big data and cognitive computing

MDPI

Article

Cumulative and Rolling Horizon Prediction of Overall Equipment Effectiveness (OEE) with Machine Learning

Péter Dobra [1,*] and János Jósvai [2]

[1] Doctoral School of Multidisciplinary Engineering Sciences, Széchenyi István University, 9026 Győr, Hungary
[2] Department of Vehicle Manufacturing, Széchenyi István University, 9026 Győr, Hungary; josvai@ga.sze.hu
* Correspondence: dobra.peter@sze.hu

Abstract: Nowadays, one of the important and indispensable conditions for the effectiveness and competitiveness of industrial companies is the high efficiency of manufacturing and assembly. These enterprises based on different methods and tools systematically monitor their efficiency metrics with Key Performance Indicators (KPIs). One of these most frequently used metrics is Overall Equipment Effectiveness (OEE), the product of availability, performance and quality. In addition to monitoring, it is also necessary to predict efficiency, which can be implemented with the support of machine learning techniques. This paper presents and compares several supervised machine learning techniques amongst other polynomial regression, lasso regression, ridge regression and gradient boost regression. The aim of this article is to determine the best estimation method for semiautomatic assembly line and large batch size. The case study presented with a real industrial example gives the answer as to which of the cumulative or rolling horizon prediction methods is more accurate.

Keywords: machine learning; prediction; OEE; rolling horizon

check for updates

Citation: Dobra, P.; Jósvai, J. Cumulative and Rolling Horizon Prediction of Overall Equipment Effectiveness (OEE) with Machine Learning. *Big Data Cogn. Comput.* **2023**, *7*, 138. https://doi.org/10.3390/bdcc7030138

Academic Editor: Min Chen

Received: 25 May 2023
Revised: 26 July 2023
Accepted: 31 July 2023
Published: 2 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Nowadays, various forecasting tools and techniques are playing an increasingly important role in industrial manufacturing companies in order to fulfill customer orders on time. In addition to traditional estimation methods, such as various trends, simulations of more effective techniques supported by machine learning have also appeared [1–4]. This is also true for Overall Equipment Effectiveness (OEE), the most frequently used efficiency Key Performance Indicator (KPI) in the domain of assembly operations [5]. The prediction of manufacturing and assembly efficiency is relevant, among others, in the fields of production planning, scheduling, investments and management decisions [6,7]. Accurate forecasting generates profit, reliability and competitive advantage for industrial enterprises.

Due to the development of information technology, many systems provide support for the recording, processing and storage of production-related data. The most frequently applied systems are Manufacturing Execution System (MES), Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) [8,9]. Currently, it is a constant challenge to reveal the patterns and relationships behind the data, which is made more difficult by the turbulent industrial environment and changes in product variances [10,11]. With the help of machine learning, it is possible to determine the expected and searched values faster, more efficiently and more accurately, including the OEE percentages. Due to the huge amount of data, machine learning is one of the best methods for processing production data.

Overall Equipment Effectiveness is measured by the machines' capability of performing a task to produce a finished product as per customer needs in a timely manner. The prediction of OEE can also be conducted in the case of machine learning in several ways, even by predicting each component (availability, performance and quality) separately. All three types of machine learning, such as supervised learning, unsupervised learning and reinforcements learning, can be used for forecasting in the field of production [12–14].

The aim of this paper is to determine the best estimation method for semiautomated assembly line and large lot size by finding out whether cumulative or rolling horizon forecasting is more accurate. In recent years, this has not been analyzed in detail, so several machine learning methods are examined and compared in this article. Considering that there are many machine learning approaches, this paper does not examine all of them but, rather, tries to outline a general state.

2. Materials and Methods

A production system and, within it, the efficiency of assembly operations can be evaluated in many ways [15]. The most common method in automotive practice is to use Key Performance Indicators (KPIs) [16]. Overall Equipment Effectiveness (OEE) as a standard and best practice indicator was introduced within the Total Productive Maintenance concept by Nakajima [17]. The original formula for calculation of OEE is written as:

$$OEE = a\ p\ q \quad (1)$$

where:

a—availability (%);
p—performance (%);
q—quality (%).

At the domain of press-hardening process in manufacturing, Lejon et al. used three machine learning methods, such as autoencoder neural network (ANN), one-class support vector machine (OCSVM) and isolation forest (IF), for anomaly detection. The ANN method was the best performing candidate in a similar dataset based on precision, recall and accuracy [18]. Fast angle-based outlier detection (FABOD) and K-nearest neighbor (KNN) methods performed better compared to histogram-based outlier score (HBOD), local outlier factor (LOF), isolation forest (IF) and one-class support vector machine (OCSVM) techniques when examining and detecting anomalies occurring on the assembly lines [19].

A quality factor affecting OEE was investigated by Peres et al. with machine learning classifiers. In the frame of multistage quality control, the following algorithms were examined: Gaussian naïve Bayes, K-nearest neighbor, XGBoost, random forest, support vector machine and logistic regression. After the evaluation of accuracy, recall, precision and F1 score, the XGBoost tuned model performed the best [20].

A hybrid prediction model was proposed to estimate whether the automotive assembly process is functioning normally or abnormally. The conception utilizes an outlier detection based on density-based spatial clustering of applications with noise and random forest classification model. The presented model achieved higher accuracy than other examined models, such as naïve Bayes, logistic regression and multilayer perceptron [21].

Wang et al. applied support vector machines algorithm to estimate the quality of welding in a high-power disk layer. The proposed quality control system worked in real-time mode [22]. Regarding the quality factor, Lee et al. compared four prediction methods as decision tree, random forest, artificial neural network and support vector machine at the area of metal casting. The best accuracy was at the ANN with 93.84% [23].

