Advanced Logistic Systems – Theory and Practice, Vol. 18, No. 3 (2024), pp. 19-28
https://doi.org/10.32971/als.2024.025

THE ALWAYS CHANGING DATA PROBLEM OF USING AI IN MANUFACTURING - USING SYNTHETIC DATA FROM THE DIGITAL TWIN TO FEED AI MODELS

ZSOLT MOLNÁR[1] – PÉTER TAMÁS[2] – BÉLA ILLÉS[3]

Abstract: *Production is becoming increasingly flexible, which also requires the flexibility of the support system for the production. And the key here is the speed of decisions, in which the support of modern artificial intelligence systems can be crucial. Flexible production is based on a well-planned control of production, which increasingly uses some artificial intelligence component. Artificial intelligence can already be useful in the early stages of planning the production line, and of course it can also control the daily operation of the production line after the installation of the production place or line. The biggest problem is supplying the neural network that controls the artificial intelligence with training data. The production lines typically change every 2-4 months, new products appear, the layout changes, and the main process data also changes due to the development of the processes. This results in the training data becoming outdated or obsolete very quickly and thus cannot be used to train models anymore. High-quality learning data can be produced by digital twin models of production lines. Such synthetic data has several advantages over data collected from production. In this article, we investigate how useful this synthetic data is during the life cycle of the production line.*

Keywords: *flexible manufacturing, layout, systematic layout planning, digital twin, artificial intelligence, decision table*

1. FLEXIBLE MANUFACTURING SYSTEMS AND THE LIFECYCLE OF A PRODUCTION LINE

Increasing the flexibility of production is the result of demand from two sides. On the one hand, the development of IT and automation in recent decades and the geopolitical changes of recent years appeared as an internal demand of manufacturing companies to be able to respond to market demands as flexibly and resiliently as possible. On the other hand, customers are increasingly demanding customized, unique products, which also requires an increase in the flexibility of production by being able to produce as large a variety of products and product portfolios as possible on a given production line or production area [1].

Flexible manufacturing systems consist of three main components: machines, material handling devices and control logic [2, 3]. The complexity of the system is usually at the control logic, which must change and adapt adaptively during the life cycle of the production line [4].

Examining the life cycle of the product and the production line is very important to understanding the problem (Fig. 1). In the case of flexible production systems, the life cycle

[1] PhD student, University of Miskolc, Institute of Logistics, Hungary
zsolt.molnar.zsolt@outlook.com
[2] university professor, University of Miskolc, Institute of Logistics, Hungary
peter.tamas@uni-miskolc.hu
[3] university professor, University of Miskolc, Institute of Logistics, Hungary
bela.illes@uni-miskolc.hu

of a product is on average 2 years, while the life cycle of production lines can be 10 years, during which 10-100 different products can be manufactured.

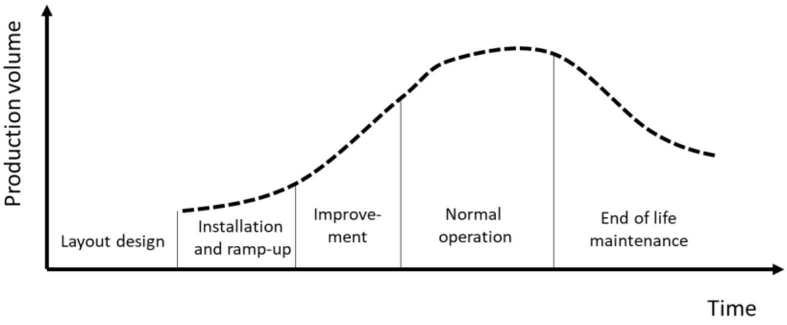

Figure 1. *The life cycle of a manufacturing line*

The Fig. 2 shows that in the case of flexible production systems, several types of products are continuously manufactured, while these products may be in different phases of their life cycle. In addition, the manufactured product mix may depend on several parameters: seasonality, changes in customer demand, difficulties in purchasing raw materials, etc. These challenges all require the flexibility of the production line [5].

