Factors Analysis and Prediction in Die-casting Process for Defects Reduction

Pavee Siriruk, Titiwetaya Yaikratok
System Engineering, School of Industrial Engineering
Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
pavee@g.sut.ac.th; titiwetaya@gmail.com

Abstract

Defect reduction has always been the continuous improvement topic that is being addressed in the manufacturing industry. Even nowadays, that the world is moving into the industrial 4.0, such a particular topic still has never outdated, only the new approaches have been introduced for the better achievement of defect reduction. This research aims to reduce the defects in die-casting process of the Hard Disk Drive (HDD) component manufacturing company, focusing on the effects of various machine parameters on the defects occurring in casting products. Predictive maintenance approach and machine learning have been introduced to determine the suitable data modelling technique.

- The most related independent factors can be identified through Feature Importance method.
- Decision Tree (DT) performed the best results among other classification methods.
- The 91.18% accuracy can be obtained by decision tree algorithm.

However, the ratio of labelled data still needs to be reviewed and optimized for the future work as well as continue the actual checking on the frontline production results with the Subject-Matter Expert (SME) also required in order to obtain the best prediction results.

Keywords

Big Data Analytics, Classification, Defects Prediction, Machine Learning, Predictive Maintenance.

1. Introduction

Predictive maintenance (PdM) has been used widely in many fields of industry, the main common goals are to reduce unscheduled downtime, improve productivity, reducing waste and unnecessary scrap, which finally led to benefits improvement for the company. Likewise in the HDD, a digital storage industry, ranging from upstream to downstream, there are many players who supply raw materials and components for the whole industry. This research focus on Motor baseplate manufacturing (3rd tier supplier), particularly on die-casting process. Various types of defects have been generated here, but the effort will be put solely on the outer surface porosity defect. This kind of defect cannot be detected 100% at the manufacturer site due to the inspection technique limitation. It will be found once passing through customer processes, and it has affected a lot of quality issues at customer's manufacturing (HDD). The occurrence of HDD failure at the end users is highly dangerous for HDD manufacturers, since the loss of their customer's crucial information might reduce the competitiveness in the digital storage market (Su and Huang 2018).

However, In the Supply Chain Management (SCM) point of view, this issue has been continued to discuss among multi-level suppliers for a while, a few scenarios have been proposed. The 1st scenario is to improve the inspection method at the 3rd tier supplier, but this action needs a huge investment which will increase the selling price at the upstream factory, and customer is still not ready to absorb this cost. Therefore, the 2nd scenario of multi-level production data analytics has been proposed in order to find the relationship between defects and machine parameters that might lead to the better ways to control the defect occurrence instead of improving the inspection process.

Typically, in the 2nd to 3rd tier suppliers, they are not interested to invest in the new machine and technology, those fully equipped with the sensors, costly, but enabling the multi-dimensional data from the machine. However, once their customers have made a big move into the industry 4.0 and predictive maintenance, the critical machine parameters, historical data have been requested for customer's machine learning projects, making this kind of future investment is undeniable, but still leaving the questions of how to utilize and gain the benefits from those data for the factory itself.

1.1 Objective

This research aims to reduce the defects in die-casting process of the Hard Disk Drive (HDD) component manufacturing company, determine the suitable data modelling techniques for predictive maintenance, focusing on machine parameters that caused the defects on casting products, several machine learning algorithms will be applied in this study, using Python, a standard programming software for the data modelling and prediction.

2. Literature Review

Not only a single algorithm has been studied in one particular project, Chen X. et al. (2021) have done the research in the Steel industries, like TATA steel company, Partial Least Squares Regression (PLSR), Artificial Neural Network (ANN) and Random Forests (RFs) techniques had been trained and tested. The result showed that Partial Least Squares Regression (PLSR) yielded the best. It can predict bush wear with a good accuracy and also can produce stable results. Durbhaka et al. (2016) presented the combination technique also found in the Wind Turbine that applied the PdM to predict the failure of bearings from vibration signals. K-Nearest Neighbor, k-means, and SVM are able to provide 78.8% - 87.0% accuracy. The Collaborative Recommendation Approach (CRA) is also applied on each technique, with this CRA model can achieved 93% accuracy. Apart from ML techniques selection, Rønsch et al. (2021) have pointed out that the selection of data sources extracted from the machines or processes are also important, different types of sensor may give the same predicted results, the cost of data collection needed to be taken into account.

