Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 00 (2017) 000–000

www.elsevier.com/locate/procedia

The 50th CIRP Conference on Manufacturing Systems

Towards Industry 4.0 Utilizing Data-Mining Techniques: a Case Study on Quality Improvement

Harley Oliff[a]*, Ying Liu[a].

[a]Mechanical and Manufacturing Engineering, Cardiff University, Cardiff, UK, CF22 3AA

* Corresponding author. Tel.: +44(0)7446 103980; *E-mail address:* OliffH@cardiff.ac.uk

Abstract

The use of data-mining as an analytical tool has been increasing in recent years; and the emergence of new manufacturing paradigms such as the Industry 4.0 initiative have led many smaller manufacturers to look at utilizing these powerful techniques; however, practical applications are still in their infancy, and remain out of reach for many of these small manufacturing enterprises (SME's). This paper focuses on methods to integrate these emerging paradigms into existing manufacturing processes, specifically, how data-mining principles may be used to begin to explore the concept of Intelligent Manufacturing under Industry 4.0; with a focus on improving product and process quality.

In collaboration with an industrial partner; a respected manufacturer of household electronic appliances, techniques were developed using open-source and freely-available software, running on readily available hardware and using only existing data-collection points, that were able to provide actionable feedback which could be used to make improvements to the manufacturing operations; and to increase product quality. This paper serves as evidence that the ability to utilise these techniques is now within reach of numerous smaller manufacturing operations, and provides a further understanding of how moves towards fully Industry 4.0 ready factories may be made in the years to come.

Keywords: Industry 4.0; Data-Mining; Intelligent Manufacturing.

1. Introduction

The principles of data and text mining are long established and well understood. However, the resurgence in popularity of the field -due to recent successes in the field of machine learning algorithms [31] has paved the way for this powerful new tool to be adopted by industry. This forward thinking, combined with the development of the relevant technologies [28] and the entry into the age of 'big-data' [27]; has shifted the ability to make use of these powerful methodologies beyond the bounds of academic institutions.

The 'Industry 4.0' methodology laid out at the 2013 Hamburg world fair [42] has developed into a focused and unique objective. The term, coined by the German government to describe their ongoing vision for manufacturing, whilst this concept is multi-faceted and often open to interpretation, there exists a clear theme of *intelligent manufacturing*; which makes use of advanced computational technologies, and the advancements in digital systems and machine learning processes to support decision making, run self-sufficiently via networks of distributed control, and to self-adjust and self-correct should problems arise.

The objectives have been prompted, in part, by the mounting pressures and challenges facing manufacturing industries in the new era. There is a massively increased demand for high quality, bespoke products [33], developed using sustainable and efficient methodologies. To meet this demand, intelligent, reconfigurable systems need to be developed. Estimates by government agencies put the potential gains in efficiency of such processes as high as 30%.

Initial steps have been taken in the implementation of intelligent systems, and many companies with large manufacturing requirements have begun to explore the potential of this area [4], however, the vast investments needed [26] both in capital and skills present a significant obstacle. Through this research, we outline a methodology to adopt the principles of data-mining and utilize them to support decision making with respect to quality, at both the component and control levels.

2212-8271

2. Literature Review

2.1. Big Data & Industry 4.0

Since its introduction, the Industry 4.0 initiative has been widely discussed and in the author's opinion has migrated into the role of a popular science buzzword to simply relate emerging digital and advanced manufacturing technologies. The initiative is broad, and a significant volume of focused research remains until such a paradigm can be realized if indeed it is possible to develop a system in line with the Industry 4.0 initiative in its current form at all.

Much of the shift towards Industry 4.0 has been driven by the emergence of 'Big-Data', and the issues associated with the way industrial operations collect, manage and interpret their data remain prevalent [7]. The concept of big data and considerations of how to deal with such large datasets is an intrinsic challenge of any system operating in an Industry 4.0 environment as it typically renders traditional statistical processing methods useless due to its complexity and sheer size.

Hilbert [13] outlines five main characteristics with which to describe big data: *Volume*, the quantity of generated and stored data. *Variety*, the type, and nature of the data. *Velocity*, the speed at which the data is generated and processed. *Variability*, The consistency of the data. *Veracity*, the quality of the captured data, which can vary massively between devices or even individual sensors. These five features of big-data present substantial challenges [5, 39], but are the source of its massive potential.