During the life cycle of the production line, different products are produced at different periods. Thus, the sum of those quantity curves is the actual number of pieces to be produced. Similarly, the profit curve also adds up. In the case of profit, the goal is to be able to maintain the expected profit level during the entire life cycle of the production line [6].

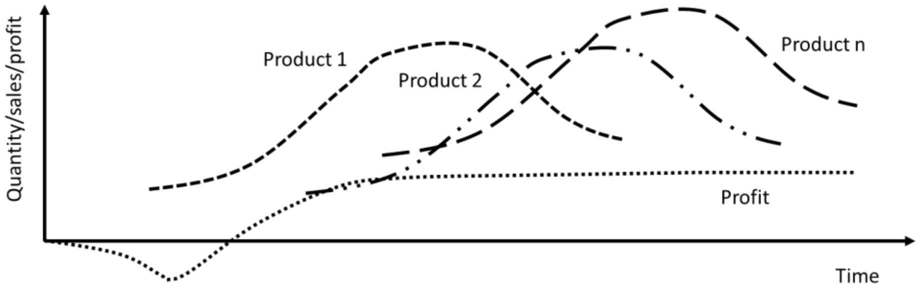

Figure 2. *Manufacturing product mix with products in a different lifecycle stage*

If we want to support the decisions on a production line with some kind of AI tool, then we will soon be faced with the specialties resulting from the life cycle of the production lines. As mentioned earlier, production lines are constantly developing and changing. Accordingly, the characteristics of the data generated on them can also change significantly. For example, in the case of an average electronics production line, a new product appears every 2 months, and the production parameters change every 1 month due to some process improvement. The production line simply does not have a long enough stable state of data collection that can be used to train an AI model. From a data collection point of view, we can already talk about a new production line every 1-2 months.

Fig. 1 contains the classic four-element product life cycle curve of the production lines, supplemented at the beginning with the layout design phase. The life cycle curve shown in the figure clearly shows that if we are talking about real data collection, we can start it only when the line is installed and ramped up. It means that larger amounts of data are only available during the development period. From the point of view of data quality, the quality of the data collected during the ramp-up is often questionable, since in that phase there are usually a lot more extra problems and shutdowns resulting from the start-up of the line. When we want to install some kind of artificial intelligence-based decision-making element in the queue, we quickly fall into a trap, because the AI component needs data to function, but there is no data until the system is working normally. If we want to integrate an AI component into a line that has been operating for a long time, then we are usually in an easier situation. We are in a much better position if we use synthetic data, since even in the design phase of the line we can create synthetic data based on the design data, and during the lifecycle we can continuously improve the data quality. Thus, by the time the line is physically installed, we can already have a working AI model. Fig. 3 shows the availability of the amount of data and their relation to the product life cycle.