Since the response variables (output) of this study are labelled data, OK/NG (discrete type), therefore the study will put the effort on supervised learning with classification method, Decision trees (DT), Logistic regression (LR), and Random Forest (RF) will be used for further study.

2.1 Classification Trees (DT)

Classification Trees or sometimes called Decision Trees (DT), is one of supervised learning methods. It is gradually using if - else conditions to split the messy data into two parts, layer by layer, until those remaining data cannot split, this is according to the entropy level. The weak points of classification trees are the model overfitting and it provides low accuracy since it came from a single tree. Based on the literature review with the keywords of Predictive Maintenance and Machine Learning since 2009 onwards, it was found that classification trees has been used gradually in many field of industries as follows; Kolokas et al. (2018) presented the forecasting faults of industrial equipment, in order to set the alarm before the incident occurs. Bukhsh et al. (2019) also applied predictive maintenance for the railway infrastructure, in order to reduce the unplanned maintenance cost. Kaparthi S. and Bumblauskas D. (2020) used DT in designing predictive maintenance systems to improve the uptime of machinery in the agriculture industry. Hsu et al. (2020) proposed the fault diagnosis and predicted the failure of wind turbines in Taiwan. Moreover, Aliyan et al. (2020) also used DT to predict the early stage of power system blackouts.

2.2 Logistic Regression (LR)

LR is used for classification problems. It will be applied only when the output variable is a discrete value, which could be binary classification like the probability of the event, Yes/No, Pass/Fail, however, the output variables can also be more than two types of classifications. Regarding Sigmoid function, the best fit model appears to be the "S curve"; it is not a straight line like we have seen in linear regression. A few cases in the past have been done on Logistic regression but the results obtained may not be that satisfactory. Hence, it has not been used widely compared to other classification methods. Liao et al. (2006) had developed the Remaining Useful Life (RUL) prediction of bearings, by accelerating the degradation and failure in order to obtain the vibration signals used for modelling. Phillips J. et al. (2015) proposed the outperform prediction results of Logistic Regression over Artificial Neural Network (ANN) and Support Vector Machines (SVM), in predicting the machinery's health through oil samples collected from mining trucks. Zhang and Pengzhu Zhang (2015) presented the failure prediction in which the various types of data collected from wind turbine sensors of 100 companies in the wind energy industry USA, mapped out the input data to turbine faults that related to alarming type, that link to the failure record.

2.3 Random Forests (RFs)

RFs is the assemble learning algorithm which combines many decision trees to be trained on the same datasets, then the final prediction will be made based on the most voted results from each tree. RFs can be used in both regression

and classification, the advantages of RFs seem to be avoiding the overfittings and providing highly accurate results. Random Forests (RF) is the most popular algorithm that has been using nowadays; Prytz et al. (2015) used RF to predict the required period of maintenance of air compressors in the trucks and buses. Canizo et al. (2017) introduced the prediction of wind turbine failures, Spain. Mathew et al. (2017) presented the prediction of the Remaining Useful Lifetime (RUL) of Turbofan Engine of NASA aircraft, 10 classification methods have been evaluated, and RF performed the best results. Another successful case proposed by Su and Huang (2018) to predict HDD failure in the data center, the system named "HDPass", consists of two types of models, 1) batch training by using historical data and 2) Real-time prediction by using real-time data, at 85.84% accuracy can be achieved. Lasisi and Attoh-Okine (2018) introduced the defect prediction for quality improvement by using track geometry data in the Railways. Amihai et al. (2018) used RF for the industrial assets' health prediction by using vibration data from pumps in the chemical plant. Kolokas et al. (2018) studied the forecasting faults of industrial equipment in anode production, it is possible to set the warning system 5-10 mins before the failure happens. Behera et al. (2019) have successfully predicted the Remaining Useful Life (RUL) and fault detection of individual equipment in the turbofan aircraft engine. Another example case that was found in Die - casting manufacturing in South Korea, Kim Ji Soo et al. (2020) presented a defect prediction and diagnosis system using the process condition data; over 89% accuracy of prediction can be obtained from this study.

