Big Data for predictive maintenance of industrial machinery

Antti Koistinen

Control Engineering, Faculty of Technology
P.O. Box 4300, FI-90014 University of Oulu, Finland
E-mail: antti.koistinen@oulu.fi

Abstract

The operation of industrial manufacturing processes can suffer greatly when critical components fail suddenly. Large manufacturing processes can have plenty of critical components whose failure can interfere with the process operation. Typically these parts are changed periodically according to preventive maintenance strategy. Industry is eager to move towards predictive maintenance in order to make savings in spare parts and lower downtime. Predictive maintenance requires several measurement campaigns from a single part in order to make a working model or finding condition thresholds. A single measurement campaign from a certain part can take lots of time and give limited information about developing condition in certain environment. Multiplying the amount of this measured data leads to a more reliable estimate for the aspects affecting the condition and thresholds. The idea is to gather condition monitoring data from several similar machines or machine parts from a wide range of different environmental and stress conditions. This data can be used to generate models for several varying fault types. Data used for this system can include condition monitoring data from the target, automation system data describing operating conditions, metadata for describing environmental factors and maintenance reports in standardized form, including pictures of faults and events.

1. Introduction

Developing advanced predictive methods for condition based maintenance (CBM) and prognostics purposes is difficult and time consuming, since collecting single dataset from a newly installed part until part failure can take half a year or even more time. Gathering enough data from a single target for a useful predictive algorithm can take too long in order to build a useful application. Spare parts, driving conditions or environmental variables can change and that leaves the newly constructed model outdated. Other problem is that at the end of the measurement campaign, we have data for a single target in some specific environment and with certain operating parameters but we cannot use this data for similar targets in different environment. What we need is simultaneous measurements from similar targets in different environments and operating conditions gathered for common use. This enables the construction of failure models for varying different environments simultaneously. Data sharing promotes developing predictions since we can do proactive maintenance and algorithm development simultaneously.

Changes in the operating conditions of environmental factors can be accounted while using this system.

A vast amount of failure models and information from different failures can be used in the decision support system (DSS) for CBM. Mainenance crew can plan ahead for minimizing the unnecessary factory down time and operators can learn how to operate machinery at the optimal level to prolong their remaining useful life (RUL). In addition to maintenance crew, the spare part and machine manufactures can use this feedback information to design their products to be optimal for varying environments. This information could encourage operators to share their best practices in process operation across different sites. Different failure models can be easily compared after we have a suitable platform for the model use.

This article aims in collecting useful material and ideas for future case study where Big data is used in machine health monitoring. There are already some studies made in this area concerning machine fleet prognostics and health monitoring (PHM) but real life implementations still need lots of work to be done. Some of the related work done in the University of Cincinnati:

- Big Data from machine fleets for informative decision making. (1)
- Design framework for cloud-based health monitoring. (2)
- Degradation patterns for RUL estimation. (3)
- Spare parts inventory management for geographically distributed assets. (4)

2. Data for the monitoring system

This study assumes that all of the available data is at our disposal and does not concern data security and confidentiality issues. In real life, these aspects are very important and operators do not want to give any data about their process to be shared publicly to everyone. Data in industrial locations are more or less structured or semi-structured making analysis possible with only small effort. Data is also usually reliable since it is originated from industrial standard sensors and actuators. Reports are in digital form with easily standardisable fields. That being said the data can have several gaps or outliers because of some unknown influence. This needs to be accounted also by the CM system. Using imperfect data and Big data in industry is described in (5) with several techniques and methods for different areas of data processing.

2.1 Big Data

Big Data in this study consists of automation data from machine environment and operation, condition monitoring data (vibration data etc.), written reports about the events or failures including photos and other relevant information. One general way to characterize the Big Data is by describing it with three Vs (5):

- Volume: The amount of data is huge.
- Velocity: The continuous data streams generate more data all the time.
- Variety: Structured data, unstructured data, semi-structured data. Sensor data, text, pictures, video, and so on.

Some include additional fourth V for value and veracity. It is used for describing challenges in conclusions from Big data analysis.

