

Predictive Analytics in a Pulp Mill using Factory Automation Data—Hidden Potential

Nykyri Mikko, Kuisma Mikko, Kärkkäinen Tommi J., Junkkari Tero, Kerkelä Kari,
Puustinen Jouko, Myrberg Jesse, Hallikas Jukka

This is a Author's accepted manuscript (AAM) version of a publication

published by IEEE

in 2019 IEEE 17th International Conference on Industrial Informatics (INDIN)

DOI: 10.1109/INDIN41052.2019.8972070

Copyright of the original publication: © 2019 IEEE

Please cite the publication as follows:

Nykyri, M. Kuisma, M., Kärkkäinen, T.J., Junkkari, T., Kerkelä, K., Puustinen, J., Myrberg, J., Hallikas, J. (2019). Predictive Analytics in a Pulp Mill using Factory Automation Data—Hidden Potential. In: 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 2019. pp. 1014-1020. DOI: 10.1109/INDIN41052.2019.8972070

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses.

**This is a parallel published version of an original publication.
This version can differ from the original published article.**

Predictive Analytics in a Pulp Mill using Factory Automation Data—Hidden Potential

Mikko Nykyri
School of Energy Systems
LUT University
Lappeenranta, Finland
mikko.nykyri@lut.fi

Mikko Kuisma
School of Energy Systems
LUT University
Lappeenranta, Finland
mikko.kuisma@lut.fi

Tommi J. Kärkkäinen
School of Energy Systems
LUT University
Lappeenranta, Finland
tommi.karkkainen@lut.fi

Tero Junkkari
Kaukas Pulp Mill
UPM
Lappeenranta, Finland
tero.junkkari@upm.com

Kari Kerkelä
Kaukas Pulp Mill
UPM
Lappeenranta, Finland
kari.kerkela@upm.com

Jouko Puustinen
Kaukas Pulp Mill
UPM
Lappeenranta, Finland
jouko.puustinen@upm.com

Jesse Myrberg
UPM
Helsinki, Finland
jesse.myrberg@upm.com

Jukka Hallikas
School of Business and Management
LUT University
Lappeenranta, Finland
jukka.hallikas@lut.fi

Abstract—Industrial automation systems have collected vast amounts of data for years. Data analytics and machine learning can be used to reveal different phenomena and anomalies, which may be otherwise impossible to see. However, the opportunities offered by the data are not currently utilized even though the technology is available. In this paper, a the potential use of the data analytics and machine learning of automation system data is presented. A case study on indirect measurement and predictive analysis of electric motor overcurrent was carried out in a pulp mill. Predictive models reached accuracy up to 98,85 %. The methods presented can be generalized to other processes. Since automation systems store data in most industrial sites, no additional hardware is necessarily needed for industrial internet of things (IIoT) systems, making a factory scale IIoT system possible.

Index Terms—Factory and process automation, Fault detection, diagnostics and prognostics, Intelligent Digital ecosystems

I. INTRODUCTION

The Internet of Things (IoT) is taking ground worldwide. The technology allows physical things to collect data and to be connected to each other and the internet. Data, when refined is valuable as it can help to detect phenomena which might not otherwise be observable. Applying IoT technology in industries forms the Industrial Internet of Things (IIoT), which improves productivity, increases predictability, allows new business opportunities, and creates new value for industry operators. Factories and industrial plants are vast sources of data. However, the potential of stored data is not widely utilized. One promising solution for data utilization is the use of machine learning and data analytics, which can emerge new information from the data available in a factory [1]. Collected factory data is also presented to be used in predictive

maintenance to avoid unplanned breakdowns [2], [3] and in remaining useful life predictions [4].

In this paper, the usage of machine learning, data analytics and factory automation data to detect and predict electric motor overcurrent is presented as a case study of harnessing the potential of data available in industrial environment. The case study represents a single point of a large, factory-sized IIoT system, where data is used to gain advantage in production. The IIoT system works in parallel with the existing automation system of the factory, and they share the source of the data. The detection and prediction of failures or abnormal functioning of process equipment is a typical need in industrial plant. Malfunctioning process equipment and machinery can cause the factories to be forced to halt the production for the time for repairs. This causes financial losses, as the factory does not make money when the production is stopped. If, for example, an electric motor failure could be predicted, the downtime caused by unexpected events could be minimized.

This paper takes also a glance on what are the driving factors and business enablers for building an IIoT system. Also, the generalization the methods presented in the pulp mill case study is discussed.