It is well known that the rate of data generation, capture and storage is continually increasing [14], and soon the volume of data generation in this field will require the consideration of potentially unbound datasets and continuous data streams [21]. However, despite the vast amounts of information that is being generated, relatively few companies involved in the manufacturing sector are utilizing this data [16].

Current research efforts [2, 26] have attempted to provide comprehensive definitions of the necessary 'criteria' that need to be met, across all areas of the business. Others, have attempted to illustrate how the paradigm will be implemented in the future [41, 23]. However, Industry 4.0 is a multi-faceted problem, and it is unlikely that all aspects of it will be applicable to all businesses.

2.2. Intelligent Manufacturing

Intelligent Manufacturing describes any manufacturing processes which involve a degree of computational intelligence. This can be via the use of embedded sensors as in the case of IoT technologies [3], and cover the use of analytical techniques on historical process data to provide Knowledge Discovery and support decision making within manufacturing systems [15, 19] or, ultimately, the development and implementation of full Cyber-Physical-Systems [20], a synthesis of physical and digital technologies across the entire manufacturing system; and necessary associated technologies and frameworks.

Intelligent manufacturing itself encompasses many emerging technologies and processes that are considered 'part of' Industry 4.0, and both Theoretical and Technological advancements are being seen at an ever increasing rate. Main research focuses include: novel automation control systems, with a focus on, decentralization, virtualization, reconfiguration, and adaptability [29, 9, 18, 35]; the development and application of machine learning and artificial intelligences [32]; and virtual and augmented reality systems, which are being used to bridge gaps in geography, knowledge and skill level [24]. In addition, other enabling technologies and associated fields of research have also seen renewed interest and novel ideas. Algorithm development and software engineering [36] have both seen a variety of successful advances in previous years.

Current implementations have demonstrated adaptive scheduling, real-time modelling of processes, and Decision Support Systems, that have been used to refine processes and component design. Indeed, significant studies in this area have resulted in a variety of frameworks by which to classify and evaluate such emerging systems [30]. The 5C's architecture, proposed by Lee [17], outlines the different intelligence levels achievable, and their associated technologies and capabilities. This architecture is illustrated in Figure.1.

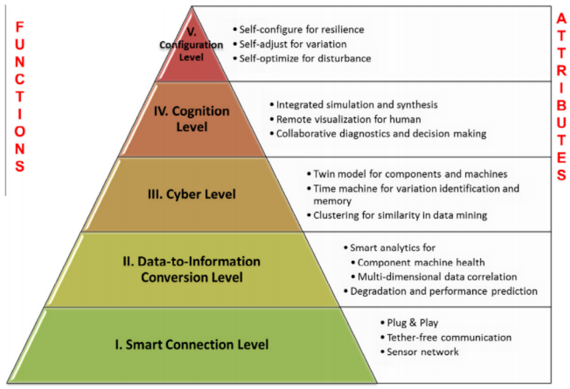

Figure.1. The 5C's architecture. Figure reproduced from [17]

Many of the aforementioned technologies still require significant development before they become realistic to use on an industrial scale, and as such are of limited use to those without the funding to conduct their own research. Indeed, many obstacles to the revolution will become apparent only once the research reaches a commercial level. Issues such as standardization [37] and validation [10] of such novel architectures are likely to further impede progress for those manufacturing facilities without the necessary resources; as are ethical concerns and political interventions [40].

Whilst the area of Intelligent Manufacturing is itself a multi-faceted problem, the recurring element that underpins much of this revolution is the collection, utilization and understanding of data, or the study of 'Informatics'; almost all of the areas linked with the intelligent manufacturing research area rely on the capture and analysis of data in some way. To this end the use of advanced data analytics and machine learning is a key technology to develop to further these other technologies; and the next step in this chain lies in utilizing the vast reserves of data through data mining and knowledge discovery, to better understand these manufacturing processes.

2.3. Industry 4.0 in SME's

Initial steps have been taken in the implementation of intelligent systems [12], and their appearance in manufacturing operations globally is ever increasing.

Small-scale implementations have achieved success, but tend to focus on specific tasks, such as control of motors and actuators to maintain process parameters. With a lack of understanding at the process scale.

The reality is that few companies have the necessary systems and capital in place to make leaps such as these in their operational processes, and find themselves presented with substantial barriers with respect to access. Due to the vast scope of the technologies and methodologies, and substantial costs involved and lack of understanding and competence with advanced manufacturing techniques, at the employee level [1].