A fleet of machines produces plenthora of data to be processed and certain preprocessing has to be done since transferring all of the data is neither possible nor sensible. Automation data and unstructured or semi-structured data from reports needs to be connected with corresponding condition monitoring data and it is important that all of this data is temporally synchronized. Every target can have some metadata implemented into its information when it is installed. This helps in the determining of important environmental factors that cannot be determined from the automation data alone. A factory with several parallel machines in the same operating state could have profiles ready for different machine environmental factors. This makes new machine installations fast and easy in terms of connecting them to the condition monitoring network.

The automation system data needs to be defined carefully since it can be difficult and laborious to add more data sets after the initial implementation. When this type of system is widely used, it needs standardized data sets from all the data providers (targets) and in addition to that, the data needs to be temporally comparable. One possibility is to convert the data into comparable form near the centralized database after transfer from the source but this would require lots of processing power when working with large fleets. The optimal way of reaching this goal would be that the data is sent in a standardized format from every provider. This way we could distribute the computational need across the condition monitoring network.

Event data concerning all the reported actions during the machine operation should be linked to CM data with timestamps. Event data includes all maintenance actions and other events worth reporting about the machine usage. Human control actions are important part of the machine usage information and should be included [(1)]. Event reporting can be rather diverse if done in open form. Reporting could use a standardized format with descriptive cells that can be used for event classification for describing related CM data.

Centralized data handling requires lots of processing power and with a fleet of targets this is not possible since dense CM data alone can be impossible to transfer. Preprocessing CM data should be done near the targets, data converting to standardized form at the local CM database, and further analysis and classification at the centralized database connected to other machines in the same fleet. Centralized computing can use predetermined classification information to roughly divide different targets. Outliers and gaps in recorded data need to be accounted in the centralized processing with the help of historical data sets from the target in question and if needed, use the comparable targets in value estimation. Data streams need working communication protocols and OPC for example can be used to acquire controller signals [(1)].

2.2 Data handling

2.2.1 Big data processing

Big Data needs automatic data handling and analysis tools since manual data analysis is impossible for vast amount data with varying types and structures. Some general Big data analysis techniques are listed in Table 1 [(6)].

Table 1. Big data analysis techniques

A/B testing	Machine learning	Spatial analysis
Association rule learning	Natural language processing	Statistics
Classification	Neural networks	Supervised learning
Cluster analysis	Optimization	Simulation
Crowdsourcing	Pattern recognition	Time series analysis
Data fusion and data integration	Predictive modelling	Unsupervised learning
Data mining	Regression	Visualization
Ensemble learning	Sentiment analysis	
Genetic algorithms	Signal processing	

More information on Big data handling tools and aspects can be found in [6].

Freely written event descriptions and comments can be analysed with text mining methods such as natural language processing (NLP) [7]. Free NLP toolkits can be found online:

- Stanford Topic Modeling Toolbox (nlp.stanford.edu/software/tmt/tmt-0.4/)
- Machine Learning for Language Toolkit MALLET (mallet.cs.umass.edu/)
- RTextTools (www.rtexttools.com)

Text mining methods benefit from preliminary dictionary building which involves analysing large datasets for discovering the groupings of terms. The standardized lists of terms can work as a prerequisite for classifying event records and the additional text mining can be used for finding more correlations. Correlations may be found between different fault groups which are useful in fault type comparison and finding different reasons for faults. Correlations inside the same fault group can reveal patterns leading to the fault.

2.2.1 Condition monitoring data

The condition monitoring data can consist of various data streams depending on the monitored target. It is possible to have several condition monitoring data streams from vital targets. Certain targets can be monitored using automation data already available e.g. moment or power consumption data from motors. In several cases there is a need for target specific monitoring methods such as algorithms for vibration [8], sound data [9], modelling based on laser scanning [10], electromagnetism [11].

The CM database should use a somewhat pre-designed format for describing targets belonging to the same fleet. There is a standardized structure defined for condition based maintenance data called The Open System Architecture for Condition Based Maintenance (OSA-CBM) provided by the Machine Information Management Open Systems Alliance (Mimosa). [12, 13] This standardized structure includes all the information necessary for locating the origin of the described data like location, operator, machine, part, sensor, timestamp etc. Standardized metadata format for datasets ensure that we are using right dataset from the correct target.