The factory selected for this study is UPM Kaukas pulp mill in Lappeenranta, Finland. The pulp mill has a high production rate of 740 000 tonnes of softwood and birch pulp per year. Therefore the mill represents a large industrial plant, and the observations presented in this paper can be generalized for applying in other industrial instances, also.

A. Internet of Things

Internet of Things (IoT), connecting things to the internet, allows the collection of data from different locations. Gathering raw data, however, is a not new technology and data may not be valuable in its raw form. Present day industrial automation systems collect and visualize data for the factory

operators, not utilizing the full value of the process data. By refining gathered data, new information can be created, which may prove to be valuable. The refined data has potential to, for example, yield savings for the operator, improve productivity and improve the quality of the product. The driving force of digitalization of industries is to harness the full potential of all of the available data and move towards smart manufacturing. IIoT is a key tool in that transformation [5].

The potential of IoT technology arises from new business opportunities associated to the industrial production and service environments. IoT provides more tools to support fact-based decision making based on the real-time information [6]. The automatization of the processes has increased the effectiveness of production systems in recent years, however, IoT is an essential technology enabler to transfer companies into the fourth industrial generation. In [7] it is described that in Industry 4.0 “manufacturing systems are vertically networked with business processes within factories and enterprises and horizontally connected to dispersed value networks that can be managed in real time – from the moment an order is placed right through to outbound logistics. In addition, they both enable and require end-to-end engineering across the entire value chain.” This vision highlights the integrated processes between firms and digitalization as an enabler for integration, automation and real time information sharing.

B. Data Analytics

The process of extraction of information from data sets is called machine learning [8]. This can be done, if a sufficient amount of data is available. In addition to the raw data, depending on the use case, detailed description of the data may be required. If, for example, prediction of failures is desired, data of both flawless and faulty operation is needed to find the correlations of data which lead towards the failure detection.

Machine learning happens by utilizing different algorithms to produce models, find patterns and predict a user-defined target output [8]. An algorithm is, simply put, a set of instructions to solve a problem. Machine learning algorithms are a collection of these instructions, which solve an algorithm-specific problem. There is a vast amount of such algorithms, which work in different kinds of problems. The correct algorithm for the task has to be selected in each case.

A trained machine learning algorithm is called a model. The model is an entity which can create predictions on data. Models are created by allowing a training algorithm to determine the internal weights or other parameters of the model based on the data. The model is then able to make predictions of the value or quantity which the model was trained to predict.

The trained model can be integrated into the factory systems to provide predictions using factory automation data (Fig. 1). To utilize machine learning in an industrial environment, necessary data connections need to be made, mainly to the automation system and its data storage for model training and continuous data flow. The predictions made with the model can

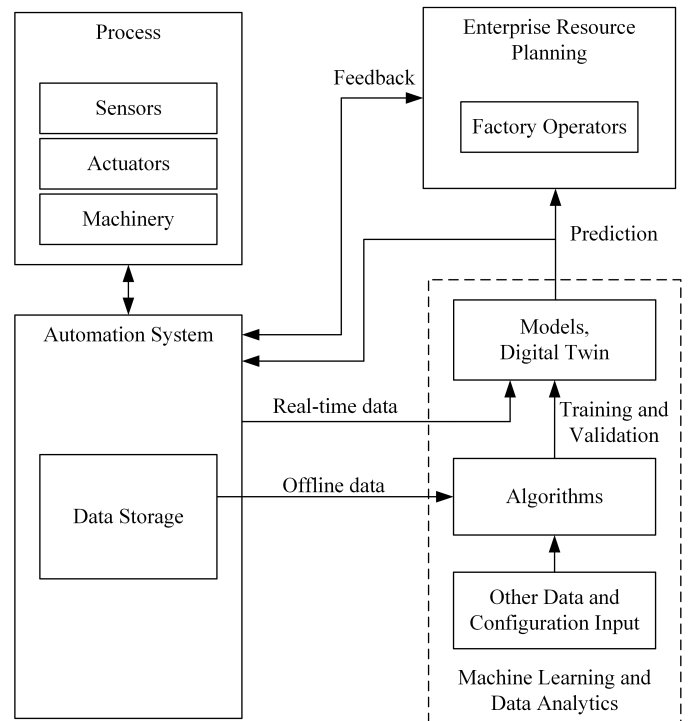


Fig. 1. The architecture and data flow used in machine learning. The data flows continuously from the automation system to the model, which gives predictions for the operators and back to the automation system. The model is trained using history data from the factory.

then be used by the factory operators, to control the process with the new information, when necessary.

