of the system and significant events making it easier to debug the causes. For this reason, system logs are excellent sources of data to perform online monitoring and anomaly detection [8].

The work presented in [2] developed a solution for anomaly detection in a supercomputer using semi-supervised learning. Thus, by using machine learning techniques with an autoencoder, the algorithm is able to identify the normal behavior of computing nodes, which minimizes the training reconstruction error for anomaly detection. After the first training stage, the autoencoder receives new unseen data to evaluate the error of its inference phase. In further experiments, historical data from a real supercomputer was tested. The results show that this approach obtained an anomaly detection rate between 88 and 96%. A similar autoencoder approach was applied in a supercomputing production environment using error injection tests and frequency configuration, by [3]. The authors obtained detection rates between 87 and 98%.

To predict failures in large-scale storage systems in data center environments, [42] used a fully connected neural network with backpropagation, configured with three layers with 19, 30 and 1 neurons, respectively. This solution obtained detection rate of up to 95% against standard SMART tools for self-monitoring and failure analysis, typically these storage management tools are only able to obtain rates between 50 and 60%.

A case study in Google clusters was developed by [6] to evaluate the use of neural networks for the failure prediction of jobs. The authors used a recurrent neural network architecture to process the data by job. The experiments show a true positive rate of 84% and a false positive rate of 20%. Estimations suggest that between 6% and 10% of computing resources were preserved with the proposed approach.

A machine learning framework for divergence diagnosis of HPC environments was presented in [34]. Resource utilization data are processed by random forest classifiers and evaluated according to the F1-score metric in two different environments, obtaining the minimum value of 0.97. A similar method for performance anomaly and variation detection in apps is used in [20]. Descriptive statistics and supervised machine learning methods are used to create a prediction model from computing nodes monitored data. The random forest classifier is applied over the data with two

classes: normal and critical. Results for a 30 min window frame show a precision of 98% and a recall of 91%. However, during the validation process in a production environment this method presented a high number of false positives, achieving up to 79% failure detection rates. The authors do not recommend relying on this method only for anomaly detection because of the low precision scores, and for this reason additional diagnosis are required.

The DeepLog framework developed by [8] uses LSTM networks for online anomaly detection in system logs. The training process uses only log entries considered normal by the system, and new entries may be streamed using a mechanism of user feedback, especially if the detection is a false positive, that is, if a normal log entry is incorrectly classified as an anomaly. This way, DeepLog uses this reinforced method to dynamically adjust weights online, and adapt to new execution patterns. The method presented results of 99% precision for anomaly detection.

In the work published by [27], the authors carry out experiments for failure prediction of a virtualized hardware stack for cloud computing. Data from a series of system components were collected for a total of five years between 2001 and 2006. The failure events labeled in the final time-series presented five possible origins, software error, hardware error, human mistake, network error and undefined. The data were submitted tested against a set of machine learning algorithms, random forests, linear discriminant analysis, support vector machine, K-nearest-neighbor, and classification and regression trees. Conclusive results points toward support vector machines as being the superior alternative in this case, with the highest accuracy of 91% and the lowest RMSE of 0.1718.

The technique presented in [31] applies deep learning methods for failure diagnosis and classification over two publicly available databases of industrial rotating machinery. The method, built upon convolution neural networks, was able to classify the rotating patterns of the equipment and advise operators when maintenance was needed. Sensor data from each equipment were converted to the frequency domain with the fast Fourier transform technique, and inputted to the deep learning model. When tested against the public datasets, MaFaulDa and CWRU, the model presented accuracy of 99.58% and 97.3%, respectively.

Predictive maintenance allows minimizing the maintenance costs of equipment, maximizes operation time and preserves system integrity by reducing the risks of failure, thus allowing preventive measures to avoid asset loss.

Considering the results of machine learning models for event prediction in time series and the continuous monitoring of mission critical equipment, the use of neural networks and deep learning provides a promising path for this application domain. [10, 22]. However, for the successful application of these models it is important to overcome the implementation challenges in these environments that arise from the set of equipment involved in the system. These challenges also require alignment among stakeholders to sponsor the collaboration for the development of adequate business models in favor of all involved partners [23].

Error reports and sensor data collected by field teams or tools are frequently enough to predict the behavior of mission critical equipment. This way, the automation and standardization of the data collection as a preparatory stage for data

treatment and quality becomes important to potentialize deep learning techniques, since these methods have shown promising results in prognosis and health management of industrial equipment [5, 10, 23].