The current literature highlights a gap in the application of these technologies. The rate of technological advancement in this area is outpacing its adoption in the manufacturing sector, as the challenges associated with practical use prevent many of the smaller operations from utilizing these advancements. With this in mind, the following process was developed and validated using a case study, to seek to overcome many of the common barriers to access.

3. Process Development

This section outlines the approach taken by the authors to develop a system within the confines of the existing system to implement data-mining to focus on the discovery of patterns and knowledge with which to provide a decision support system to the production engineers, with a focus on improved quality. A system model that provides insights to support decision making meets the necessary criteria of the *cognition* level of the 5C's architecture, demonstrating a level of intelligence. The research was conducted with the support of an industrial partner, a small manufacturing enterprise that produces washing machines and tumble-dryers, in a range of models. Discussions with our Industrial partner led to a list of criteria to meet as follows: The process must be built around the use of archived data; as automated digital collection of the data would require significant investment; The process must be developed to utilize readily available tools; The process must run without interference on established computing hardware within the facility.

Following discussions, a dataset was provided which consisted of a collection of *re-work records*, consisting of brief, textual descriptions of observed faults, and the actions taken to correct these faults. This was supported by supplementary, nominal attributes; such as the model number, the date etc. A full list of the attributes can be seen in Table.1.

Each instance in the dataset was representative of a single fault, and the data could be provided at a daily rate or any specified combination thereof, and with significant historical archived data to support and train. Based on the nature of the available dataset, a process was hypothesised that would enable the necessary preparation and knowledge extraction of the data. Once validated through the case-study, the process would be applicable and implementable to many manufacturing processes.

Table 1. Attribute descriptions for the provided dataset.

Attribute	Data Type	Description
Line	Nominal	The production line used for manufacture
Model	Nominal	The identifying model code
Date	Date	Datestamp of each instance
Fault Group	Text	The Group of faults into which the specific fault falls, typical values: scratch/damage, electrical, fit, etc.
Fault	Text	Details of the specific nature of the fault
Remedy	Text	Details of corrective action/disposal
Remedy Detail	Text	Additional notes on corrective action
Serial Number	Nominal	The unique serial number of the affected product
Surname	Nominal	The surname of the quality engineer entering the data

The process focuses on building an analytic model to produce a set of rules to be used as a decision support system, thus targeting the *Cognition* level in Lee's 5C's architecture. A flowchart illustrating the steps of the proposed process can be seen in Figure.2.

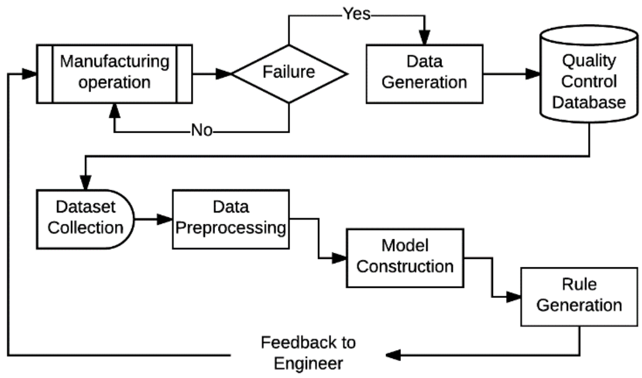

Figure.2. Flowchart illustrating the proposed system.

The manual data-collection methods resulted in data delivery in a discrete time period, typically daily. As such, software with which to perform this analysis was chosen on the basis of its ability to handle datasets rather than continuous data streams.

The process will use the WEKA (Waikato Environment for Knowledge Acquisition) data-mining software. WEKA is a powerful, java based, analytical tool focused on the application of data-mining techniques to datasets. Whilst many professional and supported data-mining software packages exist, WEKA offers distinct advantages to SME's in that it is quick and easy to implement and access; easy to use; and requires zero financial investment.

3.1. Case Study

Pre-processing of collected data is frequently necessary to improve accuracy and reliability of predictions. No standard data-collection methodologies exist, and the approach was necessarily heuristic. It was decided after careful examination of the dataset that as the dataset was only concerned with failures and problems that had occurred, there was little variation in the sentiment and style of the textual information.