Predicting the OEE value at a production line with six machines, decision tree regression algorithm was the more robust and had the best result in terms of mean square error than K-nearest neighbors, support vector machine and artificial neural network [24]. El Mazgualdi et al. presented the use of various machine learning algorithms under different configuration to predict OEE of an automotive wiring factory. Support vector regression (SVR), support vector regression cross-validation (SVRCV), support vector regression genetic algorithm (SVRGA), random forest (RF), random forest cross-validation (RFCV), extreme gradient boost (XGB), extreme gradient boost cross-validation (XGBCV) and deep learning (DL) were evaluated based on mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE). It was concluded that neither method was effective enough due to the small size of training data [25].

In terms of transmission assembly line predicting OEE, using Bayesian ridge regression shows higher accuracy (99%) than other machine learning methods among others such as adaptive naïve Bayes-based algorithm (96%), logistic regression model (84%), support vector machine (97%) and decision tree (89%) [26,27]. However, it is worth using more detailed minimum daily or shift-level basic data than the monthly OEE percentages used here. Martinez et al. attempted to estimate OEE by separately predicting availability, performance and quality. A new approach was taken in which they tried to combine the predictions of completely different machine learning algorithms. It was a new ensemble which combines the best of each method, is able to operate in all the cases, and reduces the error [28].

Khdoudi et al. compared four machine learning methods to predict welding process parameters. It was concluded that the convolutional neural network (CNN) was the most accurate for energy prediction, the support vector regression (SVR) model for amplitude prediction, and the regression model for pressure prediction. The best prediction method was a combination of machine learning techniques [29,30].

A deep reinforcement learning framework was proposed using three methods as deep Q-network, proximal policy optimization and advantage actor–critic algorithms for ensuring the product quality and minimizing the overall energy consumption of an industrial glass manufacturing process [31]. In the area of quality forecasting, 14 algorithms were compared (ridge regression, linear regression, light gradient boosting machine, lasso regression, random forest regressor, artificial neural networks, gradient boosting regressor, extra trees regressor, elastic net, Bayesian ridge, K neighbors regressor, AdaBoost regressor, least angle regression and orthogonal matching pursuit); the ridge regression algorithm presented the best overall predictive performance for the test examples [32].

Zouhri et al. used polynomial, sigmoid and (radial basic function (RBF) kernels as genetic-based SVM for chemical and rolling process quality data classification. The RBF kernel function was the recommended for classification with 87.15% (chemical data) and 99.08% (rolling process data) accuracy [33].

Based on the above-mentioned scientific literature, it can be seen that many methods can be used to predict OEE percentages; however, rolling horizon and the application of the cumulative method are not mentioned.

Selection of Independent Variables for Machine Learning

In this article, the authors examine the estimation of OEE from a prediction point of view, so this is the dependent variable. The basis for the selection of the dependent variables is a cause-and-effect diagram published in a previous article [34]. The factors affecting OEE are analyzed according to the following six aspects: man, environment, method, material, machine and measurement. A total of 150 influencing factors were revealed and described. Further examining the independent factors, Table 1 as an excerpt shows which can be measured and recorded with MES, ERP, SQL or log files. In addition to all this, it shows an example and indicates the effect of the factors on availability, performance and quality. Appendix A contains more details.

Based on the measurability and occurrence of the factor characteristics of assembly lines, the following 12 independent variables were selected: process failure downtime, break downtime, technical downtime, changeover downtime, quality reason downtime, logistics reason downtime, not planned downtime, other downtime reason, number of changeover, average cycle time, number of assembled units and number of scrap units. OEE percentage, availability percentage, performance percentage and quality percentage are considered as dependent variables. These variables will be used in the next chapters. The basis for selecting the independent variables was that they should be characteristic of each assembly line, be objectively measurable and occur in large quantities. In addition, it was important not to select too many independent variables.

Table 1. Based on Ishikawa diagram, an excerpt of OEE contributors and measurability of attributes.

Attribution		Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Qualification	Practical experience	Yes	Variable cycle time per station per person		x	
Motivation	Goals	Yes	Availability, performance, quality, OEE target		x	
Organization	Improvement	Yes	Trends (OEE, scrap, etc.)		x	
Production technology	Assembly process	Yes	Manual or nonmanual (automated) assembly		x	
Measurement, control	Maintenance	Yes	Downtime, reason: maintenance	x		
Work process	Process parameters	Yes	Time and duration data		x	
Material and information flow	Available workforce	Yes	Staff (operator, setter, etc.)	x		
Material failure	Material quality	Yes	Downtime, reason: quality problems			x
Material handling	Not available, not accessible	Yes	Downtime, reason: logistics problem	x		
Machine and tool adjustment	SMED, OTED	Yes	Downtime, reason: changeover	x		
Product control	Sampling frequency	Yes	Number of checked products			x
Checking of assembly process	SPC	Yes	In MES: SPC report			x

The elements of the complete Ishikawa diagram and effects on OEE are presented in Tables A1–A6.

3. OEE Prediction with Machine Learning

After presenting the real work environment and the data used, this chapter examines the possibilities of OEE prediction using different machine learning methods. The authors are aware that there are many machine learning techniques; however, the described elements give a complete picture of the prediction processes.

3.1. The Real Work Environment of Applied Machine Learning

This article illustrates the presented methods through a real industrial example. The selected semiautomatic assembly line is a metal seat structure assembly line for the automotive industry located in Central Europe. Figure 1 shows the simplified layout of the hybrid line.