In [11], similar data analytics on electric motors were performed. In this paper, the analysis is conducted further with feature importance analytics for all the algorithms and the method of feature importance calculation was changed due to issues of built-in feature importance tool of scikit-learn presented in [12]. The feature importances in this paper are calculated with Python library called Skater, which measures the entropy in changing of the predictions. The results presented in this paper therefore are more reliable than the ones presented in [11].

II. METHODS

An electric motor is a crucial piece of equipment in a pulp mill, and the number of electric motors in a pulp mill can easily be in the thousands. The motors provide power to many different parts of the pulp making process, such as material pumping and stirring. The motors also come in various sizes and power ratings. Therefore, an electric motor is suitable for a case study for IIoT, as they are common in different industries.

Malfunctioning electric motors can cause large-scale production losses in pulp mills. The faults in induction motors can be divided into three categories [13]:

- 1) Electrical faults (unbalanced supply voltage, overcurrent, overvoltage, overload, earth fault, inter-turn short circuit etc.)

TABLE I

CONFUSION MATRIX FOR THE PREDICTION. THE ROWS REPRESENT ACTUAL VALUES, AND THE COLUMNS REPRESENT PREDICTED VALUES.

		Actual overcurrent	
		No	Yes
Pred.	No	8683	35
	Yes	0	66

Accuracy: 99,60 %
Precision: 100,00 %
F1 Score: 79,04 %

whether the motor has been running on overcurrent in the past 10 minutes.

To predict into the future, the model needs to be re-trained. The data set remained mostly the same as with the present-time overcurrent detection. Because this model predicts into the future, the present time current information could be included into the model training dataset. Therefore the future prediction is not an indirect measurement, like the overcurrent detection presented above is. The output feature was set to be the generated overcurrent true/false flag from the next ten minutes of data.

Due to future predicting being a more challenging task than the indirect present-time detection, the future prediction was tested with six different classification algorithms: random forest, gradient boosting, logistic regression, Multi-layer Perceptron Classifier (MLPC), Gaussian Naive Bayes (NB) and Linear discriminant.

C. Model evaluation

To evaluate the models, the testing set is used to test how well the created models work. For the model, the ratio of correct predictions to total observations (accuracy) and the ratio of correctly predicted overcurrents to total predicted overcurrents (precision) are calculated. A metric considering both precision and recall of a test called F1 score is calculated, also. It is the harmonic average of precision and recall (the ratio of correctly predicted overcurrents to all overcurrent observations). Along the numeric performance metrics, the performance of the models are also presented with receiver operating characteristic curves and confusion matrices. If the model is deemed capable enough, it can be deployed into use and, for example, real-time data can be fed into the model to get predictions in real time.

D. Feature importance calculation

The input features have different levels of importance in the model—some of the data is more significant than other. Knowing which feature is most important may prove to be valuable information, as the factory operators are often interested in which parts of the process are the most crucial and have the possibility to form a bottleneck.

The data importances can be calculated by measuring the entropy in the change of predictions as the features are

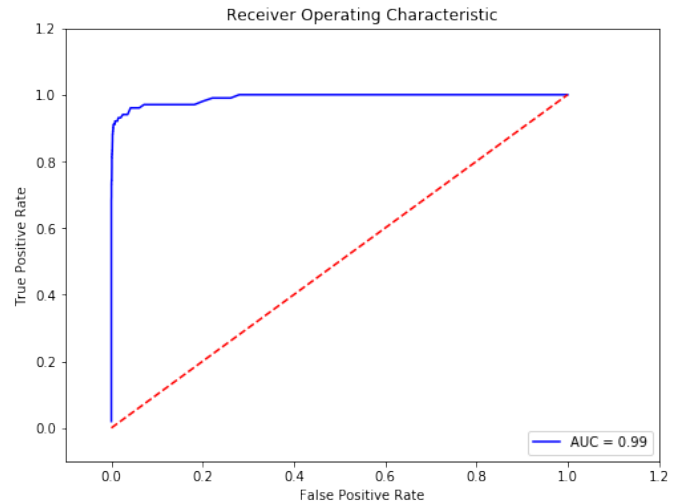


Fig. 3. The receiver operating characteristic curve of the model. The y-axis represents the true positive rate and x-axis represents the false positive rate. The red line represents a pure random guess, so the area between the curve and the red line should be as large as possible. The curve raises very sharply, indicating high accuracy.

individually perturbed. This can be done with skater, and the results can be presented as a bar chart, where each feature is given an importance between 0 and 1, while the sum of all importances being 1.