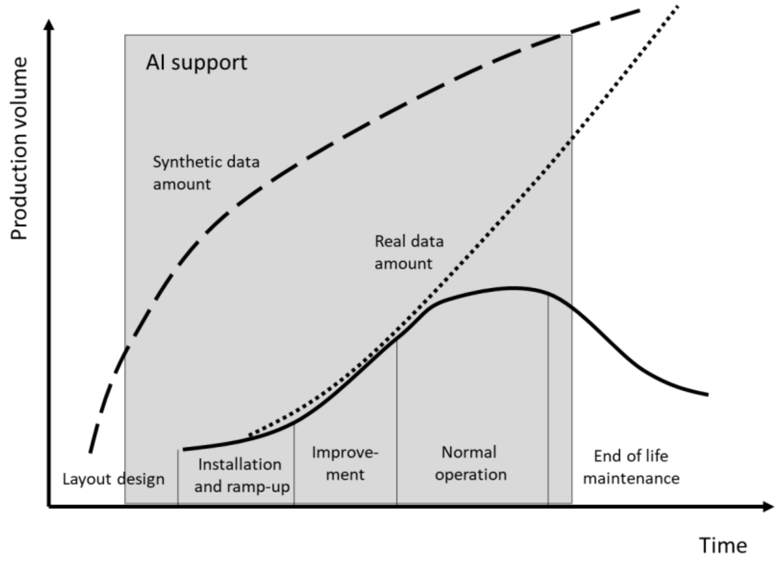

Figure 3. *The availability of the real world collected and the synthetic data during the lifecycle*

Based on Fig. 3, in the traditional way the collection of data from the production line can start from the ramp-up period of the line. At that time, however, the quality of the data is generally not yet suitable for use in machine learning due to biases in it. We can only collect better quality data during the normal operation. However, this is often already too late if we are working on an AI model whose task is to control the production line. The situation is much better for synthetic data. Even in the design phase of the production line, the design data can be used to generate the initial batch of the synthetic data. Later, the quality of the synthetic data can be improved as the planning data becomes more accurate.

The parameters measured on the installed line can further refine the data. Data mixing can also be used during data generation. In data mixing the data package used to train the AI model contains both real measured data and synthetic data. Such mixed data often better characterize the extreme states of the system.

Synthetic data is data generated by some model to replace real data. By controlling the data generation process, the quality of the synthetic data can be better for machine learning purposes. Synthetic data can be created on any scale, quantity, and for any duration. It is essential that the synthetic data reflect the nature, composition, and proportions of the real data. During the preparation of the synthetic data, we have to model the distribution and structure of the real data [7]. The biggest advantage of synthetic data is that we can completely control its creation. Another major advantage of synthetic data is that it usually contains less personal or private data, so the problem of handling personal data (for example GDPR related) appears less often. For example, a real data collection system usually contains which worker performed the work by worker name, while in the synthetic implementation of the same, the workers appear only as serially numbered objects [8].

The most widespread methods of generating synthetic data are the following [9, 10]:

- rule-based methods: this is a well-used method if the generated data must comply with some constraint or rule.
- generative methods: these methods are very common for generating images, for example, where the data is created by processing previous real data and generating new data based on it.
- simulation-based methods: in the case where real systems have to be modelled, a mapping of the real world has to be created (digital twin), this method can be used very well. The data is generated by the simulation that maps the real world and behaves identically to it. In the next part, we examine the concept of the digital twin and, in relation to the research, we choose simulation as a tool for creating synthetic data.
- methods based on data expansion: in this case, the real data is expanded with generated data to have a sufficiently large set of data available. The resulting data will be a mixture of real and generated data.

It is worth comparing the synthetic data with the real data based on some data-specific aspects. Since we work with a large amount of data in production, we can examine how the two data types fit into the 5V model of big data (Table I.).

Table I.

Comparison of the real and synthetic data from the Big Data 5V point of view

Big data parameters	**Real collected data**	**Synthetic data**
Velocity	Constantly increasing data after the line is installed.	Data generation can already start during the line planning.
Volume	The amount depends on the complexity of the process.	The quantity depends on the complexity of the process, more data due to earlier data collection.
Veracity	When the line changes, the data can become invalid or outdated, in which case a new data collection cycle must be started, which gives	The quality and accuracy of the data depends on the accuracy of the digital twin model. Although the data becomes obsolete here as well with the change of

	results after a long time. The line can operate with a poorly trained model during the transition period. In many cases, the data require cleaning and could contain duplication.	the line, new data - the data mapping of the changed situation - can be generated quickly. The data generation process can take care of the data quality.
Variety	The structure of the data depends on the production data collection system, in many cases data come from different systems and in different formats.	The data structure created in the digital twin model is uniform, well-structured. The data generation process is fully controlled.
Value	If the data becomes outdated, their business value will be zero, and even outdated training data can lead to incorrect production and business decisions.	The well-built and constantly updated digital twin.