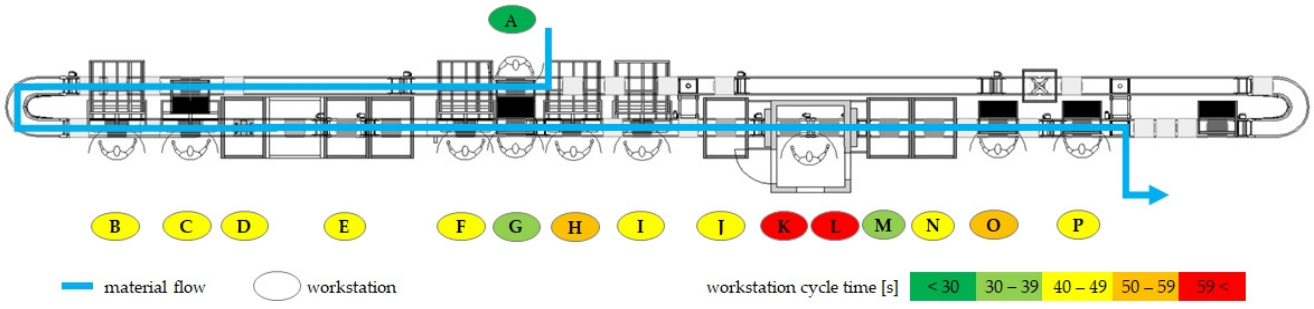

Figure 1. Semiautomatic assembly line layout with cycle time groups.

The assembly line consists of 16 work stations named with letters from A to P. The blue arrow shows the direction of material flow. The cycle time of each station is marked with different colors. Assembly operations are performed by human or machines according to Table 2.

Table 2. Human and machine operations at the hybrid assembly line.

Assembly Station	Operation by	Assembly Station	Operation by
A	human	I	human
B	human	J	machine
C	human	K	human
D	machine	L	human
E	machine	M	machine
F	human	N	machine
G	human	O	human
H	human	P	human

This production unit has been operating for more than 10 years and, in two shifts, produces generally 380–440 products per shift. Since the beginning of the production, the assembly line has produced the same main product; there is no significant difference between the product variations.

In order to accurately describe the production environment, it is first necessary to calculate and determine the type of production process based on Equation (2) and Table 3 [35].

$$T_f = \frac{QT}{I_{pr}} \quad (2)$$

where:

T_f—production rate;
Q—annual volume of the production task (unit/year);
T—working time requirement of the production task assigned to the production unit (hour/unit);
I_{pr}—actually available productive time base (hour/year).

Table 3. Production types and rates.

Types of Production Process	Production Rate
Mass production	$0.8 < T_f \leq 1.0$
Large batch size	$0.6 < T_f \leq 0.8$
Medium batch size	$0.4 < T_f \leq 0.6$
Small batch size	$0.2 < T_f \leq 0.4$
Job-shop or project production	$0 \leq T_f < \leq 0.2$

Based on these, the analyzed hybrid assembly line with T_f values (0.76, year 2021; 0.78, year 2022) can be classified as large batch class.

3.2. Production Data for Machine Learning

All the data related to the semiautomatic assembly line mentioned in the previous subsection are available with the support of the factory MES and the SQL database system. The production data used are real, accurate and continuous. A set of assembly data can be considered as Big Data. Every second, hundreds of data are generated on the line, which are stored in different systems, for example, a separate system handles process data and a separate system handles product-specific data or logistics barcode data. During the processing and sorting of the data of these systems, 769 records are created and one record contains the data of an entire eight-hour shift. Assembly operations run in two shifts, so that means 10 records per week. The data of the examined period come from the years 2021 and 2022 and are scaled with no factors. In addition, the extreme or outlier values are not excluded so that the real industrial environment is depicted as well as possible. In the following, the goal is only to predict the OEE values; the individual components (availability, performance and quality) are not examined separately. The values of the

dependent variable are shown in Figure 2. The OEE values are shown on the vertical axis and the individual records are shown in chronological order on the horizontal axis.

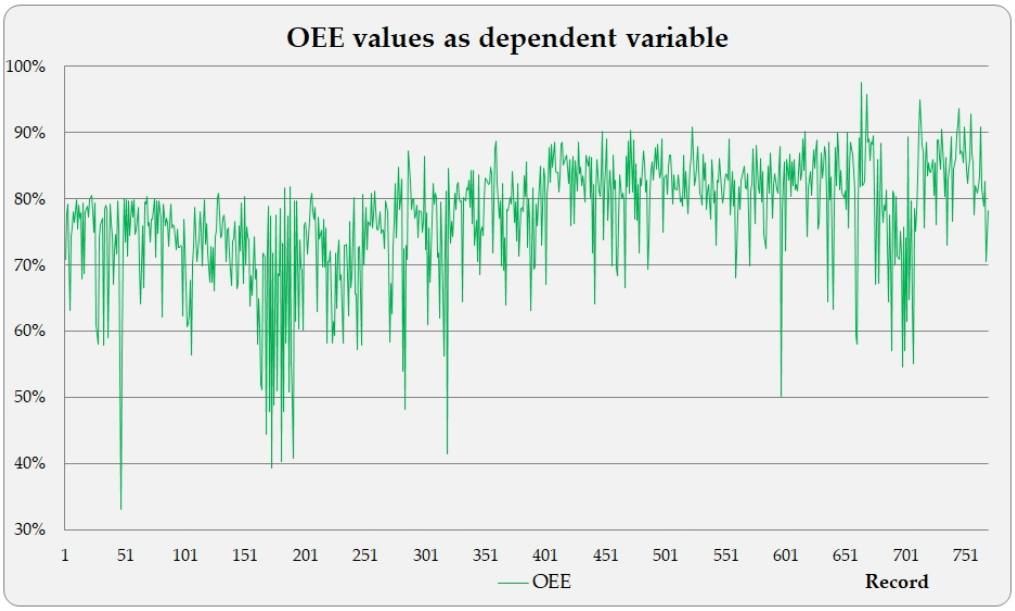

Figure 2. OEE values as dependent variable.

3.3. Cumulative and Rolling Horizon Prediction

During machine learning prediction of OEE, the following two main cases can be distinguished according to how the training period is selected:

- Cumulative approach;
- Rolling horizon approach;
  - Fix rolling horizon;
  - Changing rolling horizon.

In the cumulative case, the amount of training data increases continuously over time so more and more data are used as part of the training set. This has the advantage that all occurring data are taken into account; however, overfitting can occur. In the rolling horizon approach, the selected training data use a different time window for each prediction, which moves forward continuously. In general, rolling horizon prediction pushes the time window; cumulative does not.