III. RESULTS

A. Overcurrent detection

The model accuracy was tested using the test set, and the model scored an accuracy score of 99,60 %. Based on the confusion matrix (Table I), it can be seen that the model is very good at detecting true negatives—the situations where the motor is not under overcurrent. This combined with zero false positives yields a precision score of 100 %. The rows represent predictions (pred.) and the columns represent actual values. The ideal result would be that the top right and bottom left corners would be zero, since those are either false positives or false negatives. The motor was under overcurrent in 101 samples and the model managed to predict 66 of those samples, missing none of them. However, the model gave quite a few false positives, 35 of them. The amount of false positives is quite high, when the number of true positives is taken into account.

The receiving operating characteristic curve (Fig. 3) shows the model functioning well. The steeper the curve, the better the model is detecting true positives. The ideal curve would be shaped like a step function, rising immediately to 1,0. The red line indicates a pure random guess - if the curve falls below the red line, it indicates that the model gives more false positives than true positives. Therefore, the area between the curve and the red line should be as large as possible.

B. Overcurrent prediction

The confusion matrices (Tables IIa-IIf) and receiving operating characteristic curves (Fig. 4a-4f) show that the different

TABLE II
CONFUSION MATRICES AND ACCURACY SCORES FOR SIX DIFFERENT
CLASSIFICATION ALGORITHMS.

		Actual overcurrent	
		No	Yes
Pred.	No	8669	92
	Yes	14	9
		Accuracy: 98,76 %	
		Precision: 39,13 %	
		F1 Score: 14,52 %	

(a) Random Forest

		Actual overcurrent	
		No	Yes
Pred.	No	8604	56
	Yes	79	45
		Accuracy: 98,83 %	
		Precision: 36,29 %	
		F1 Score: 40,00 %	

(b) Gradient Boosting

		Actual overcurrent	
		No	Yes
Pred.	No	8447	85
	Yes	236	16
		Accuracy: 96,35 %	
		Precision: 6,35 %	
		F1 Score: 9,07 %	

(c) Logistic Regression

		Actual overcurrent	
		No	Yes
Pred.	No	8668	101
	Yes	0	10
		Accuracy: 98,79 %	
		Precision: 0,00 %	
		F1 Score: 0,00 %	

(d) MLPC

		Actual overcurrent	
		No	Yes
Pred.	No	1241	1
	Yes	7442	100
		Accuracy: 15,27 %	
		Precision: 1,33 %	
		F1 Score: 2,62 %	

(e) Gaussian NB

		Actual overcurrent	
		No	Yes
Pred.	No	7584	30
	Yes	1099	71
		Accuracy: 87,15 %	
		Precision: 6,07 %	
		F1 Score: 11,17 %	

(f) Linear Discriminant

models yield different results. The highest model accuracy score was MLPC with a score of 98,85 %. This is due to the low amount of false negatives (zero instances). However, the model did not detect any overcurrents correctly, and all positive results were false positives. Even with slightly lower accuracy percentages, random forest or gradient boosting provides better results. Although both result in false positives and negatives, the models were able to predict overcurrents. For example, with gradient boosting model, when the model predicts an overcurrent, the prediction is correct 36,29 % of the time.

The best algorithm for each job depends purely on the data and the characteristics of the phenomena predicted—for example, on some cases neural network might perform best, and on other cases random forest might. The accuracy scores of the future prediction algorithms are quite high, if Gaussian NB is not taken into account. However, the precision scores alternate to some degree and are not comparable to the accuracies. The shape of the receiving operating characteristic