3. Method

3.1 Machine Learning

Machine learning (ML) has been used increasingly nowadays, which provides positive impacts to many businesses and industries. It also gives the insights that enable the manufacturing patterns that help create a faster real-time decision making for even more targeted action and response. Rai R., et al. (2021) have explained that this PdM system can support various manufacturing tasks such as intelligent and continuous inspection, predictive maintenance, predictive defects, machinery's health diagnosis, quality improvement, process optimization, etc.

3.2 Framework

The general framework of ML has been introduced by Carvalho T. et al. (2019) as shown in Figure1. It is consisted of 4 main steps as follows;

1) Data collection, selecting the relevant data sources for ML.
2) Data pre-processing, this step includes the data exploration and contribution analysis that will help in data transformation, data cleansing, data reduction, this step is preparing data to be ready for ML.
3) Model selection, training and validation, in this step a few algorithms may be selected to be trained, then tested to see which algorithm provides the best predictive results and robustness, due to the process may changes over time.
4) The model maintenance is also needed in the ML framework.

Normally, there are four main techniques that have been used in PdM, the most model being deployed are (1) Random Forest followed by (2) Neural Network based methods (i.e. ANN - Artificial NN, CNN - Convolution NN, LSTM - Long Short-Term Memory Network, and Deep Learning), (3) Support Vector Machine (SVM), and (4) *k-mean*s.

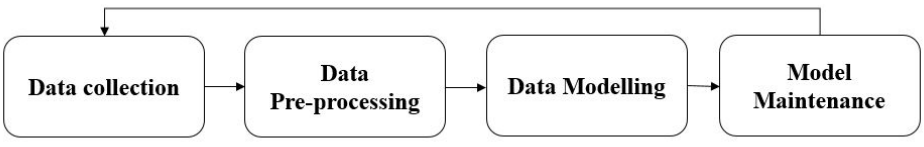

Figure. 1 General framework of machine learning

4. Data Collection

The real data from production line over 5 months were collected from machine sensors, one machine was setting up as a prototype; it can be said that those data represent all the process variation. There are 35 attributes of machine parameters such as mold temperature, other temperature related, rising up times, running value, speed, velocity, time stamp, etc., all these (continuous data) were generated real-time from the sensors of this prototype machine, each casting product has its own data, recorded by the serialization method, as shown in Table 1.

Table 1. Dataset schema

Index	Data fields	Types	Descriptions
1	S/N	Discrete	S/N
2	Date time	Continuous	Date time
3	Cycle time	Continuous	Cycle time
…	…	Discrete	…
…	Factor 1	Continuous	Sensor data
…	Factor 2	Continuous	Sensor data
…	…	Continuous	Sensor data
34	Factor 28	Continuous	Sensor data
35	Output (OK/NG)	Discrete	Inspection data

After that, each piece of casting product will be passed through the whole processes, the defects will be checked and identified at the final VMI station, all goods and defects will be recorded here with the serial number.
This casting process uses aluminum ingot as a raw material, with the well control of incoming quality, the effect of insufficient quality can be neglected from this study. There are 17 types of casting defects that have been classified in this process, but only surface porosity will be further analyzed.

Ahmad Nourian-Avval and Ali Fatemi (2020) had explained the mechanism of porosity that this defect can be caused by the remaining gas bubbles inside the liquid aluminum, those bubbles could not escape or purge out before the aluminum liquid cooled down, therefore the defect formed during the solidification.

Porosity defect can be caused by many factors, i.e. mold quality, mold temperature, liquid aluminum temperature, casting pressure, pouring rate, runner inadequate, low/high speed range inadequate, etc. (Sangwoo Park et al. 2019).

4.1 Data collection

Two groups of real data were collected: 1) machine data from machine sensors, and 2) defects data after classification from the visual inspection process, these two groups of data were matched together through serial number.
The 141,000 data sets collected from one machine in the real die-casting production over the 5 months period, after data cleansing, removing the duplicate and missing data, there are 92,000 datasets remaining for data modelling these datasets representing all the variations in the process.

4.2 Data pre-processing

In this study, correlation factors screening was not considered, all the independent factors were used as the “input” of the model. Before constructing the model, the data exploration and distribution analysis of each input factor also needs to be checked to make sure those raw data are suitable and ready for the modelling.

4.3 Performance evaluation

Confusion Matrix or sometimes called Error Matrix is the table that showing the performance of machine learning classification problem comparing the actual value to the prediction value, as shown in Figure 2.