The vast majority of instances all described some form of failure, the variance being the cause, and they were typically written by the same operators describing the same issues, and hence were syntactically and stylistically similar. To overcome this, it was decided to treat each instance of textual information nominally; that is, instead of separating each instance out and using a *bag of words* approach, the text contained would act in the same way as any other nominal value. The rule-based learning algorithms would then be able to build a classification model based on the frequency of the attribute values; attribute values that occur frequently within the same instance indicate a relationship between the attributes.

The *Fault Group* attribute was selected to act as our class attribute, and describes the category of fault recorded. Typical values include: ‘fit’, for faults involving assembly failures, and ‘scratch/damage’, where components are damaged and unusable. Multiple factors may contribute to the occurrence of these faults, and an accurate model would produce a set of rules, indicating how the different attribute values influence the faults observed to be occurring. This rule set can then be used as a decision support system, supplying information to the process engineers about the observed processes, through the construction of an Ishikawa diagram, an established quality and process control technique. The presentation of the analysis in this way will enable multiple rules to be visualized as the causal factors of each category branch on the diagram.

It was necessary to convert the textual data to the lowercase and remove all spaces, to prevent the algorithm distinguishing between different capitalizations or descriptions of the same problem. The WEKA software considers the same value written in the upper case a different value.

Consideration must also be given to outlier detection; infrequently occurring events that may lead to inaccuracies in the model. Using the *RemoveFrequentValues* filter, instances with unique attribute values that occur only once in the dataset can be removed. The *InterquartileRange* filter was then applied to isolate and remove any other infrequent instances.

Feature Selection in this instance was deemed unnecessary due to the limited number of attributes within the dataset. However, consideration was given to the attributes that would be used to extract information. Several attributes exhibited little variation, and others, it was clear, had little information to offer in terms of assembly faults. As such these attributes should be removed to minimise noise.

3.2. Algorithms

Extensive literature exists covering a wide range of the techniques that may be utilised as a part of the data mining process. A major family of algorithms are those which focus on Rule Based learning. As explained by [22], these algorithms are the oldest; some of the simplest; and work by using mathematical relationships to determine a set of ‘rules' by which to classify the data. As the computing power available continues to increase, these algorithms are becoming increasingly complex. These types of algorithm were considered preliminarily for this research, due to their ease of construction and interpretation.

Two main variations were tested: the PART algorithm, an implementation within WEKA of the C4.5 algorithm [25] which uses a *divide-and-conquer* approach to build a decision tree, before ‘pruning’ the unnecessary structures within the tree; and the JRip or RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) algorithm [8].

An initial sample of the dataset containing 1000 instances of quality control entries had been used to validate the pre-processing techniques, however, it was necessary to determine the optimum dataset size, as both too many instances and too few could lead to inaccurate models. A 6000 instance dataset was prepared using the relevant pre-processing techniques, the PART and JRip algorithms were then run and evaluated using a 10-fold-cross-validation, and the number of instances reduced between iterations. The percentage of correctly classified instances, when evaluated, is plotted for both algorithms in Figure.3a.

The results of this preliminary assessment indicated that a dataset size exceeding 5000 instances leads to negligible gains in model accuracy for both algorithms. This corresponds to approximately 10 days; an approximate working fortnight's worth of records.

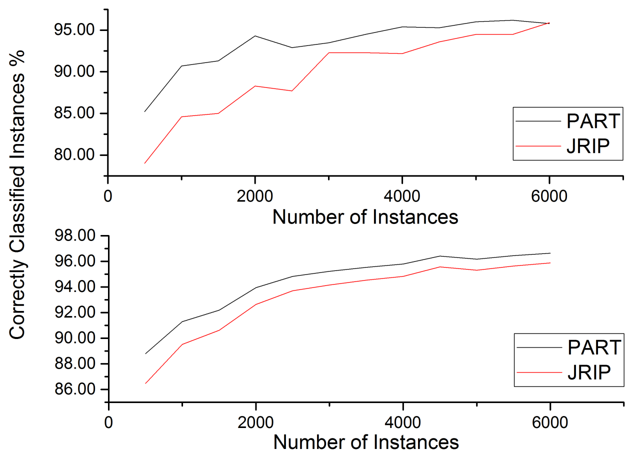

Figure.3. Percentage of correctly classified instances against the number of instances. a) Evaluated using 10-fold-cross-validation. b) Evaluated using isolated Test Set.