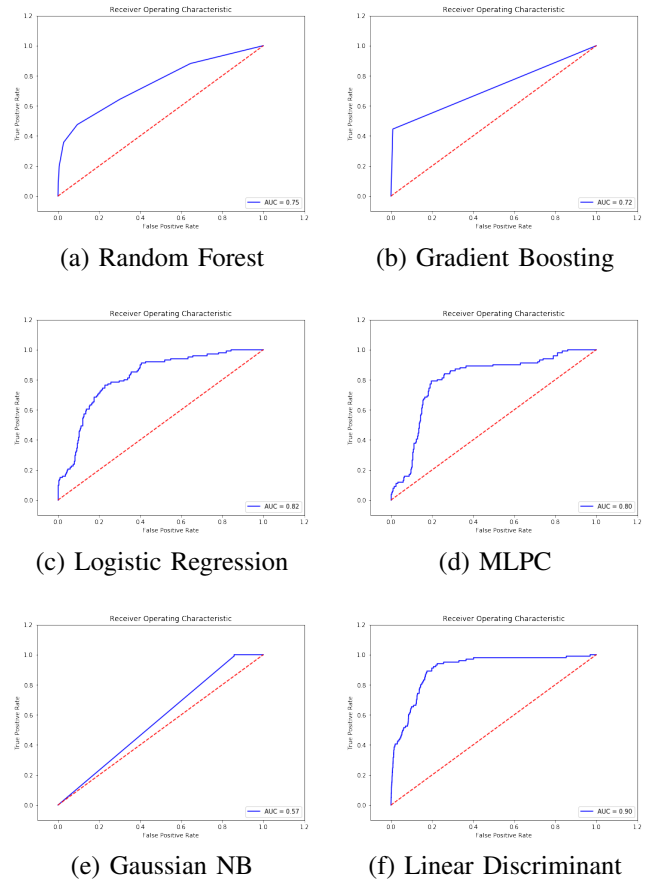


Fig. 4. Receiving operating characteristic curves of six different classification algorithms. The curves are not as ideal as in 3. The Gaussian NB curve (e) is the worst in performance, due to the low area between the red pure guess line and the curve.

curves itself does not tell everything. The performance of linear discriminant algorithm (Figure 4f), seem quite good when looking only at the receiving operating characteristic curve, but the amount of false positives in confusion matrix in Table II f is quite high, 1099 samples.

IV. DISCUSSION

As demonstrated in the case study, data analytics applied to automation data allows the creation of useful information for operation and maintenance of an industrial site. The pulp mill case shows that motor overcurrents can be predicted far ahead enough to take action in order to try prevent the upcoming motor overcurrent. For example, the mass pumped by the motor studied can be diluted with water to increase its viscosity and therefore make the task of pumping the pulp mass easier. The refined new information can also be used to, for example, indicate possible bottlenecks. As an example, the motor studied in this case study might have been misfitted or have been properly sized in the past, but today, when the mill produces higher amount of pulp, the motor became too small.

As shown in this paper and in [1]–[4], data analytics can be utilized in factories to gain new value. This paper also shows that the data for analytics purposes can be retrieved

from existing industrial automation systems. The models use the raw automation data, which the automation systems have been collecting for years. The data collected has a huge hidden potential, and the existing data collection infrastructure is suitable for data analytics. If the existing hardware and data flows are connected to a machine learning model, predictions could be made in real time. Eventually this develops towards a digital twin of the factory, where each piece of machinery and equipment is monitored and its performance is predicted.

Along with the operational improvement, IoT technology and data analytics can bring in new business opportunities, also. It is obvious that Industrial IoT has huge potential especially in predictive maintenance where machines are fitted with sensors and their conditions can be monitored and predicted [15]. The data can be used to forecast future events and optimize processes accordingly. Furthermore, IoT increases the traceability of production and services to follow a product and the processes it goes through. A good traceability system provides more transparency by offering specific information to stakeholders [16].

Despite the obvious potential in IoT and data analytics, many businesses may prove reluctant in deployment of the technology. A major concern is the security or the data sharing between different stakeholders. Secure data transactions and trust is necessary in IoT system deployment [17], [18]. In many applications, a cloud platform like Amazon Web Services, Microsoft Azure or IBM Cloud is used in data analytics and machine learning tasks. Especially in those cases, the flow of data between the industrial site and the cloud service must be secured so that no third parties may access the data. Also, the ownership of the data raises a lot of concern, as businesses do not want to give away their data, which may be of much value. If utilizing a cloud platform, for example, the data ownership issues must be resolved beforehand and complete trust between the stakeholders, the user of the platform and the provider of the platform must be formed. The field is quite conservative when it comes to data sharing, as the technologies are new and the methods are therefore not widely used.

V. CONCLUSION

With data analytics and machine learning models, predictions can be made using the process data collected by the automation system. The models presented in this paper show that data analytics can create new value in a pulp mill environment.

In our case study, predictive models were created using factory automation data to indirectly detect electric motor overload in real time and predict overcurrents ten minutes into the future. The case study proves that machine learning algorithms can be used as a tool for indirect measurements and a tool for forecasting industrial equipment performance. This can bring both operational and business benefits in industries.

A key advantage on factory IoT system development is the fact that, in many cases, the data necessary data for modeling already exists because the factory automation systems have stored the process data for years. Therefore, there is no need

for data acquisition before modeling, which makes the process faster. Also, no additional IoT hardware, for example, sensors and gateways, is needed.