		Actual Value	
		Positive	**Negative**
Prediction Value	**Positive**	**True Positive (TP)**	False Positive (FP)
	Negative	False Negative (FN)	**True Negative (TN)**

Figure 2. Confusion matrix

True Positive (TP): the actual value is positive and predicted value was positive.

True Negative (TN): the actual value is negative and predicted value was negative.

False Positive (FP): the actual value is negative but the predicted value was positive.

False Negative (FN): *the actual value is positive but the predicted value was negative.*

Calculation example; From Figure 2.

Precision or *Positive Predictive Value* measures how many % of the model correctly predicted the true positive?

$$Positive\ Predictive\ Value\ (Precision) = \frac{TP}{(TP+FP)}\ x\ 100\% \quad (1)$$

Recall or *Sensitivity* measures how many % of the actual positive cases were correctly predicted?

$$Recall = Sensitivity = \frac{TP}{(TP+FN)}\ x\ 100\% \quad (2)$$

Negative Predictive Value measures how many % of actual negative were correctly predicted?

$$Negative\ Predictive\ Value = \frac{TN}{(FN+TN)}\ x\ 100\% \quad (3)$$

Specificity measures how many % of negative prediction were correctly made?

$$Specificity = \frac{TN}{(FP+TN)}\ x\ 100\% \quad (4)$$

The percent *accuracy* is the ratio of total true prediction (positive & negative) on the total prediction results, the percent accuracy of prediction can then be calculated as follows;

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} x100\% \quad (5)$$

F-measure or *F-Score* measures the harmonic mean of Precision and Recall, it is the most common use to evaluate the imbalance classification problem.

$$F - Score = 2x\ \frac{(Recall\ x\ Precision)}{(Recall+Precision)} \quad (6)$$

G-mean indicates the balance of classification and prediction performance in both of major classification and minor classification.

$$G - mean = \sqrt{\frac{TP}{TP+FN} x \frac{TN}{TN+FP}} \quad (7)$$

5. Results and Discussion

Based on the feature importance (Extra Tree Classifier) analysis running on Python programming the results shown in Table 2. It was found that the Factor 26 (pressure releasing factor) revealed the highest score at 0.053, followed by Factor 3, Factor 27, Factor 8, and Factor 16, with the coefficient value of 0.049, 0.047, 0.043 and 0.042 respectively. The score indicates the contribution level of independent variables (input) on the output, the higher score, the more contributed to the output (NG case), as shown in Figure 3.

Table 2. Feature importance analysis results

Features	Score
Factor 26	0.053
Factor 3	0.049
Factor 27	0.047
Factor 8	0.043
Factor 16	0.042

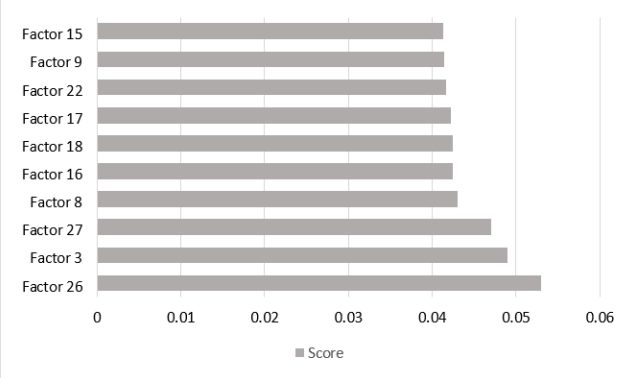

Figure. 3 Comparing score of feature importance analysis results

In addition, when compare the accuracy of Logistic regression (LR) with Decision tree (DT) and Random forest (RF), it was found that LR and RF can provide the same accuracy at 95.85%, while DT can provide 91.18% which is not much different from LR and RF. But when consider G-mean value (see equation 7) that implied the balance of classification, we found that only DT algorithm be able to predict both sides of positive and negative output, as the results shown in Table 3.

Table 3. Performance evaluation of each algorithm

Algorithms	Accuracy (%)	G-mean
Decision trees (DT)	95.85	0.28
Logistic Regression (LR)	95.85	0.00
Random Forest (RF)	91.18	0.00

Confusion Metrix compare the evaluating results of positive prediction and negative prediction of each algorithm, it was found that only Decision Trees (DT) can predicted both sides, as shown in Figure 4. While, LR and RF be able to predicted 100% of positive side but unable to predict the negative side, as shown in Figure 5 and Figure 6.