MIMOSA OSA-CBM defines a standard way for condition monitoring data transfer when using the same data in various differing systems. This definition includes six functional blocks which are defined in ISO 13374-1 and ISO 13374-2 standards:

1. Data acquisition (DA)
2. Data manipulation (DM)
3. State detection (SD)
4. Health assessment (HA)
5. Prognostics assessment (PA)
6. Advisory generation (AG)

Recorded data in the first step (DA) can be locally calculated to meaningful descriptive values in step 2 (DM). Third step can be done in cloud where several comparable datasets and models can aid in state detection (SD) for determining different abnormalities or generating new states for future use. This third step can be developed continuously as more information and analysis methods is presented for the fleet in question. Fourth step (HA) provides information on occurring faults and rates for condition development by using all gathered health, operational, and maintenance information [14]. Fifth step (PA) gives RUL predictions based on health assessment and predicted future stress loads. Comparable data from the fleet helps in forming predictions for the remaining target lifetime when used with known stress level. Cloud computing service can use both data and models in prognostics assessment depending on the characteristics of the machine fleet. Final step (AG) provides optimization information for the machine use in order to prolong its life or plan best time for maintenance. Maintenance details and spare part handling can be integrated into this step for automated spare part orders.

2.2.2 Stress monitoring using vibration data

Parts with cyclic action and repeated stress in waves can be monitored using the durability of the material according to Wöhler's curve or S-N curve [15]. Vibration monitoring is a useful tool for this type of targets. Vibration data from accelerometers can be processed into describing features or so called generalized norms [8] to represent stress levels. Scaling these norms for predefined linguistic levels of significance using nonlinear scaling enables the determination of material stress durability [16, 17]. These scaled stress indices can be used in building of cumulative stress curves that can indicate changes in condition with the increasing slope. This is demonstrated in the curve presented on the right side of Figure 1. Only indices in high stress area affect the material properties and are used in building of the cumulative stress curve.

Methods for local calculation needs to be pre-determined when using destructive data compression methods. There are also lossless methods which allow the restoring of data to its original form but using these methods leads to very large data streams. Some available compression methods are listed in [18] and one reference case for lossless compression in the wireless sensor network with the average compression of 59% was introduced in [19].

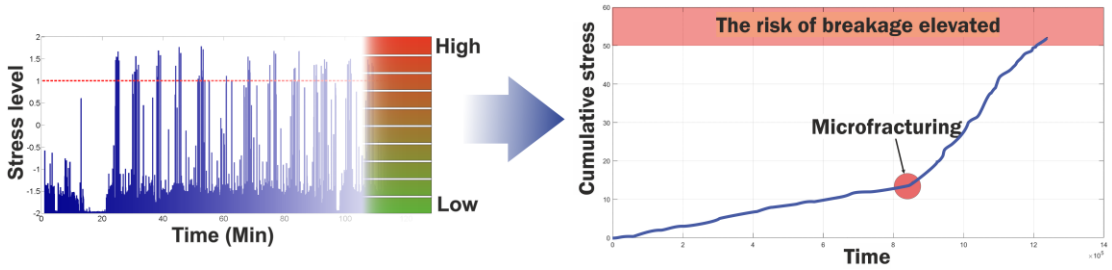

Figure 1. Generalised norms scaled to linguistic levels and used to build cumulative stress curve. Slope change in curve indicates changes in part condition.

Effective data processing is vital in creating robust and working networked system with distributed data handling capabilities. Data acquisition (DAQ) systems with field programmable gate array (FPGA) chips are able to effectively process large amounts of data in real time and calculate several different values from the same source data if needed [20, 21]. FPGA chip can do certain simple filtering and calculations more effectively than digital signal processors (DSP) [22]. Locally pre-processed data is easy to transfer onwards to local analysis and after that to CM database for state detection. DAQ systems can have a trigger for saving original raw data around unusual events for more detailed analysis of new failures.