Gartner already drew attention years ago to the fact that in a short time, it is possible that up to 85% of AI applications and projects will provide inadequate results in the future due to data quality problems [11].

2. DIGITAL TWIN

One of the most widespread methods of creating a digital twin in the case of mapping production processes is discrete event-based simulation (DES) [12]. These systems include building blocks for material flow mapping, production process data management, and usually several tools that are used to evaluate the simulation run results [13].

Most simulation systems have an integrated artificial intelligence component or can be easily integrated with common artificial intelligence systems [14].

The process of the building of the digital twin and using it with an AI application (Fig. 4):

- goal definition and abstraction: defining the goal of the project is crucial. In addition, based on the goal, it is necessary to determine here how detailed a model is needed. The level of detail affects the running speed of the model, and through it the learning speed of the AI model.
- process data collection: collecting the data for the detailed model defined in the previous step. The main data are product, quantity, routing, supporting services, timing (PQRST) [15]. It is important that the stochastic elements in the data and processes are also collected.
- model building or modification: in this step, the simulation model is created, and the AI model is taught by running it.
- verification and validation of the model: this is one of the most important steps since the correct functioning of the AI model depends on the correct functioning of the model. In addition to checking the data, it is also necessary to validate the model with example runs. Of course, this is only possible if the given production line already exists and works in physical reality. This step can be performed several times during the life cycle.
- data collection from the model: in this step, the model is run with the appropriate input data and the data required for model training is collected based on them. In

principle, the quality of the data is sufficient to eliminate all distortions and be used without editing.

- AI model training: training the model. Most simulation and digital twin building tools provide some sort of neural network component that can be trained.
- using of the AI model: the trained model is used in this step, first in the digital twin and then implemented in the real system.
- handling the change: as we wrote earlier, production is constantly changing. In the event of a change, it must be examined whether the change may have an effect on the operation of the model and, through this, on the data generated by the model. If the answer is no, then the existing AI model can continue to be used. If the answer is yes, then you have to go back to the model building phase and make the modifications. Of course, the changed model must be validated and verified in the same way before the new teaching cycle begins. The updated AI model and application can be used after the new teaching cycle.

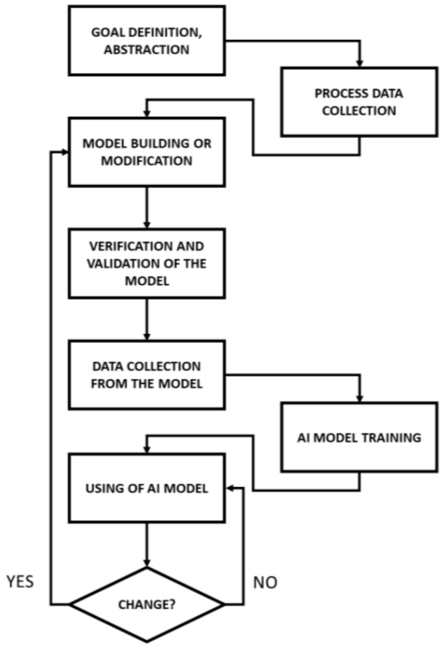
GOAL DEFINITION, ABSTRACTION
PROCESS DATA COLLECTION
MODEL BUILDING OR MODIFICATION
VERIFICATION AND VALIDATION OF THE MODEL
DATA COLLECTION FROM THE MODEL
AI MODEL TRAINING
USING OF AI MODEL
YES
NO
CHANGE?

Figure 4. *The process of the creation of an AI enhanced digital twin*

3. APPLICATION EXAMPLE

In the example below, we use the discrete event-based simulation system from Siemens called Plant Simulation [16, 17]. The application example is based on one of the built-in examples of Plant Simulation, a loop layout flexible production line.