Table 1 presents the publications that employed machine learning techniques with the goal of predictive maintenance of equipment.

Finally, we have observed a growing interest in the previous years of research works aimed toward predictive maintenance in many fields using artificial intelligence, in special with machine learning and deep learning methods. It is possible to verify that LSTM, CNN and hybrid (CNN+LSTM) architectures have been largely applied because of their positive results in the prediction of time series. The increase in computing power of commodity hardware is to be considered one of the reasons behind this growth, since robust AI techniques historically require large datasets (Big Data) and computing power to be executed effectively.

6 Final remarks

This work shows an in depth literature review on deep learning techniques allied to failure prediction, with promising results performing better than previous human based solutions. We have also identified the high popularity of convolutional networks and recurrent networks in the context of failure prediction and classification.

Moreover, in the context of predictive maintenance, we have observed a growing trend in the use of machine learning and deep learning techniques in industrial assets. The introduction of Industry 4.0 and the development of modern equipment with sensors provide a viable way of integrating data storage and processing with computing clusters. For this reason, it becomes feasible to project and validate a predictive maintenance strategy for mission critical HPC systems.

The research questions described in the planning protocol of this paper allowed the mapping of the main publications in the field of predictive maintenance with artificial intelligence in supercomputing systems. We have verified that this is still an environment with open problems and very few publications, and many of them solely focus on log data generated by these computing systems to detect localized problems such as node failure, for example. To better devise an efficient predictive maintenance program, besides log data, it is important to consider sensor data from support equipment (such as power and cooling devices) as well. The inclusion of historical data of preventive and corrective maintenance is another valuable piece of information that once integrated in a sensor data processing pipeline, could help diagnose early failures; however, none of the papers found in this review followed this approach. In addition, we identified a lack of publications addressing the remaining useful life of HPC systems, thus characterizing another gap in this field.

Finally, in a mission critical environment there are still aspects that require further investigation, especially regarding the accuracy of results. A human operator supported by an automated anomaly detection system must be aware of prediction and classification parameters such as precision and recall to minimize the

Table 1 Summary of selected publications that used machine learning techniques for predictive maintenance. Results are: precision/recall/accuracy. RD (real data), SD (synthetic data)

References	Method	Results	Purpose	Equipment	Data source	Data type[1]
Bin hu et al. [18]	Backpropagation + Genetic algorithm	81,3/–/–	Failure prediction	Semiconductor factory	System logs and sensors	RD
Susto et al. [33]	SVM, KNN	–/–/98,52 –/–/98,51	Failure prediction	Semiconductor factory	System logs	RD
Kraus et al. [21]	Bayesian inference + LSTM	–/–/–	Equipment remaining useful life	Turbofan engines	Sensors	SD
Zhang et al. [37]	LSTM	78/85,6/–	Time-lagged correlation	Power Plant	Sensors	RD
Zhang et al. [38]	LSTM	–/–/90	Failure prediction	Clusters	System logs	RD
Susto et al. [32]	Radial basis function (RBF) SVM	–/–/–	Equipment remaining useful life	Semiconductor factory	System logs	RD
Yurek et al. [36]	Decision forest regression, Bayesian linear regression	–/–/–	Equipment remaining useful life	Turbofan engines	Sensors	SD
Essien [9]	CNN + LSTM + Autoencoder	–/–/–	Smart factoring	Metal packaging machine	Sensors	RD
Mathew et al. [26]	Random forest, gradient boost, ada boost, decision tree	–/–/–	Equipment remaining useful life	Turbofan engines	Sensors	SD
Zhang et al. [37]	Bi-directional LSTM	–/–/–	Equipment remaining useful life	Turbofan engines	Sensors	SD
Wu et al. [35]	LSTM	–/–/–	Equipment remaining useful life	Turbofan engines	Sensors	SD
Das et al. [7]	LSTM	–/85/83	Failure prediction	HPC	System logs	RD
Nguyen, Medjaher [29]	LSTM	–/–/90	Equipment remaining useful life	Turbofan engines	Sensors	SD
Li et al. [22]	CNN + MLP	–/–/–	prognosis and system health	Turbofan engines	Sensors	SD
Nie et al. [30]	Gradient reinforcement decision	–/–/–	Software errors	GPU installed in HPC	System logs and sensors	RD
Giommi et al. [13]	MLP, XGBoost and random forest	–/–/–	Anomaly detection	HPC	System logs	RD
Aydin, Guldam-lasioglu [1]	LSTM	–/–/85	Anomaly detection	Engine dataset	Sensors	SD
Luo, et al. [24]	Autoencoder + BPNN	–/–/98,1	Failure prediction	Rotating machines	Sensors	RD