3. From datasources to monitoring system

Artificial intelligence (AI) and machine learning are vital in finding comparable information among the fleet. The CM database needs to have different datasets and models listed according to their environmental and operative definitions so that it will be able to provide different targets with valid data. Wrong datasets lead to faulty predictions and make the system unusable. This requires updated information from the machine state e.g. if spare part comes from the different vendor or if the processed material have changed. Whenever possible we should use condition monitoring methods that are dependant on minimal amount of environmental factors. This way we can rely on the measurement results even if there are larger changes in the monitored process or its environment. Missing or corrupted data could be replaced with the estimates from the neighbouring machine in certain fleets with several machines in the same operating conditions. Automatic variable selection and scaling for the condition monitoring methods can help in the adjusting of the algorithms after maintenance break or in case of changed condition. Finding new fault models and ways to define RUL is important in building of the continuously developing system.

A concept of the monitoring system is presented in Figure 2. Dense raw data is locally calculated into describing features which can be stored in the local CM database for local analysis. The local CM database can include manually inputted data consisting of pictures and reports regarding the current data gathering period. This CM data and temporally synced automation data can be connected into the data package after local analysis and RUL estimation. Data package is then sent to the centralized database for analysis and fleet-wide comparison. Central analysis gains more information from this dataset for developing its analysis and data provider receives information that improves its local RUL prediction vastly by considering lots of historian data and past experiences from

several differing failure mechanisms. Additional benefits come from providing this data for maintenance operators and machine or parts manufacturers. This improves maintenance response times and develops machinery that can adapt to different end user needs.