3.4. Applied Machine Learning Methods

The authors used R and RStudio program for the entire research work. The following machine learning methods were used for OEE prediction:

- Multiple linear regression (MLR);
- Polynomial regression (simple) (Pol 1);
- Polynomial regression (complex) (Pol 2);
- Lasso regression (Lasso);
- Ridge regression (Ridge);
- Random forest regression (RF);
- Gradient boost regression (GB);
- Mixed GAM computation vehicle with automatic smoothness estimation regression (MGCV).

The basis of the selection was to choose methods within the R environment that can also be used in industrial conditions without requiring a large computing capacity. Two main considerations guided the authors' choice of machine learning methods. The first aspect was the selection of the most suitable and widely used learning methods for the

analysis of production time series. The selection was based on an extensive literature analysis. The second main criterion was the applicability in an industrial environment; where possible, the authors preferred procedures with a relatively low computational demand and a predictable time requirement for their application in industrial practice. A separate Rstudio program was written for each of the selected machine learning methods, and these contain the most frequently used settings. The following R packages were used: xlsx, xml, readxl, tidyverse, ggplot2, readr, openxls, chron, lubridate, dplyr, caTools, corrplot, doParallel, randomForest, MASS, mlbench, Amelia, plotly, reshape2, caret, moments and mgcv.

In the following chapter, the cumulative and fix rolling horizon approaches were used for each machine learning method.

4. Discussion

OEE was predicted for two periods (testing set), a shorter period (6 weeks, 60 records), which is the simulation of production planning, and a longer period (15 weeks, 150 records), which can help estimate the necessary industrial investments. In the first case, the rolling is performed on a weekly basis (10 records), while, in the second case, it is performed on a monthly basis (50 records). These periods are derived from automotive practice, because assembly plans are reviewed and production scheduling is conducted weekly. On the other hand, it is sufficient to monitor the OEE data required for industrial investments every month for the period of the next three months. Industrial investment means whether it is necessary to convert, rent, build or buy a new machine or production line in order to meet customer needs. In this article, several training periods were examined; in the case of production planning, the range is from 10 to 100 in increments of 10, while, in the case of industrial investment, the range is from 100 to 200 also scaled by 10.

In each case, the evaluation and comparing of the predicted results were carried out using root mean squared error (RMSE) based on the following equation:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\bar{y}_i - y_i)^2}{n}} \quad (3)$$

where:

n—number of fitted points;
y_i—actual value;
$\bar{y}_i$—predicted value [36].

For the different regressions applied by machine learning, the RMSE values for production planning are shown in Table 4, while Table 5 is for industrial investments.

Table 4. RMSE values for production planning (short time prediction).

			RMSE								
Step	**Training**	**Testing**	**MLR**	**Pol 1**	**Pol 2**	**Lasso**	**Ridge**	**RF**	**GB**	**MGCV**	**Min**
10	10	60	0.3737	0.0587	0.1294	0.0786	0.0721	0.0701	-	-	0.0587
10	20	60	0.0881	0.0553	0.0924	0.0682	0.0676	0.0655	-	0.0882	0.0553
10	30	60	0.0635	0.0501	0.0777	0.0569	0.0582	0.0621	-	0.0635	0.0501
10	40	60	0.0545	0.0543	0.0644	0.0519	0.0534	0.0599	-	0.0545	0.0519
10	50	60	0.0516	0.0541	0.0611	0.0513	0.0524	0.0577	-	0.0516	0.0513
10	60	60	0.0493	0.0541	0.0580	0.0498	0.0517	0.0564	0.0587	0.0493	0.0493
10	**70**	**60**	0.0490	0.0543	0.0573	**0.0489**	0.0507	0.0549	0.0570	0.0490	**0.0489**
10	**80**	**60**	0.0489	0.0546	0.0569	**0.0489**	0.0513	0.0540	0.0559	0.0489	**0.0489**
10	90	60	0.0510	0.0547	0.0580	0.0497	0.0513	0.0536	0.0540	0.0491	0.0491
10	100	60	0.0514	0.0550	0.0581	0.0505	0.0516	0.0529	0.0524	0.0492	0.0492
	Avg.		0.0881	**0.0545**	0.0713	0.0555	0.0560	0.0587	0.0556	0.0559	**0.0545**
10	Cum.	60	0.0496	0.0603	0.0604	0.0503	0.0507	0.0498	**0.0472**	0.0496	0.0472

Table 5. RMSE values for industrial investments (long time prediction).

			RMSE								
Step	**Training**	**Testing**	**MLR**	**Pol 1**	**Pol 2**	**Lasso**	**Ridge**	**RF**	**GB**	**MGCV**	**Min**
50	100	150	0.0597	0.0573	0.0588	0.0555	0.0593	0.0573	0.0542	0.0597	0.0542
50	110	150	0.0619	0.0576	0.0593	0.0600	0.0631	0.0560	0.0551	0.0619	0.0551
50	120	150	0.0524	0.0565	0.0573	0.0525	0.0567	0.0528	0.0518	0.0524	0.0518
50	130	150	0.0513	0.0570	0.0578	0.0516	0.0557	0.0514	0.0494	0.0513	0.0494
50	140	150	0.0498	0.0563	0.0565	0.0513	0.0539	0.0514	0.0495	0.0498	0.0495
50	150	150	0.0496	0.0571	0.0565	0.0513	0.0540	0.0514	0.0515	0.0496	0.0496
50	160	150	0.0506	0.0585	0.0585	0.0524	0.0546	0.0522	0.0507	0.0506	0.0506
50	170	150	0.0504	0.0576	0.0576	0.0519	0.0540	0.0488	0.0472	0.0504	0.0472
50	180	150	0.0497	0.0571	0.0569	0.0513	0.0531	0.0471	0.0456	0.0497	0.0456
50	**190**	**150**	0.0482	0.0562	0.0559	0.0503	0.0512	0.0476	**0.0450**	0.0482	**0.0450**
50	200	150	0.0489	0.0570	0.0566	0.0513	0.0522	0.0476	0.0456	0.0489	0.0456
	Avg.		0.0520	0.0571	0.0574	0.0527	0.0553	0.0512	**0.0496**	0.0520	0.0496
50	Cum.	150	**0.0544**	0.0617	0.0619	0.0562	0.0558	0.0562	0.0562	**0.0544**	0.0544