It is a well-known fact that in the case of pallet-based production lines, two problems hinder the better performance of the system [2, 3]:

- too many pallets, when the pallets become congested, and the system therefore loses capacity

- too few pallets, when the pallets do not reach the next station in time, and the system therefore loses capacity.

In the case of production lines with pallets, it is critical to determine the correct number of pallets [18].

The production line in our example produces 6 different products (P01-P06) in a product mix, as customer demand vary greatly. It may happen that zero units of product P01 need to be produced one day, and the other products make up the full production volume. And on the next day, it is conceivable that product P01 accounts for the entire production volume. Since the production times of the product types at the automatic stations differ significantly, the number of pallets required for maximum output may also differ depending on the product mix.

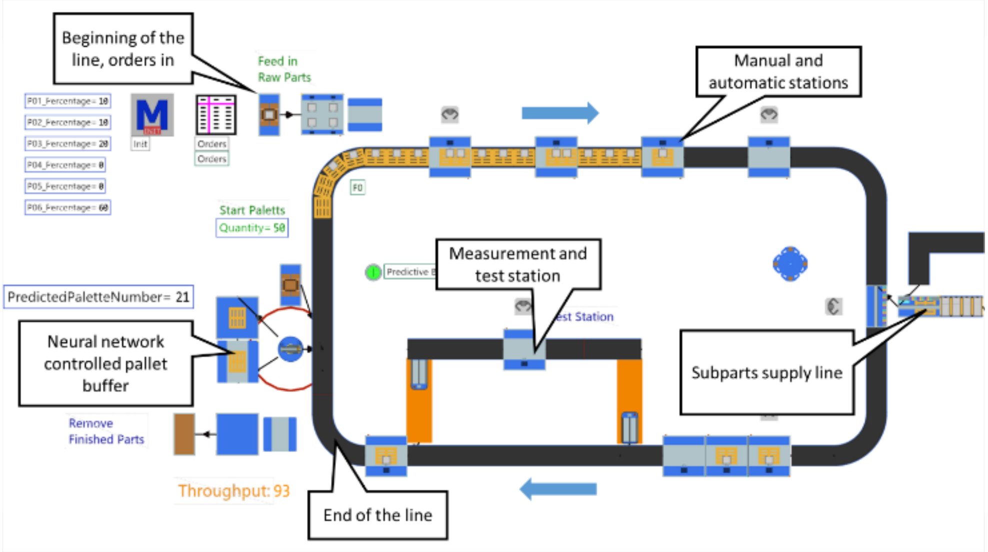

Figure 5. *Sample flexible manufacturing line and its main components*

Main characteristics of the system:

- there are automatic and manual stations on the production line.
- in the case of manual stations, the production time has a deviation.
- in the case of automatic stations, the production time depends on the product type.
- a one-piece flow occurs, the batch size is 1
- daily planning is done on the line, the product mix is specified with a given percentage value on a daily basis.
- 50 pallets are available on the line.
- the goal is to maximize the number of units produced per day.

The critical element of the system is the number of pallets, so it was decided that at the end of the line, after the finished part removal, there will be an element used to control the number of pallets. In each case, this calculates how many pallets the line can work best with, based on the product mix on the current line situation, and accordingly to that removes the pallet that arrived at the control point from the line, or, if necessary, adds a new pallet to the line from the pallet buffer. To determine the ideal number of pallets for product mixes, a large number of model runs would be required.

If we want all possible combinations, so that the 6 products take values between 0 and 100, but in such a way that their sum is 100, we should handle the following number of combinations and run the following number of experiments:

$$\textit{Number of combinations} = C(n+k-1, k-1) = C(109, 4)$$

where:

n – represents the target sum,
k – represents the number of numbers.

The calculation equation is:

$$C = n! / k! * (n - k)! \qquad \textit{where } 0 <= k <= n$$

In the case of combinations, we must count on repeated combinations since the products are different and have different production parameters.