		Actual Value	
		OK	Porosity
Predicted Value	OK	5853	359
	Porosity	239	28

Figure. 4 Confusion Metrix of DT

		Actual Value	
		OK	Porosity
Predicted Value	OK	6212	0
	Porosity	267	0

Figure. 5 Confusion Metrix of LR

		Actual Value	
		OK	Porosity
Predicted Value	OK	6212	0
	Porosity	267	0

Figure. 6 Confusion Metrix of RF

Considering Precision, Recall, and F-Measure in Table 4, LR and RF provided the very good results of Precision, Recall, and F-Score, reflecting an excellent predictive results of the positive classification. Although, DT revealed slightly lower performance of all 3 measurement values. However, the further study of imbalance classification problem is crucial for the future work in order to improvement the predictive result of the negative classification.

Table 4. Predicting results

Algorithms	Precision	Recall	F-Measure
Decision trees (DT)	1.00	0.959	0.980
Logistic Regression (LR)	0.942	0.961	0.950
Random Forest (RF)	1.00	0.959	0.980

6. Conclusion

Based on Feature importance analysis method it can be concluded the Factor 26 (Pressure releasing related) is the most contributing factor affecting the porosity defect that occurs on the outer surface of die-cast product, followed by Factor 3 (High-speed related), Factor 27 (Pressure releasing related), Factor 8 (High-speed related), and Factor 16 (Filling pressure related). The Decision Trees (DT) algorithm perform the best predictive results when considering a minor classification regarding G-Mean value, the 91.18% accuracy can be obtained by this algorithm. However, after performing investigation and analysis on the results and raw data, there is one concerning point in which the extremely low percentage of NG data were observed. Therefore, in the future work, the imbalance dataset needs to reviewed and optimized. The actual checking on the frontline production with the Subject-Matter Expert (SME) is still required. As well as continue to study the other types of classification and several feature selection methods those would also be taken into the consideration in order to obtain the most suitable and robustness prediction model for the real practice manner.

Acknowledgment

The author would like to express the gratitude to colleagues, supervisors, and top management of this manufacturing company for the full support of the essential information and technical advice provided. Most importantly, special thanks to the academic advisor who had dedicated his time to advice. This project would have not been completed without these meaningful contributions.

References

Aliyan E., Aghamohammadi M., Kia M., Heidari A., Shafie-khah, M., & Catalão J. P., Decision tree analysis to identify harmful contingencies and estimate blackout indices for predicting system vulnerability. *Electric Power Systems Research*, 178, 106036, 2020.

Amihai I., Gitzel R., Kotriwala A. M., Pareschi D., Subbiah S., & Sosale G., An industrial case study using vibration data and machine learning to predict asset health. In 2018 *IEEE 20th Conference on Business Informatics (CBI)* vol. 1, pp. 178-185, IEEE, July 2018.

Behera S., Choubey A., Kanani C. S., Patel Y. S., Misra R., & Sillitti A., Ensemble trees learning based improved predictive maintenance using IIoT for turbofan engines, In Proceedings of *the 34th ACM/SIGAPP Symposium on pplied Computing,* pp. 842-850, April 2019.

Bukhsh Z. A., Saeed A., Stipanovic I., & Doree A. G., Predictive maintenance using tree-based classification techniques: A case of railway switches, *Transportation Research Part C: Emerging Technologies*, vol.101, pp. 35-54, 2019.

Canizo M., Onieva E., Conde A., Charramendieta S., & Trujillo S., Real-time predictive maintenance for wind turbines using Big Data frameworks, *In 2017 ieee international conference on prognostics and health management (icphm),* pp. 70-77, IEEE, June 2017.

Carvalho T. P., Soares F. A., Vita, R., Francisco R. D. P., Basto J. P., & Alcalá S. G., A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024, 2019.

Chen X., Van Hillegersberg J., Topan E., Smith S. & Roberts M., Application of data-driven models to predictive maintenance: Bearing wear prediction at TATA steel, *Expert Systems with Applications*, 186, 115699, 2021.

Durbhaka Gopi Krishna, and Barani Selvaraj, Predictive maintenance for wind turbine diagnostics using vibration signal analysis based on collaborative recommendation approach, *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE,* 2016.