The following conclusions can be made about predictions in the case of production planning:

- In terms of the examined training periods, simple polynomial regression (Pol 1) showed the best average RMSE result with a fix rolling horizon (0.0545), followed by lasso (0.0555) and gradient boost regression (0.0556), although, with GB, the 10–50 training period cannot be interpreted due to the specificity of the method;
- In the case of a rolling horizon, the lowest RMSE value (0.0489) can be achieved with lasso regression set to the training parameters 70 and 80;
- Analyzing the cumulative methods, the gradient boost regression showed the best RMSE result (0.0472), followed by multiple linear regression (0.0496) and MGCV (0.0496);
- Considering the average speed of calculation, the fastest method is simple polynomial regression (4.7 s), followed by multiple linear regression (5.1 s) and random forest (9.9 s). The individual values are shown in Table 6. (The laptop used with Intel ® Celeron ® CPU N2840 @ 2.16 GHz 2.16 GHz, 4.00 GB RAM);
- The RMSE values of the examined cumulated methods are generally better than those of the fix rolling horizon methods;
- In the case of production planning, the recommended OEE prediction method is the gradient boost cumulated approach due to the most accurate RMSE value and the moderately long calculation requirement.

Table 6. Average calculation time of each method.

Applied Regression	**Production Planning**		**Industrial Investments**	
	Roll. Hor.	**Cum.**	**Roll. Hor.**	**Cum.**
Multiple linear regression	5.1 s	7.0 s	**2.0 s**	5.0 s
Polynomial regression (simple)	**4.7 s**	**6.0 s**	2.7 s	3.9 s
Polynomial regression (complex)	56.3 s	70.3 s	14.4 s	15.0 s
Lasso regression	546.9 s	765.9 s	86.3 s	117.4 s
Ridge regression	17,727.3 s	18,906.0 s	2758.2 s	3506.2 s
Random forest regression	9.9 s	33.9 s	4.0 s	6.8 s
Gradient boost regression	511.2 s	1016.7 s	94.1 s	152.8 s
MGCV	10.5 s	13.3 s	2.9 s	**3.7 s**

The following conclusions can be made about predictions in the case of industrial investments:

- Regarding the examined training periods, the gradient boost regression showed the best average RMSE result with fix rolling horizon (0.0496), followed by random forest (0.0512), multiple linear regression (0.0520) and MGCV with the same result (0.0520);
- In the case of fix rolling horizon, the lowest RMSE value (0.0450) can be achieved with gradient boost regression set to the training parameters 190;
- Analyzing the cumulative methods, the multiple linear regression showed the best RMSE result (0.0544), followed by MGCV (0.0544) and ridge regression (0.0558);
- Considering the average speed of calculation, the fastest method is multiple linear regression (2.0 s), followed by simple polynomial regression (2.7 s) and MGCV (2.9 s);
- The RMSE values of the examined fix rolling horizon methods are always better than the cumulated approach;
- In the case of industrial investment, the recommended prediction method is the gradient boost fix rolling horizon approach with 180–200 training sets, due to the most accurate RMSE value and the medium–long calculation requirement.

The machine learning methods used have the advantage of being able to estimate the real production time series under study with an average RMSE rate of approximately 4–7%. Based on our analyses, it is clear that the so-called over-learning does not occur for the methods under investigation. A weakness identified is the inaccuracy of the ability to react to sudden changes in the real world, which is an important and pronounced feature in the real environment. Based on our observations, the range of the appropriate learning set is quite wide, and it is difficult to clearly determine based on these real-world data sets how many shift cycles are required for machine learning methods to achieve an RMSE prediction value of 5% or below.

During the regressions presented so far, in addition to the fix rolling horizon (10 records for production planning, 50 records for investment) and fix testing period (60 records for production planning, 150 records for investment), only the training periods were modified. It follows that, in addition to these, there are many variation options and even smaller RMSE values can be achieved if the testing periods are changed. As an example, for multiple linear regression, Figure 3 shows the additional options.

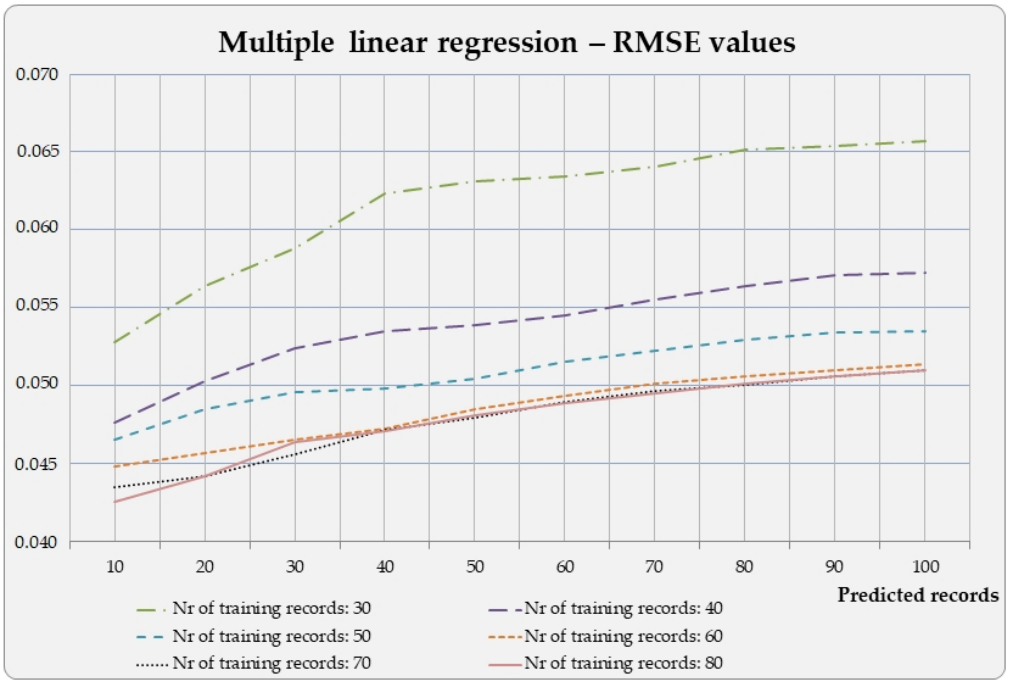

Figure 3. Different RMSE values in case of multiple linear regression.