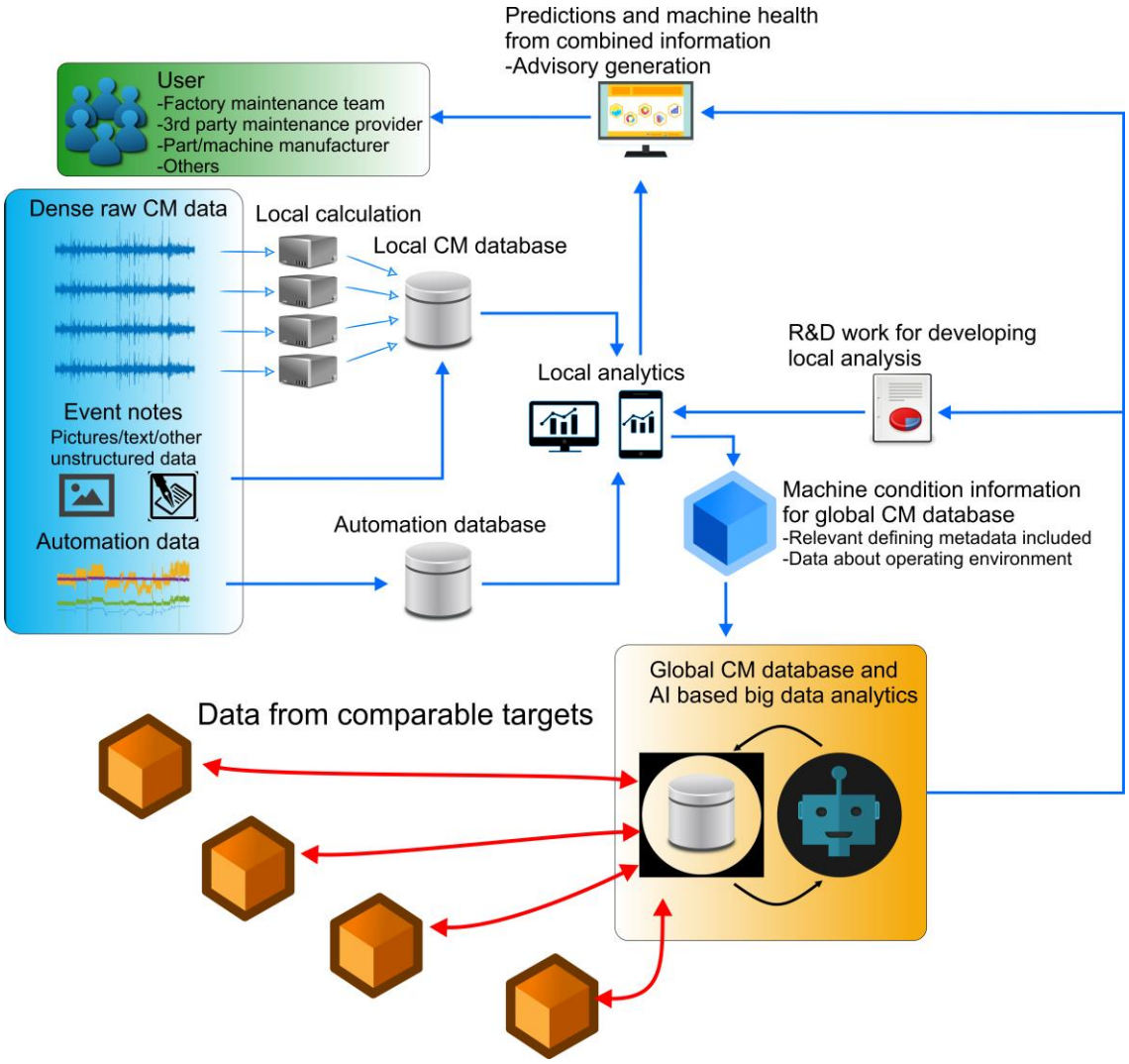

Figure 2. Condition monitoring ecosystem with centralized database for Big data analytics and failure models.

The monitoring system with the fleet of targets provides continuous data for updating predictions in all of the monitored locations. Historian data from previous faults improves prediction accuracy in a single monitored target and when used in conjunction with data from comparable targets with reliability information attached the prediction can be further improved. Increasing amount of data expands the library for varying fault types and mechanisms behind them. Lee et al.[(1)] identify the same problem in prognostics where most methods only apply for a single machine in defined working conditions and do not consider comparable machines as a fleet. They also identify several other issues concerning transfer towards intelligent machines like using production quality as an indication for machine health. Machine fleet event data can be clustered based on

different performances and working conditions for building a self-learning knowledge base. [1]

The 'Global CM database and AI based big data analytics' box in Figure 2 includes cloud based computing and one take on these analytics can be found in [2]. Yang et al. studied cloud based computing in machine health monitoring and presented a framework with guidelines for system implementation. Their Prognostics as a service (Paas) framework includes data acquisition, local computing, data transfer, database with analytics, and reporting. They found out that system has difficulties in handling irregular human based event data and this should be developed in future. [2]

When starting to receive several similar faults from comparable targets, we can forward this readily analysed data in form of report to machine or part manufacturer in order to give feedback how their product is performing and how it can be further developed. Alternatively, the manufacturer or spare parts provider can be connected to this monitoring system for automated spare parts delivery for minimizing down time in a factory. Numerous failures from similar actions can also indicate the development need in operating practices and promote the change in machine operating parameters. For example, a start up procedure could be too intense and therefore reduce RUL significantly. Regularily updated cloud based analysis needs to have feedback to local calculation for developing local analysis when required. This local analysis cannot change frequently if the changes in the analysis method have impact on the usability of gathered historical datasets. This can be avoided by processing data with several local analysis methods as the same source data can be easily processed with different parameters or algorithms without over burdening computational capacity or the transfer network. Unsupervised learning can help in developing new diagnostics methods and it can be used for finding new features [23]. New analysis methods needs to be accounted when introducing new fleets to the system.

Centralized CM database could handle the RUL calculations and failure characterization work since the data and models are stored in its database. Advisory generation for control actions should be done after information has been returned to target locations after centralized calculation. Local data users can be personnel in machine location, associate companies providing services for target machine users, or machine and parts manufactures gathering information for their product development purposes. The standardized data structure and format ensure that the processed data can be widely used by various different operators.

4. Preliminary guidelines for the use case scenario

The best way of testing this method would be piloting a moderately sized target fleet of comparable targets. The monitored fleet could include the maximum of hundred targets for effective data handling during the small scale testing. The data stream to be analysed can be estimated by setting a few initial numbers. When preprocessing vast amounts of condition monitoring data to a few easily transferrable describing features, we can reduce the data stream from possible 20GB per day to mere 100kB per day in a single target [24].