Table 1 (continued)

References	Method	Results	Purpose	Equipment	Data source	Data type[1]
Zheng et al. [41]	GAN	–/–/–	Failure prediction	Air pressure system and Turbofan engines	Sensors	RD
Nakka Agrawal, Choudhary [28]	Random forest, decision tree	73,9/81,3/–	Failure prediction	HPC	System logs	RD
Zhao et al. [40]	CNN	–/–/–	Failure diagnosis	Pump system	Sensors	RD
Martinez, Brewer et al. [25]	Genetic algorithm	–/–/95	Failure prediction	Rotors	Sensors	SD
Caponetto et al. [4]	CNN	99/–/–	Equipment remaining useful life	Rotors	Sensors	SD
Borghesi et al. [2]	Autoencoder	88–96/–/–	Anomaly detection in computational nodes	HPC	System logs	RD
Borghesi et al. [2]	Autoencoder	87–98/–/–	Anomaly detection in computational nodes	HPC	System logs	RD
Zhu et al. [42]	Backpropagation MLP	95/–/–	Anomaly detection in hard drives	Large-scale datacenter	Device logs	RD
Chen et al. [6]	RNN	84/80/–	Job failure detection	Cluster	System logs	RD
Tuncer et al. [34]	Random forest	–/–/–	Performance Loss	HPC	Performance metrics	RD
Du et al. [8]	LSTM	99/–/–	Anomaly detection in computational nodes	Cluster	System logs	RD
Klinken- berg et al. [20]	Descriptive statistics + Random forest	98/91/99	Anomaly detection	HPC	System logs and sensors	RD
Mohammed et al. [27]	SVM	–/0.67/91	Failure Prediction	HPC/Cloud	System logs and sensors	RD
Souza et al. [31]	CNN	99./99.5/99.6	Failure Prediction and Diagnostics	Rotating Machines	Sensors	RD

Some authors have developed convenient metrics to measure the performance of their algorithms.

misinterpretation of results. This field would benefit from future works that compare multiple deep learning techniques with the inclusion of multiple sensors belonging to a same supercomputing environment and the experimentation of newer deep learning methods, besides convolutional neural networks and recurrent neural networks.

Acknowledgements We would like to thank ATOS BULL, as well as the Supercomputing Center for Industrial Innovation (CS2I) and the Reference Center on Artificial Intelligence (CRIA), both from SENAI CIMATEC, for providing the infrastructure and environment for the execution of this research.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aydin O, Guldamlasioglu S (2017) Using LSTM networks to predict engine condition on large scale data processing framework. In: 2017 4th International Conference on Electrical and Electronic Engineering (ICEEE). IEEE, pp 281–285. https://doi.org/10.1109/iceee2.2017.7935834
2. Borghesi A, Bartolini A, Lombardi M, Milano M, Benini L (2019) Anomaly detection using autoencoders in high performance computing systems. Proc AAAI Conf Artif Intell 33:9428–9433. https://doi.org/10.1609/aaai.v33i01.33019428
3. Borghesi A, Libri A, Benini L, Bartolini A (2019) Online anomaly detection in hpc systems. In: 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS). IEEE, pp 229–233. https://doi.org/10.1109/AICAS.2019.8771527
4. Caponetto R, Rizzo F, Russotti L, Xibilia M (2019) Deep learning algorithm for predictive maintenance of rotating machines through the analysis of the orbits shape of the rotor shaft. Ergonomics and applied human factors. International conference on smart innovation. Springer, pp 245–250. https://doi.org/10.1007/978-3-030-22964-1_25
5. Carvalho T P, Soares F A, Vita R, Francisco R d P, Basto J P, Alcalá S G (2019) A systematic literature review of machine learning methods applied to predictive maintenance. Comput Ind Eng 137:106024. https://doi.org/10.1016/j.cie.2019.106024
6. Chen X, Lu CD, Pattabiraman K (2014) Failure prediction of jobs in compute clouds: a google cluster case study. In: 2014 IEEE international symposium on software reliability engineering workshops. IEEE, pp 341–346. https://doi.org/10.1109/ISSREW.2014.105
7. Das A, Mueller F, Siegel C, Vishnu A (2018) Desh: deep learning for system health prediction of lead times to failure in hpc. In: Proceedings of the 27th international symposium on high-performance parallel and distributed computing. pp 40–51. https://doi.org/10.1145/3208040.3208051
8. Du M, Li F, Zheng G, Srikumar V (2017) Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp 1285–1298. https://doi.org/10.1145/3133956.3134015
9. Essien A, Giannetti C (2020) A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders. IEEE Trans Ind Inform 16(9):6069–6078. https://doi.org/10.1109/TII.2020.2967556
10. Fink O, Wang Q, Svensén M, Dersin P, Lee WJ, Ducoffe M (2020) Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng Appl Artif Intell 92:103678. https://doi.org/10.1016/j.engappai.2020.103678
11. Francois C (2017) Deep learning with python. Apress, Berkeley
12. Ghiasvand S, Ciorba F.M (2019) Anomaly detection in high performance computers: a vicinity perspective. In: 2019 18th international symposium on parallel and distributed computing (ISPDC). IEEE, pp 112–120. https://doi.org/10.1109/ISPDC.2019.00024