The predicted or testing records are shown on the horizontal axis and the RMSE values on the vertical axis. The different colored lines show the number of used training records. Figure 3 shows that, in all cases, if the amount of testing data is reduced, the RMSE value decreases. For production planning, the use of the figure provides guidelines as to how much prediction accuracy the used training set will result in during the testing period. Based on Figure 3, the main conclusion is that multiple linear regression fits better with more training records and less testing data. However, it is important to monitor the RMSE values, depending on how accurately it is necessary to predict the OEE values.

5. Conclusions

At the area of the manufacturing industry, it is essential that the follow up of assembly operations' performance and efficiency takes in real-time mode and the estimation works quickly and reliably. Based on supervised machine learning, assembly efficiency metrics, e.g., Overall Equipment Effectiveness (OEE) can be predicted, but it is not clear by which method.

This paper collected and evaluated the factors as initial information or data affecting OEE from different perspectives. The individual elements were examined according to whether they affect availability, performance or quality and how they can be measured by Manufacturing Execution System (MES), SQL query or another way. Based on the measurability and occurrence of the factor characteristic of assembly lines, 12 independent variables were selected among others, including process failure downtime, technical downtime, changeover downtime, average cycle time and number of assembled units. OEE percentage, availability percentage, performance percentage and quality percentage are considered as dependent variables.

Based on the selected data, the prediction of OEE was analyzed using different machine learning methods, such as multiple linear regression, simple and complex polynomial regression, lasso regression, ridge regression, random forest regression, gradient boost regression and MGCV regression. Using real industrial assembly line data, each machine learning method has been demonstrated for production planning and investment. Two approaches were used, fix rolling horizon and cumulative way. The results were evaluated using root mean squared error (RMSE) and the computation time.

In the case of production planning, the recommended OEE prediction method is the gradient boost cumulated approach due to the most accurate RMSE value and the moderately long calculation requirement. In the case of industrial investment, the recommended prediction method is the gradient boost fix rolling horizon approach with 180–200 training sets, due to the most accurate RMSE value and the medium–long calculation requirement. In addition to these, further optimization possibilities were outlined using the extended testing period, which was demonstrated with multiple linear regression.

The authors plan to examine further machine learning methods and search for additional optimization options to predict the dependent variable even more accurately. In the future, it will be possible to combine the cumulative approach with the fixed rolling horizon way for a lower value of RMSE.

Author Contributions: Conceptualization, P.D.; methodology, P.D.; software, P.D.; validation, P.D.; formal analysis, P.D.; investigation, P.D.; resources, P.D. and J.J.; data curation, P.D.; writing—original draft preparation, P.D.; writing—review and editing, P.D.; visualization, P.D.; supervision, J.J.; funding acquisition, P.D. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Széchenyi István University, Győr, Hungary, Publication Support Program.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. The Ishikawa diagram "Man" effect on OEE contributors and measurability of attributes.

Attribution		Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Qualification	Education	No	-			
	Internal training	Yes	Other downtime, reason: planned training	x		
	Special knowledge	No	-			
	Practical experience	Yes	Variable cycle time per station per person		x	
	External training	No	-			
Skills, abilities	Seeing	No	-			
	Hearing	No	-			
	Fine movement	No	-			
	Stamina	No	-			
	Communication skills	No	-			
Personality, character	Punctuality	Yes	Number of assembled bad products			x
	Speed	Yes	Variable cycle time per station per person		x	
	Compliance	No	-			
	Monotony tolerance	No	-			
	Conscientious work	No	-			
Motivation	Goals	Yes	Availability, performance, quality, OEE target		x	
	Expectations of employee	No	-			
	Rewards, condemnations	No	-			
	Team, company	No	-			
	Work conditions	No	-			
Organization	Available staff	Yes	Other downtime, reason: missing operator	x		
	Improvement	Yes	Trends (OEE, scrap, etc.)		x	
	Leading, management	No	-			
	Support	No	-			
	Planning	Yes	Production and resource plan, scheduling		x	

Table A2. The Ishikawa diagram "Environment" effect on OEE contributors and measurability of attributes.

Attribution		Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Work environment	Safety	No	-			
	Health protection	No	-			
	Ergonomics	No	-			
	Perceived environment	No	-			
	Automatization	Yes	Manual or nonmanual (automated) assembly		x	
Production environment	Complexity of technology	Yes	Assembly lines, stations, products		x	
	5S	Yes	Downtime, reason: cleaning	x		
	Concerns of technology	No	-			
	Production without waste	Yes	Downtimes, scrap	x		
	Visual support	Yes	OEE Andon board		x	

Table A2. *Cont.*

	Attribution	Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Market environment	Oder	Yes	Production and resource plan, scheduling		x	
	Takt time, cycle time feasibility	Yes	Takt time, cycle time data		x	
	Pull system	Yes	Production plan, batch size, changeover, scheduling		x	
	Competition	No	-			
	Production plan feasibility	Yes	Fulfilment of production plan		x	
Company environment	Company, team	Yes	Staff (operator, setter, etc.)		x	
	Shift schedule, breaks	Yes	Downtime, reason: changeover	x		
	Overtime	Yes	Production and resource plan, scheduling		x	
	Motivation, commitment	No	-			
	Employee expectations	No	-			
Worker environment	Social situation	No	-			
	Plant availability	No	-			
	Social acceptance	No	-			
	Benefits	No	-			
	Expectations of employer	No	-			

Table A3. The Ishikawa diagram "Method" effect on OEE contributors and measurability of attributes.