A fleet can also consist of simulated machines that produce comparable data with previously made or acquired fault database. Acquiring as wide fault library as possible with real fault characteristics is important for estimating system operation using real data. The initial decision for condition monitoring method is important as it needs to be able to detect different fault mechanisms. Other possibility is to use several methods for detecting specific faults. Big data methods testing would require plenty of varying data from several sources. It would be very useful to have a case where the reason for the fault is already found and there is plenty of varying information from several targets to be used as learning data.

4.1 Hose pump case

Oulu Mining School (OMS) has pilot size minerals grinding and enrichment circuits with over 30 comparable hose pumps that operate in similar conditions and pumps the same material. This setup could be used as a good platform for testing these predictive methods with the machine fleet. Automation data is recorded continuously from pump operation and this data could reveal changes in hose condition and changes in motor operating moment could be linked in wear and breakages. Motor information can be connected with grinding circuit operation data for estimating environmental factors such as feed slurry density that can affect pump RUL dramatically. There is currently data from two breakages caused by too dense slurry feed and continuous data gathering enables building of a small CM database in future. Datasets leading to faults needs to be linked with pictures taken from the faults and operator descriptions. The feed material changes from time to time and this needs to be accounted in pump and fault categorization by an algorithm which uses standardized operator descriptions. The pump categorization here can divide pumps for example between different process circuits based on different feed slurry densities. This setup also enables possibility to test this system with artificial simulated faults for developing CM system architecture and centralized cloud calculation with AI methods and Big data handling.

5. Conclusions

This paper presented several aspects concerning digitalization possibilities in advanced condition monitoring. This study provides some insight for the fleet-wide data processing system. First pilot is required to guide the development of this idea into the robustly working condition monitoring system.

Varying data that is available for describing machine operation and condition should be used more effectively than it currently is. Effective model and data gathering requires viewing comparable machines as fleets for increased information.

Centralized CM database with advanced data handling capabilities could be used for finding reasons for faults occurred in past at monitored machine fleets and developing machine operation prolonging its lifetime. Machines which are joined to this CM database in fleets can have their fleet specific datastreams, analysis methods, and thresholds but Big data analysis in the database could reveal new information that can be used between different fleets.

Database analysis can also use production quality data and user notes from machine operating time for finding fault root causes. The human based data should be shaped to as standardized format as convenient for automated data use. Standardization in data format, type, and signals is vital for creating a working system. This fleet level cloud based analysis should be done as a service for the manufacturing operators in order to ensure up-to-date analysis and system support.

Some benefits of this type of system are increased prediction accuracy for machine RUL, possibility to develop machines that are more suitable for their use, and decreased response time in maintenance operations. This study is a preliminary step for future research in implementing Big data into condition monitoring.