A lot of questions are raised when discussing the ownership of the data. As data is shared between stakeholders, it is often not certain who owns the data and who decides, for example, how and where it is kept safely, to whom it is shared and how it is destroyed when necessary.

To reach the full potential of all of the data available, the solution presented in this paper should be scaled up to cover the whole pulp mill. This requires all of the motors in the mill to be modeled. This is possible, since the methods presented in the case example can be applied to all other equipment in the mill. The tools for building mill-wide systems exist, and building such system requires wide-scale knowledge about data analytics and process engineering. A possible solution for mill-wide system building is a cloud platform, such as Microsoft Azure, Amazon Web Services or IBM Cloud.

ACKNOWLEDGMENT

We would like to thank our colleagues from UPM Kaukas mill who provided expertise and the factory automation data that was the main research material in this paper.

REFERENCES

- [1] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone and A. Beghi, "Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," in *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812-820, June 2015.
- [2] P. K. Illa and N. Padhi, "Practical Guide to Smart Factory Transition Using IoT, Big Data and Edge Analytics," in *IEEE Access*, vol. 6, pp. 55162-55170, 2018.
- [3] C. P. Gatica, M. Koester, T. Gaukster, E. Berlin and M. Meyer, "An industrial analytics approach to predictive maintenance for machinery applications," 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Berlin, 2016, pp. 1-4.
- [4] H. Yan, J. Wan, C. Zhang, S. Tang, Q. Hua and Z. Wang, "Industrial Big Data Analytics for Prediction of Remaining Useful Life Based on Deep Learning," in *IEEE Access*, vol. 6, pp. 17190-17197, 2018.
- [5] E. Sisinni, A. Saifullah, S. Han, U. Jennehag and M. Gidlund, "Industrial Internet of Things: Challenges, Opportunities, and Directions," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724-4734, Nov. 2018.
- [6] Li, B. and Li, Y., "Internet of Things Drives Supply Chain Innovation: a Research Framework," *International Journal of Organized Innovation* Volume 9 Number 3, pp. 71-93, January 2017.
- [7] Kagermann, H., Helbig, J., Hellinger, A. and Wahlster, W., "Recommendations for implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry; final report of the Industrie 4.0 working group," *Forschungsunion*, 2013.
- [8] Bakshi, K. and Bakshi, K., "Considerations for artificial intelligence and machine learning: Approaches and use cases," 2018 IEEE Aerospace Conference, pp. 1-9.
- [9] Anderson, N., et al, "The Industrial Internet of Things Volume T3: Analytics Framework," url: <https://www.iiconsortium.org/industrial-analytics.htm>.
- [10] Sasikala, B.S., Biju, V.G., and Prashanth, C.M., "Kappa and accuracy evaluations of machine learning classifiers," 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pp. 20-23, May 2017.
- [11] Nykyri, M., "Data analytics for predictive maintenance in a pulp mill—case electric motors. Master's Thesis, LUT School of Energy Systems, Lappeenranta University of Technology, 2018.
- [12] Parr, T., Turgutlu, K., Csiszar, C. and Howard, J., "Beware default random forest importances", 2018, url: <https://explained.ai/rf-importance/index.html>

- [13] Karmakar, S., Chattopadhyay, S., Mitra, M. and Sengupta, S, Induction Motor Fault Diagnosis: Approach through Current Signature Analysis. Springer Singapore, 2016.
- [14] Ransom, D.L. and Hamilton, R, "Extending motor life with updated thermal model overload protection," IEEE Transactions on Industry Applications Volume 49 Issue 6, pp. 2471–2477, June 2013.
- [15] Ferber, S, "How the Internet of things changes everything," Harvard Business Review, 2013.
- [16] Wognum, P.M.N., Bremmers, H., Trienekens, J.H., van der Vorst, J.G.A.J. and Bloemhof, J.M., "Systems for sustainability and transparency of food supply chains–Current status and challenges," Advanced Engineering Informatics Volume 25 Issue 1, pp. 65–76, January 2011.
- [17] M. Frustaci, P. Pace, G. Aloï and G. Fortino, "Evaluating Critical Security Issues of the IoT World: Present and Future Challenges," in IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2483-2495, Aug. 2018.
- [18] R. Roman, P. Najera and J. Lopez, "Securing the Internet of Things," in Computer, vol. 44, no. 9, pp. 51-58, Sept. 2011.
doi: 10.1109/TII.2018.2890203