	Attribution	Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Production technology	Assembly process	Yes	Manual or nonmanual (automated) assembly		x	
	Repair, rework process	Yes	Downtime, reason: quality problems			x
	Checking process	Yes	Cycle time of checking station		x	
	Packaging process	Yes	Cycle time of packaging station		x	
	Automatization	Yes	Manual or non-manual (automated) assembly		x	
Measurement, control	Maintenance	Yes	Downtime, reason: maintenance	x		
	SPC, 100% checking	Yes	In MES: SPC report			x
	Six sigma	No	-			
	Failure analysis, PDCA, Pareto	Yes	Pareto analysis			x
	Poka yoke	Yes	Downtime, reason: Poka yoke check			x
Work process	Standard operational procedure	Yes	In MES: documents module		x	
	Process parameters	Yes	Time and duration data		x	
	Material workflow	No	-			
	Best practices	Yes	In MES: documents module		x	
	Planned cycle time	Yes	Cycle time data per stations		x	
Lean methods	Goals	Yes	Availability, performance, quality, OEE target		x	
	Expectations	No	-			
	Rewards, condemnations	No	-			
	Company, team	No	-			
	Work conditions	No	-			
Material and information flow	Available workforce	Yes	Staff (operator, setter, etc.)	x		
	Improvement	Yes	Trends (OEE, scrap, etc.)		x	
	Leading, organization	No	-			
	Support	No	-			
	Planning	Yes	Production and resource plan, scheduling		x	

Table A4. The Ishikawa diagram "Material" effect on OEE contributors and measurability of attributes.

Attribution		Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Material failure	Material quality	Yes	Downtime, reason: quality problems			x
	Surface	Yes	Downtime, reason: quality problems			x
	Deficiency	Yes	Downtime, reason: quality problems			x
	Surplus	Yes	Downtime, reason: quality problems			x
	Color	Yes	Downtime, reason: quality problems			x
Size error	Width, length, height	Yes	Downtime, reason: quality problems			x
	Diameter	Yes	Downtime, reason: quality problems			x
	Out of tolerance	Yes	Downtime, reason: quality problems			x
	Deformation	Yes	Downtime, reason: quality problems			x
	Position problem	Yes	Downtime, reason: quality problems			x
Quantitative error	Too much, too little	Yes	Downtime, reason: logistics problem	x		
	Not available	Yes	Downtime, reason: logistics problem	x		
	Stuck together	Yes	Downtime, reason: quality problems			x
	Mixed	Yes	Downtime, reason: quality problems			x
	Batch failure	Yes	Downtime, reason: quality problems			x
Material handling	Damaged	Yes	Downtime, reason: logistics problem	x		
	Contaminated	Yes	Downtime, reason: logistics problem	x		
	Temperature, warranty	Yes	Downtime, reason: logistics problem	x		
	Not available, not accessible	Yes	Downtime, reason: logistics problem	x		
	Not identified	Yes	Downtime, reason: logistics problem	x		
Design failure	Function problem	Yes	Downtime, reason: design failure			x
	Comfort problem	Yes	Downtime, reason: design failure			x
	Not controllable, not repairable	No	-			
	Not durable	No	-			
	Not or difficult to assemble	Yes	Significantly different cycle time, takt time		x	

Table A5. The Ishikawa diagram "Machine" effect on OEE contributors and measurability of attributes.

Attribution		Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Maintenance	Maintenance plan	Yes	Downtime, reason: maintenance	x		
	Fulfilment of planned maintenance	Yes	Downtime, reason: maintenance	x		
	Assemble ability, repairability, maintainability	Yes	Downtime, reason: maintenance	x		
	Standard parts	No	-			
	Manuals, drawings	No	-			
Machine and tool adjustment	Adjusted, validated	Yes	Downtime, reason: settings	x		
	Setting documentation	No	-			
	Simple, fast, standardized	Yes	Downtime, reason: settings	x		
	Failure catalogue available	No	-			
	SMED, OTED	Yes	Downtime, reason: changeover	x		
Stability	Trouble proof	No	-			
	Energy supply continuous	Yes	Downtime, reason: technical issue	x		
	Machine capability, process capability	Yes	Significantly different cycle time, takt time		x	
	Operable	Yes	Operating and non-operating time	x		
	Reliable PC, PLC network	Yes	Downtime, reason: technical issue	x		

Table A5. *Cont.*

	Attribution	Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Standardization	Standard parts, components	No	-			
	Spare parts available	Yes	Downtime, reason: maintenance	x		
	Available documents	No	-			
	Parts of MES	Yes	Assembly line and machines exist in MES		x	
	Traceable, identifiable	Yes	In MES: traceability reports		x	
Safety	Can be stopped, interruptible	No	-			
	Ergonomic	No	-			
	Free of hazardous materials	No	-			
	Easy to handle	No	-			
	Not accidental	No	-			

Table A6. The Ishikawa diagram "Measurement" effect on OEE contributors and measurability of attributes.