References

1. J Lee, H-A Kao, and S Yang, 'Service Innovation And Smart Analytics For Industry 4.0 And Big Data Environment', Procedia CIRP, Vol 16, pp 3–8, January 2014.
2. S Yang, B Bagheri, H-A Kao, and J Lee, 'A Unified Framework And Platform For Designing Of Cloud-Based Machine Health Monitoring And Manufacturing Systems', J. Manuf. Sci. Eng., Vol. 137, No 4, pp 6, 2014.
3. T Wang, J Yu, D Siegel, and J Lee, 'A Similarity-Based Prognostics Approach For Remaining Useful Life Estimation Of Engineered Systems', in 2008 International Conference on Prognostics and Health Management, pp 1–6, 2008.
4. C Jin, D Djurdjanovic, H D Ardakani, K Wang, M Buzza, B Begheri, P Brown, and J Lee, 'A Comprehensive Framework Of Factory-To-Factory Dynamic Fleet-Level Prognostics And Operation Management For Geographically Distributed Assets', in 2015 IEEE International Conference on Automation Science and Engineering (CASE), pp 225–230, 2015.
5. S J Qin, 'Process Data Analytics In The Era Of Big Data', AIChE J., Vol 60, No 9, pp 3092–3100, September 2014.
6. J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, and A H Byers, 'Big Data: The Next Frontier For Innovation, Competition, And Productivity', McKinsey Global Institute, 2011.
7. K A Neuendorf, 'The Content Analysis Guidebook', SAGE, 2016.
8. S Lahdelma and E Juuso, 'Signal Processing And Feature Extraction By Using Real Order Derivatives And Generalised Norms. Part 1: Methodology', Int. J. Cond. Monit., Vol 1, No 2, pp 46–53, November 2011.
9. H V Ravindra, Y G Srinivasa, and R Krishnamurthy, 'Acoustic Emission For Tool Condition Monitoring In Metal Cutting', Wear, Vol 212, No 1, pp 78–84, November 1997.
10. P P Rosario, R A Hall, and D M Maijer, 'Liner Wear And Performance Investigation Of Primary Gyratory Crushers', Miner. Eng., Vol 17, No 11–12, pp 1241–1254, November 2004.
11. A Sorsa, K Leiviskä, S Santa-aho, and T Lepistö, 'Quantitative Prediction Of Residual Stress And Hardness In Case-Hardened Steel Based On The Barkhausen Noise Measurement', NDT E Int., Vol 46, pp 100–106, March 2012.
12. 'MIMOSA | An Operations And Maintenance Information Open System Alliance'. [Online]. Available: http://www.mimosa.org/mimosa/. [Accessed: 19-May-2016]

13. T Sreenuch, A Tsourdos, and Ian Jennions, 'Distributed Embedded Condition Monitoring Systems Based On OSA-CBM Standard', Comput. Stand. Interfaces, Vol 35, No 2, pp 238–246, February 2013.
14. A Arnaiz, B Iung, A Adgar, T Naks, A Tohver, T Tommingas, and E Levrat, 'Information And Communication Technologies Within E-Maintenance', in *E-maintenance*, Springer, London, pp 39–60, 2010.
15. I Marines, X Bin, and C Bathias, 'An Understanding Of Very High Cycle Fatigue Of Metals', Int. J. Fatigue, Vol 25, No 9, pp 1101–1107, September 2003.
16. E K Juuso, 'Nonlinear Scaling Of Signals For Intelligent Analyzers', in IEEE International Workshop on Intelligent Signal Processing, 2005., pp 316–321, 2005.
17. E K Juuso, 'Integration Of Intelligent Systems In Development Of Smart Adaptive Systems', Int. J. Approx. Reason., Vol 35, No 3, pp 307–337, March 2004.
18. W Guo and P W Tse, 'A Novel Signal Compression Method Based On Optimal Ensemble Empirical Mode Decomposition For Bearing Vibration Signals', J. Sound Vib., Vol 332, No 2, pp 423–441, January 2013.
19. Q Huang, B Tang, and L Deng, 'Development Of High Synchronous Acquisition Accuracy Wireless Sensor Network For Machine Vibration Monitoring', Measurement, Vol 66, pp 35–44, April 2015.
20. S K Shome, U Datta, and S R K Vadali, 'FPGA Based Signal Prefiltering System For Vibration Analysis Of Induction Motor Failure Detection', Procedia Technol., Vol 4, pp 442–448, 2012.
21. W Zheng, R Liu, M Zhang, G Zhuang, and T Yuan, 'Design Of FPGA Based High-Speed Data Acquisition And Real-Time Data Processing System On J-TEXT Tokamak', Fusion Eng. Des., Vol 89, No 5, pp 698–701, May 2014.
22. J Vite-Frias, R de J Romero-Troncoso, and A Ordaz-Moreno, 'VHDL Core For 1024-Point Radix-4 FFT Computation', in International Conference on Reconfigurable Computing and FPGAs, 2005. ReConFig 2005, pp 4 – 24, 2005.
23. Y Lei, F Jia, J Lin, S Xing, and S X Ding, 'An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data', IEEE Trans. Ind. Electron., Vol 63, No 5, pp 3137–3147, May 2016.
24. A Koistinen, 'Crusher Wear Monitoring In IIoT Framework', in Proceedings of the 1st World Congress on Condition Monitoring, Ilec conference centre, London, pp 12, 2017.