Plugging in the values C = 96,560,646 and multiplied by the number of possibilities of the pallet number (between 20 and 50 there are 31 steps to analyse), the possible number of experiments = 2,993,380,026. In addition, in the case of simulations, it is recommended to perform several runs (observations) in one case to take into account the stochastic process characteristics. With observation runs, the real run number is 5-10 times higher. Even with today's modern computers and distributed or cloud execution capabilities, this is a significant number. If we use a neural network, and we run just some distinct cases, this number can be significantly reduced.

To train the neural network, we run a limited number of cases, which are generated by combining the product mix between 0 and 100 in 10% increments. The sum of the product mix of course must be 100%. In this case, the number of combinations = C (n+k-1, k-1) = C(19, 4). Plugging in the values: C = 3 003 combinations.

In the example, in the case of pallet numbers, the pallet number range between 20-50 was examined in steps of 5. Thus, a total of 3,003*7 = 21,021 runs were needed to generate the neural network baseline.

Table II.

Neural network settings

Setting	Value
Number of training steps	200
Hidden layer dimension	7
Activation function	Sigmoid
Input	Product mix (6 numbers) and the pallet number
Output	Number of produced parts

The neural network was trained with the basic data shown in Table 2. The learning curve is shown in Figure 6. It can be seen that after 160 teaching cycles, the value of the error is low, it does not decrease further significantly, and in addition, the value of the maximum error has also decreased to an acceptable level. After 200 training steps the training was finished at an error of 2.531%.

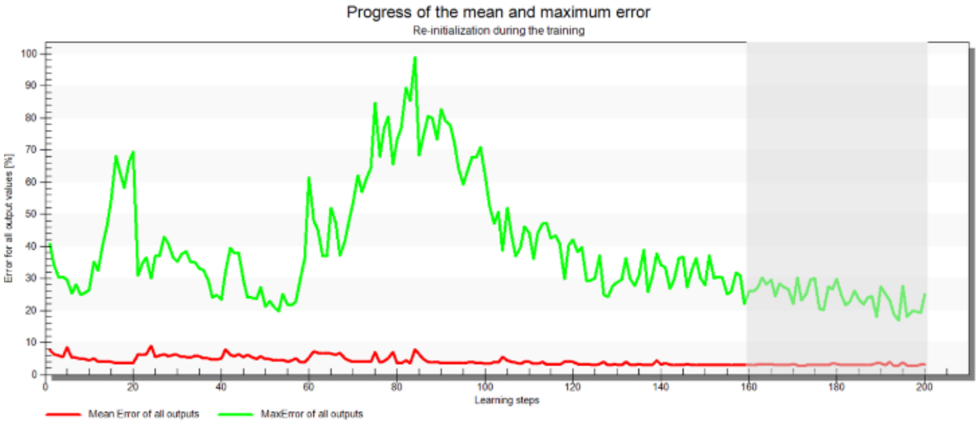

Figure 6. *The progress of the neural network training.*

We integrated the neural network created in this way into the model and ran the original parameter mixes (the 21,021 combinations) again. The result is that in more than 25% of the cases we can produce more with the continuously optimized number of pallets, and in the remaining 75% we do not produce less. For example, in the case of the 10%, 10%, 20%, 0%, 0%, 60% product mix, 73 can be produced with a fixed pallet number, while 93 can be produced in one day with the neural network-controlled system. This is a significant improvement of the line.

The developed neural network-based decision system can be used as long as

- no new product launches on the line
- there are no changes in the production parameters (operation time, machine parameters, layout...)

If such a change does occur, then the digital twin model must be modified in accordance with the changes and the training data required for the neural network must be recreated.

4. Summary and conclusion

The article examines the impact of changes in the data underlying artificial intelligence on the control of production processes. It’s clearly visible that due to frequent system changes in production, the collected data cannot be used or can only be to a limited extent. Therefore, by mapping the system, a digital twin must be built, from which synthetic data can be collected. The article presents the process through a concrete industrial example.