13. Giommi L, Bonacorsi D, Diotalevi T, Tisbeni S.R, Rinaldi L, Morganti L, Falabella A, Ronchieri E, Ceccanti A, Martelli B (2019) Towards predictive maintenance with machine learning at the INFN-CNAF computing centre. In: international symposium on grids & clouds (ISGC). Taipei, Taiwan: Proceedings of Science, p 17. https://doi.org/10.22323/1.351.0003
14. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, vol 1. MIT press Cambridge, Cambridge
15. Guan Q, Zhang Z, Fu S (2012) Ensemble of bayesian predictors and decision trees for proactive failure management in cloud computing systems. J Commun 7(1):52–61. https://doi.org/10.4304/jcm.7.1.52-61
16. Haykin S (2007) Neural networks: a comprehensive foundation. Prentice-Hall Inc, New Jersey
17. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
18. Hu B, Pang CK, Luo M, Li X, Chan HL (2012) A two-stage equipment predictive maintenance framework for high-performance manufacturing systems. In: 2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, pp 1343–1348. https://doi.org/10.1109/ICIEA.2012.6360931
19. Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering-a systematic literature review. Inf Softw Technol 51(1):7–15. https://doi.org/10.1016/j.infsof.2008.09.009
20. Klinkenberg J, Terboven C, Lankes S, Müller MS (2017) Data mining-based analysis of hpc center operations. In: 2017 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, pp 766–773. https://doi.org/10.1109/CLUSTER.2017.23
21. Kraus M, Feuerriegel S (2019) Forecasting remaining useful life: interpretable deep learning approach via variational bayesian inferences. Decis Support Syst 125:113100. https://doi.org/10.1016/j.dss.2019.113100
22. Li X, Ding Q, Sun JQ (2018) Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf 172:1–11. https://doi.org/10.1016/j.ress.2017.11.021
23. Lima ALDCD, Aranha VM, Sperandio EG (2019) Manutenção preditiva aplicada a ambientes de missão crítica de supercomputação utilizando inteligência artificial: Uma revisão sistemática de literatura. In: Anais do V Simpósio Internacional de Inovação e Tecnologia. Blucher Engineering Proceedings, pp 657–664. https://doi.org/10.5151/siintec2019-82
24. Luo B, Wang H, Liu H, Li B, Peng F (2018) Early fault detection of machine tools based on deep learning and dynamic identification. IEEE Trans Ind Electron 66(1):509–518. https://doi.org/10.1109/TIE.2018.2807414
25. Martínez D, Brewer W, Strelzoff A, Wilson A, Wade D (2020) Rotorcraft virtual sensors via deep regression. J Parallel Distrib Comput 135:114–126. https://doi.org/10.1016/j.jpdc.2019.08.008
26. Mathew V, Toby T, Singh V, Rao B.M, Kumar M.G (2017) Prediction of Remaining Useful Lifetime (RUL) of turbofan engine using machine learning. In: 2017 IEEE International Conference on Circuits and Systems (ICCS). IEEE, pp 306–311. https://doi.org/10.1109/ICCS1.2017.8326010
27. Mohammed B, Awan I, Ugail H, Younas M (2019) Failure prediction using machine learning in a virtualised HPC system and application. Cluster Computing 22(2):471–485. https://doi.org/10.1007/s10586-019-02917-1
28. Nakka N, Agrawal A, Choudhary A (2011) Predicting node failure in high performance computing systems from failure and usage logs. In: 2011 IEEE international symposium on parallel and distributed processing workshops and Phd Forum. IEEE, pp 1557–1566. https://doi.org/10.1109/IPDPS.2011.310
29. Nguyen KT, Medjaher K (2019) A new dynamic predictive maintenance framework using deep learning for failure prognostics. Reliab Eng Syst Saf 188:251–262. https://doi.org/10.1016/j.ress.2019.03.018
30. Nie B, Xue, J, Gupta S, Patel T, Engelmann C, Smirni E, Tiwari D (2018) Machine learning models for GPU error prediction in a large scale HPC system. In: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, pp 95–106. https://doi.org/10.1109/DSN.2018.00022
31. Souza RM, Nascimento EGS, Miranda UA, Silva WJD, Lepikson HA (2021) Deep learning for diagnosis and classification of faults in industrial rotating machinery. Comput Ind Eng 153:107060. https://doi.org/10.1016/j.cie.2020.107060
32. Susto G.A, McLoone S, Pagano D, Schirru A, Pampuri S, Beghi A (2013) Prediction of integral type failures in semiconductor manufacturing through classification methods. In: 2013 IEEE 18th