By logically partitioning the dataset into a decided time step, both short-term and long term patterns may be discovered, by considering the duration over which the data to be analysed was collected. The process also aims to be self-validating: As rules are uncovered and used to make decisions and take corrective actions, where the source of failure is resolved, the rules will change to reflect different patterns, as the prevalence of the resolved fault will decrease in the dataset. In order for this to remain true, each subsequent set of rules produced via the method outlined must be considered to supersede the previous set in terms of validity; the most recent set is the most accurate analysis of the current state.

One potential issue with using a *cross-fold-validation* technique for model evaluation is that of overfitting, models tested on the data used to train them often learn the patterns within that dataset, but perform worse when tested on data collected at a different instance in time. To remove this factor, and isolated *Test-Set* consisting of 1000 instances was created and used to evaluate the PART and JRip algorithms in the same manner as before. The results are shown in Figure.3b.

The results of the Test-Set evaluation supported the conclusion that increases in accuracy are negligible when the dataset size begins to exceed 5000 instances. The accuracy of the JRip and PART algorithms for the 5000 instance tests were 94.5% and 96.0% respectively.

3.3. Improving Model Accuracy

Two methods of improving classifier accuracy are bagging and boosting which are both methods that fall into the category of *ensemble learning.* Standard classifiers build simple models of the data, whereas, in ensemble learning, multiple base models are combined to produce an amalgamated model.

Bagging involves the creation of new datasets for multiple classifiers. In bagging, a dataset of *N* instances (where *N* is the size of the original dataset) is created by randomly drawing with replacement. The replacement means that instances from the original dataset may occur more than once, or not at all. These models then utilise a voting system to fully develop the final classifier.

Boosting involves the creation of a *series* of classifiers, where each in the series is given a different training set that is based on the performance of the preceding classifiers, and their prediction errors. Instances that were incorrectly predicted in previous models are given a greater weighting than those classified correctly and are more likely to be chosen for future datasets. In this way, the classifier becomes iteratively better, by focusing more heavily on the weaker areas of its learning at successive stages. To implement these two ideas into WEKA, algorithms exist in the WEKA toolkit, the *AdaBoostM1* algorithm to boost a classifier, and the *bagging* algorithm [6,25].

The 5000 instance dataset was tested using bagged and boosted versions of the algorithms. Whilst running the computations, the AdaBoostM1 algorithm, when used with the PART algorithm, would cause WEKA to run out of heap memory. Whilst it is possible to work around this issue, the concept of this case study was to produce a simplified and easily implementable procedure, and the necessary understanding of computation was judged to be excessive. The results of the bagging and boosting and their effects on the accuracy of the developed models can be seen in Table.2.

Table 2. Accuracies of the Bagging and AdaBoostM1 meta-algorithms.

Algorithm	Bagging Accuracy	AdaBoostM1 Accuracy
JRip	95.4%	97.3%
PART	96.9%	-

3.4. Decision Support Generation

A validated process now exists by which to produce a model, consisting of a set of rules, that can be used to support decisions regarding product quality issues. The JRip algorithm and the Boosting technique (to improve accuracy), can be used to produce an accurate model, with a sufficient degree of confidence. The model is trained using a 5000 instance dataset. This model can then be used to make predictions about the patterns contained in any future datasets produced. To demonstrate this, a further dataset containing the entries from 3 days of runtime, a Set of Rules was produced. Those with the highest *coverage* (a metric which expresses the number of correctly classified instances in comparison to the number incorrectly classified) are shown in Table.3. Rules with a high coverage, are not only prevalent in the dataset (and hence occur frequently), but are also those which the generated model is able to predict with the highest degree of accuracy.

Table 3. Rules Generated by the final model

Rule	Class (Fault Group)	Coverage
Fault = Plinth AND Model = 85969	Fit	15.7/0.0
Fault = Timer knob AND Model = 74628	Fit	31.3/0.0
Fault = Door Assy Fault AND Remedy = Change Part	Scratch/Damage	45.5/0.0
Fault = Drum AND Model = 74628	Fit	175.0/9.0
Remedy = Retest Auto	Auto Test	213.0/0.0
Remedy = Fit Part AND Date = 13/11/2015	Missing Part	13.2/0.0
Remedy = Fit Part AND Date = 12/11/2015	Missing Part	9.0/0.0
Fault = Worktop AND Remedy = Refit	Fit	34.9/0.0

From Table.3, several insights can be found directly, without the use of additional quality tools. For instance, the model highlights that the *Plinth* component on the 85969 model produced by the company is a frequent source of failure, specifically relating to the fit of the component to the product.