The novelty of the method and the example is that it connects the life cycle of the production line with the life cycle of the digital twin. This is an exciting approach that will help to ensure that production lines can be used for longer in the future, thereby reducing investment and maintenance costs.

References

[1] Molnár, Zs., Tamás, P. & Illés, B. (2021). The Life cycle of the layouts of flexible and reconfigurable manufacturing areas and lines. *Advanced Logistic Systems – Theory and Practice*, 15(1), 20-29, https://doi.org/10.32971/als.2021.003

[2] Shivanand, M. B. V. K. H. K. (2006). *Flexible Manufacturing System*. New Delhi: New Age International Limited

[3] P. E. J. (1983). *Flexible Manufacturing Systems: An Overview*. Fall Industrial Engineering Conference, American Institute of Industrial Engineers, 639-645,

[4] S. K. R. (1986). *Flexible Manufacturing Systems: An Industry Overview*. Production and inventory management, Washington, D. C., 27(4), 1-9,

[5] Abanda, F.H., Jian, N., Adukpo, S. et al. (2024). Digital twin for product versus project lifecycles' development in manufacturing and construction industries. *J Intell Manuf*. https://doi.org/10.1007/s10845-023-02301-2

[6] Ullah I. & Narain, R. (2020). Achieving mass customization capability: the roles of flexible manufacturing competence and workforce management practices. *Journal of Advances in Management Research*, **18**(2), 273-296, http://doi.org/10.1108/JAMR-05-2020-0067, 2020.

[7] Jordon, J. et al. (2022). *Synthetic Data - what, why and how?* The Alan Turing Institute, The Royal Society, London.

[8] Rabaev, M., Pratama, H. & Chan, K. C. (2024). Leveraging Synthetic Data and Machine Learning for Shared Facility Scheduling. In: Ullah, A., Anwar, S., Calandra, D., Di Fuccio, R. (eds) Proceedings of International Conference on Information Technology and Applications. ICITA 2022. *Lecture Notes in Networks and Systems*, 839, Springer, Singapore. https://doi.org/10.1007/978-981-99-8324-7_34

[9] El Emam, K. et al. (2020). *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly Media

[10] Kerim, A. (2023). *Synthetic Data for Machine Learning: Revolutionize your approach to machine learning with this comprehensive conceptual guide*. Packt Publishing

[11] Gartner: *Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence.* Retrieved from https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence. [Accessed 10 04 2024].

[12] Sujová, E., Střihavková, E. & Čierna, H. (2018). An Analysis of the Assembly Line Modernization by Using Simulation Software. *Manufacturing Technology*, **18**(5), 839-845, https://doi.org/10.21062/ujep/187.2018/a/1213-2489/MT/18/5/839

[13] Vishnupriya, B., Cheng, C. & Jaime, C. (2024). Enhancing Manufacturing Operations with Synthetic Data: A Systematic Framework for Data Generation, Accuracy, and Utility. *Frontiers in Manufacturing Technology*, **4**, https://doi.org/10.3389/fmtec.2024.1320166

[14] Chan, K. C., Rabaev, M. & Pratama, H. (2022). Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation. *Production &amp; Manufacturing Research*, **10**(1), 337–353. https://doi.org/10.1080/21693277.2022.2086642

[15] Muther, R. & Hales, L. (2015). *Systematic Layout Planning.* Management & Industrial Research Publications, ISBN 978-0-933684-06-5

[16] Bangsow, S. (2020). Tecnomatix Plant Simulation. Modeling and Programming by Means of Examples. Second Edition, Springer, https://doi.com/10.1007/978-3-030-41544-0

[17] Siemens, *Plant Simulation version 2302 Help*, Siemens, 2023.

[18] Zubair, M. (1994). *Flexible Manufacturing Systems: Planning Issues and Solutions*. Garland Publishing Inc.