Conference on Emerging Technologies & Factory Automation (ETFA). IEEE, pp 1–4. https://doi.org/10.1109/ETFA.2013.6648127
33. Susto G.A, Schirru A, Pampuri S, McLoone S, Beghi A (2014) Machine learning for predictive maintenance: a multiple classifier approach. IEEE Trans Ind Inform 11(3):812–820. https://doi.org/10.1109/TII.2014.2349359
34. Tuncer O, Ates E, Zhang Y, Turk A, Brandt J, Leung VJ, Egele M, Coskun AK (2017) Diagnosing performance variations in HPC applications using machine learning. International supercomputing conference. Springer, pp 355–373. https://doi.org/10.1007/978-3-319-58667-0_19
35. Wu Y, Yuan M, Dong S, Lin L, Liu Y (2018) Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. Neurocomputing 275:167–179. https://doi.org/10.1016/j.neucom.2017.05.063
36. Yurek O.E, Birant D (2019) Remaining useful life estimation for predictive maintenance using feature engineering. In: Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, pp 1–5. https://doi.org/10.1109/ASYU48272.2019.8946397
37. Zhang J, Wang P, Yan R, Gao R.X (2018) Long short-term memory for machine remaining life prediction. J Manuf Syst 48:78–86. https://doi.org/10.1016/j.jmsy.2018.05.011
38. Zhang K, Xu J, Min M.R, Jiang G, Pelechrinis K, Zhang H (2016) Automated IT system failure prediction: a deep learning approach. In: 2016 IEEE International Conference on Big Data (Big Data). IEEE, pp 1291–1300. https://doi.org/10.1109/BigData.2016.7840733
39. Zhang S, Li X, Wang J, Su S (2017) Curve-registration-based feature extraction for predictive maintenance of industrial equipment. International Conference on Collaborative Computing: Networking, Applications and Worksharing. Springer, pp 253–263. https://doi.org/10.1007/978-3-030-00916-8_24
40. Zhao H, Wang J, Gao P (2017) A Deep Learning Approach for Condition-Based Monitoring and Fault Diagnosis of Rod Pump System. STIoT Editorial Board 32. https://doi.org/10.29268/stsc.2017.0003
41. Zheng S, Farahat A, Gupta C (2019) Generative adversarial networks for failure prediction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp 621–637. https://doi.org/10.1007/978-3-030-46133-1_37
42. Zhu B, Wang G, Liu X, Hu D, Lin S, Ma J (2013) Proactive drive failure prediction for large scale storage systems. In: IEEE 29th symposium on mass storage systems and technologies (MSST). IEEE, pp 1–5. https://doi.org/10.1109/MSST.2013.6558427