4. Case-Study Insights

The aim of this case study was to establish how best to utilise data mining to improve assembly and quality control processes; to allow them to be implemented into existing systems, with a minimal impact.

Validated results have been produced which can be easily be interpreted and become actionable pieces of information. To this end, the proposed system can be said to demonstrate an effective Decision Support System and qualifies at the *cognition* level in the 5C's architecture; demonstrating a level of intelligence in-line with the Industry 4.0 initiative. In addition, the case-study aims were fully met: the final system is implementable, works with archived data, and has a low computational requirement by design. The approach is adaptable, and as long as suitable care is taken to correctly partition the data and understand the effect that this partitioning will have, the method can be used to determine a vast number of different patterns depending on how the dataset is divided initially. Additional study of outlier detection and advanced algorithms could further refine the results, however the global model accuracy and high coverage of the drawn conclusions lead to considerable confidence that the results support real-world trends.

Whilst the methods developed demonstrated the possibility of using data-mining in this way, they are by no means ideal, and several challenges remain to be overcome.

The process is very much a demonstration of supervised learning, and whilst valid, it requires significant human input, in terms of both processing, and interpretation. The next logical step in the evolution of this process would be an automated system, which would perform the necessary corrective actions, or notify quality and process engineers of emerging trends.

5. Concluding Remarks

As explored previously, many manufacturing enterprises are keen to adopt principles of *intelligent manufacturing*, but are presented with a barrier to doing so. This work presents evidence, that some of these barriers preventing such adoption, may be overcome with considered use of freely available software and existing data. The industry 4.0 initiative places significant emphasis on the utilisation data to form intelligent systems and processes, and by exploring the ways in which companies may utilise their existing records, such an intelligent system has been presented. Whilst in this instance, the methodology proved useful, countless variations in manufacturing processes mean that such a problem is difficult to generalize to all processes, and significant further work is required in this field to realise the full potential of intelligent manufacturing.

References

[1] Adolph, S., Tisch, M. and Metternich, J., 2014. Challenges and approaches to competency development for future production. *Journal of International Scientific Publications–Educational Alternatives: 1001-10.*

[2] Almada-Lobo, F., 2016. The Industry 4.0 revolution and the future of Manufacturing Execution Systems (MES). Journal of Innovation Management, 3(4), pp.16-21.

[3] Atzori, L. et al. 2010. *The internet of things: A survey*. In: Computer networks. Vol 54. pp. 2787-2805.

[4] Auschitzky, E. et al. 2014. *How Big Data can improve manufacturing* [Online]. Available at: http://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing[Accessed: 16/10/2016]

[5] Babcock, B., et al. 2002. *Models and issues in data stream systems*. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems pp. 1-16.

[6] Breiman, L. 1996. Bagging predictors. *Machine learning*. Vol 24. pp.123-140.

[7] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V. and Rong, X., 2015. Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, *2015*, p.12.

[8] Cohen, W.W. 1995. Fast effective rule induction. *Proceedings of the twelfth international conference on machine learning*. pp. 115-123.

[9] da Silva, R.M., Benítez-Pina, I.F., Blos, M.F., Santos Filho, D.J. and Miyagi, P.E., 2015. Modelling of reconfigurable distributed manufacturing control systems. IFAC-PapersOnLine, 48(3), pp.1284-1289.

[10] Feth, P., Bauer, T. and Kuhn, T., 2015. Virtual Validation of Cyber Physical Systems. In *Software Engineering & Management* (pp. 201-206).

[11] Freund, Y. and Schapire, R.E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*. Vol 55. pp.119-139.

[12] Hansson, K., Yella, S., Dougherty, M. and Fleyeh, H., 2016. Machine Learning Algorithms in Heavy Process Manufacturing. American Journal of Intelligent Systems, 6(1), pp.1-13.

[13] Hilbert, M. 2016. Big Data for Development: *A Review of Promises and Challenges*. In: Development Policy Review. Vol 34. pp.135-174.

[14] J. Gantz and D. Remsel, ''The digital universe in 2020: Big data, bigger digital shadows and biggest growth in the far east,'' in Proc. IDC iView, IDC Anal. Future, 2012.

[15] Lee, J., Bagheri, B. and Jin, C., 2016. Introduction to cyber manufacturing. *Manufacturing Letters*, *8*, pp.11-15.

[16] Lee, J., Kao, H.A. and Yang, S., 2014. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia CIRP, 16, pp.3-8.