Hsu J. Y., Wang Y. F., Lin K. C., Chen M. Y., & Hsu J. H. Y., Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning. *Ieee Access*, vol. 8, 23427-23439, 2020.

Kaparthi, Shashidhar, and Daniel Bumblauskas, Designing predictive maintenance systems using decision tree-based machine learning techniques, *International Journal of Quality & Reliability, Management,* 2020.

Kim Ji Soo, Jun Kim, and Ju Yeon Lee, Die-Casting Defect Prediction and Diagnosis System using Process Condition Data*, Procedia Manufacturing,* vol. 51*,* pp.359-364, 2020.

Kolokas N., Vafeiadis T., Ioannidis D. & Tzovaras D, Forecasting faults of industrial equipment using machine learning classifiers, *In 2018 Innovations in Intelligent Systems and Applications (INISTA),* pp. 1-6, IEEE, July 2018.

Lasisi, Ahmed and Nii Attoh-Okine, Principal components analysis and track quality index: A machine learning approach, *Transportation Research Part C: Emerging Technologies,* vol. 91, pp. 230-248, 2018.

Liao Haitao, Wenbiao Zhao, and Huairui Guo, Predicting remaining useful life of an individual unit using proportional hazards model and logistic regression model*, RAMS'06. Annual Reliability and Maintainability Symposium, 2006,* IEEE, 2006.

Wu Z., Lin W., Zhang Z., Wen A., & Lin L., An ensemble random forest algorithm for insurance big data analysis. *In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),* vol. 1, pp. 531-536, July 2017.

Mathew V., Toby T., Singh V., Rao B. M., & Kumar M. G., Prediction of Remaining Useful Lifetime (RUL) of turbofan engine using machine learning. *In 2017 IEEE International Conference on Circuits and Systems (ICCS),* pp. 306-311, IEEE, December 2017.

Nourian-Avval, Ahmad, and Ali Fatemi., Fatigue life prediction of cast aluminum alloy based on porosity characteristics, *Theoretical and Applied Fracture Mechanics 109:* 102774, 2020.

Park Sangwoo, Kim Changgyun, and Sekyoung Youm., Establishment of an IoT-based smart factory and data analysis model for the quality management of SMEs die-casting companies in Korea*, International Journal of Distributed Sensor Networks 15.10,* 1550147719879378, 2019.

Phillips J., Cripps E., Lau J. W., & Hodkiewicz M. R., Classifying machinery condition using oil samples and binary logistic regression, *Mechanical Systems and Signal Processing,* vol.60, pp. 316-325, 2015.

Prytz R., Nowaczyk S., Rögnvaldsson T., & Byttner S., Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data, *Engineering applications of artificial intelligence,* vol. 41, pp. 139-150, 2015.

Rai R., Tiwari, M. K., Ivanov, D., & Dolgui, A., Machine learning in manufacturing and industry 4.0 applications, *J. of Production Research,* 2021.

Rønsch G. Ø., Kulahci M., & Dybdahl M., An investigation of the utilisation of different data sources in manufacturing with application in injection moulding, *International Journal of Production Research,* pp. 1-18, 2021.

Su C. J., & Huang S. F, Real-time big data analytics for hard disk drive predictive maintenance, *Computers & Electrical Engineering,* vol. 71, pp. 93-101, 2018.

Zhang Zhongju, and Pengzhu Zhang, Seeing around the corner: an analytic approach for predictive maintenance using sensor data*." Journal of Management Analytics 2.4,* pp. 333-350, 2015.

Biographies

Pavee Siriruk received the Ph.D. (2009), degrees in industrial and systems engineering from Auburn University, Auburn, AL. Currently, he is a lecturer in the Department of Industrial Engineering at Suranaree University of Technology, Nakhon Ratchasima, Thailand. His research focuses on stochastic process, system optimization, and supply chain and logistics.

Titiwetaya Yaikratok received the B.E. (2012), degrees in chemical engineering from Suranaree University of Technology. She is now working as an assistant manager of the Data Solution group, a sub - section of the Management Information System department in the HDD component manufacturing company, Thailand. Her research interests are data analytics, predictive maintenance, and machine learning towards a digital transformation in the manufacturing and also in the other private/government sectors.