[17] Lee, J., Bagheri, B. and Kao, H.A., 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, pp.18-23.

[18] Leitão, P., 2009. Agent-based distributed manufacturing control: A state-of-the-art survey. *Engineering Applications of Artificial Intelligence*, *22*(7), pp.979-991.

[19] Mi, M. and Zolotov, I., 2016, February. Comparison between multi-class classifiers and deep learning with focus on industry 4.0. In *2016 Cybernetics & Informatics (K&I)* (pp. 1-5). IEEE.

[20] Monostori, L., 2014. Cyber-physical production systems: roots, expectations and R&D challenges. *Procedia CIRP*, *17*, pp.9-13.

[21] Moran, K. 2012. *Data analytics for manufacturing.* Advanced Manufacturing Technology. vol.5.

[22] Nosofsky, R.M. et al. 1994. Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*. Vol 22. pp.352-369.

[23] Pan, M., Sikorski, J., Kastner, C.A., Akroyd, J., Mosbach, S., Lau, R. and Kraft, M., 2015. Applying industry 4.0 to the Jurong Island eco-industrial park. *Energy Procedia*, *75*, pp.1536-1541.

[24] Pirvu, B.C., Zamfirescu, C.B. and Gorecky, D., 2016. Engineering insights from an anthropocentric cyber-physical system: A case study for an assembly station. *Mechatronics*, *34*, pp.147-155.

[25] Quinlan, J.R., 1996, August. Bagging, boosting, and C4. 5. In *AAAI/IAAI, Vol. 1* (pp. 725-730).

[26] Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P. and Harnisch, M., 2015. Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries. *Boston Consulting Group*.

[27] Russom, P. 2011. *Big Data Analytics.* TDWI Best Practices Report. Fourth Quarter.

[28] Sebastiani, F. 2001. *Machine Learning in Automated Text Categorization.* The National Research Council, Italy.

[29] Shafiq, S.I., Sanin, C., Szczerbicki, E. and Toro, C., 2016. Virtual Engineering Factory: Creating Experience Base for Industry 4.0. *Cybernetics and Systems*, *47*(1-2), pp.32-47.

[30] Shearer, C. 2000. The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*. Vol 5, pp. 22

[31] Silver, D. and Hassabis, D., 2016. AlphaGo: Mastering the Ancient Game of Go with Machine Learning. *Research Blog*.

[32] Spezzano, G. and Vinci, A., 2015. Pattern Detection in Cyber-Physical Systems. *Procedia Computer Science*, *52*, pp.1016-1021.

[33] Tao, F. et al. 2011. *Cloud manufacturing: a computing and service-oriented manufacturing model.* Proceedings of the Institution of Mechanical Engineers. Part B. Journal of engineering manufacture. Vol 225. pp.1969-1976.

[34] Theorin, A., Bengtsson, K., Provost, J., Lieder, M., Johnsson, C., Lundholm, T. and Lennartson, B., 2016. An event-driven manufacturing information system architecture for Industry 4.0. *International Journal of Production Research*, pp.1-15.

[35] Valckenaers, P. and Van Brussel, H., 2005. Holonic manufacturing execution systems. *CIRP Annals-Manufacturing Technology*, *54*(1), pp.427-432.

[36] Vyatkin, V., 2013. Software engineering in industrial automation: State-of-the-art review. *IEEE Transactions on Industrial Informatics*, *9*(3), pp.1234-1249.

[37] Weyer, S., Schmitt, M., Ohmer, M. and Gorecky, D., 2015. Towards Industry 4.0-Standardization as the crucial challenge for highly modular, multi-vendor production systems. *IFAC-PapersOnLine*, *48*(3), pp.579-584.

[38] Witten, I.H. et al. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Ed. Elsevier.

[39] Wu, X. et al. 2014. *Data Mining with Big Data.* IEEE Transactions of Knowledge and Data Engineering. Vol 26. Pp 97-107.

[40] Zhou, K., Liu, T. and Zhou, L., 2015, August. Industry 4.0: Towards future industrial opportunities and challenges. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on* (pp. 2147-2152). IEEE.

[41] Zuehlke, D., 2010. SmartFactory—Towards a factory-of-things. *Annual Reviews in Control*, *34*(1), pp.129-138.

[42] Zühlke, D. 2014. *Industrie 4.0: From Vision to Reality.* Tech News. Vol 26. pp 10-11.