	Attribution	Can Be Measured by MES, ERP, SQL and Log Files	Example	A	P	Q
Material checking	Quantitative control	Yes	Number of scrap pieces			x
	Incoming inspection	Yes	Number of scrap pieces			x
	Quality control	Yes	Number of scrap pieces			x
	Traceability	Yes	In MES: traceability reports		x	
	Function checking	Yes	Number of scrap pieces			x
Product control	Sampling frequency	Yes	Number of checked products			x
	Sampling size	Yes	Number of checked products			x
	Checking functional operation	Yes	Scrap or good products			x
	Sampling place	Yes	Checking station in MES			x
	Documentation control	Yes	In MES: documents module		x	
Machine, tool checking	Appropriate frequency	Yes	Downtime, reason: maintenance	x		
	Machine testing	Yes	Downtime, reason: maintenance	x		
	Checking maintenance	Yes	Downtime, reason: maintenance	x		
	Calibration	Yes	Downtime, reason: maintenance	x		
	Safety control	Yes	Downtime, reason: maintenance	x		
Checking of assembly process	SPC	Yes	In MES: SPC report			x
	Simulation control	Yes	Downtime, reason: settings	x		
	First and last product control	Yes	Timestamp data of first and last products			x
	Type change checking	Yes	Timestamp data of changeover			x
	Poka yoke control	Yes	Timestamp data of Poka yoke check			x
Measurement instruments checking	Functionality	Yes	Timestamp data of Poka yoke check			x
	Reliability	No	-			
	Accuracy	No	-			
	Frequency	Yes	Timestamp data of Poka yoke check			x
	Documents control	Yes	In MES: documents module		x	

References

1. Backus, P.; Janakiram, M.; Mowzoon, S.; Runger, C.; Bhargava, A. Factory cycle-time prediction with a data mining approach. *IEEE Trans. Semicond. Manuf.* **2006**, *19*, 252–258. [CrossRef]
2. Choudhary, A.K.; Harding, J.A.; Tiwari, M.K. Data mining in manufacturing: A review based on the kind of knowledge. *J. Intell. Manuf.* **2009**, *20*, 501–521. [CrossRef]
3. Tao, F.; Qi, Q.; Liu, A.; Kusiak, A. Data-driven smart manufacturing. *J. Manuf. Syst.* **2018**, *48*, 157–169. [CrossRef]
4. Nagorny, K.; Lima-Monteiro, P.; Barata, J.; Colombo, W.A. Big data analysis in smart manufacturing: A review. *Int. J. Commun. Netw. Syst. Sci.* **2017**, *10*, 31–58. [CrossRef]
5. Muchiri, P.; Pintelon, L. Performance measurement using overall equipment effectiveness (OEE): Literature review and practical application discussion. *Int. J. Prod. Res.* **2008**, *46*, 3517–3535. [CrossRef]

6. Harding, J.A.; Shahbaz, M.; Srinivas, S.; Kusiak, A. Data-Mining in manufacturing: A review. *J. Manuf. Sci. Eng.* **2006**, *128*, 969–976. [CrossRef]
7. Wuest, T.; Weimer, D.; Irgens, C.; Thoben, K.D. Machine learning in manufacturing: Advantages, challenges, and applications. *Prod. Manuf. Res.* **2016**, *4*, 23–45. [CrossRef]
8. Mantravadi, S.; Moller, C. An overview of next-generation Manufacturing Execution System: How important is MES for Industry 4.0? *Procedia Manuf.* **2019**, *30*, 588–595. [CrossRef]
9. Beregi, R.; Pedone, G.; Háy, B.; Váncza, J. Manufacturing Execution System integration through the standardization of a Common Service Model for Cyber-Physical Production Systems. *Appl. Sci.* **2021**, *11*, 7581. [CrossRef]
10. Wiendahl, H.P.; Reichardt, J.; Nyhuis, P. *Handbook Factory Planning and Design*; Springer: Berlin/Heidelberg, Germany, 2015.
11. ElMaraghy, H.; Schuh, G.; ElMaraghy, W.; Piller, F.; Schönleben, P.; Tseng, M.; Bernand, A. Product variety management. *CIRP Ann.* **2013**, *62*, 629–652. [CrossRef]
12. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
13. Zhou, Z.H. *Machine Learning*; Springer Nature: Singapore, 2021. [CrossRef]
14. Rebala, G.; Ravi, A.; Churiwala, S. *An Introduction to Machine Learning*; Springer International Publishing: Cham, Switzerland, 2019. [CrossRef]
15. Oliveira, R.; Taki, S.A.; Sous, S.; Salimi, M.A. Global Process Effectiveness: When Overall Equipment Effectiveness Meets Adherence to Schedule. *Procedia Manuf.* **2019**, *38*, 1615–1622. [CrossRef]
16. Corrales, L.C.; Lambán, M.P.; Korner, M.E.H.; Royo, J. Overall Equipment Effectiveness: Systematic literature review and overview of different approaches. *Appl. Sci.* **2020**, *10*, 6469. [CrossRef]
17. Nakajima, S. *Introduction to TPM: Total Productive Maintenance*; Productivity Press: Cambridge, UK, 1988.
18. Lejon, E.; Kyösti, P.; Lindström, J. Machine learning for detection of anomalies in press-hardening: Selection of efficient methods. *Procedia CIRP* **2018**, *72*, 1079–1083. [CrossRef]
19. Abdelrahman, O.; Keikhosrokiani, P. Assembly line anomaly detection and root cause analysis using machine learning. *IEEE Access* **2020**, *8*, 189661–189672. [CrossRef]
20. Peres, R.S.; Barata, J.; Leitao, P.; Garcia, G. Multistage quality control using machine learning in the automotive industry. *IEEE Access* **2019**, *7*, 79908–79916. [CrossRef]
21. Syafrudin, M.; Alfian, G.; Fitriyani, N.L.; Rhee, J. Performance analysis of IoT-based sensor, Big Data processin, and Machine Learning model for real-time monitoring system in automotive manufacturing. *Sensors* **2018**, *18*, 2946. [CrossRef]
22. Wang, T.; Chen, J.; Gao, X.; Qin, Y. Real-time monitoring for disk laser welding based on feature selection and SVM. *Appl. Sci.* **2017**, *7*, 884. [CrossRef]
23. Lee, J.H.; Noh, S.D.; Kim, H.J.; Kang, Y.S. Implementation of Cyber-Physical Production System for quality prediction and operation control in metal casting. *Sensors* **2018**, *18*, 1428. [CrossRef]
24. Souza, B.V.; Santos, S.R.B.; Oliveira, A.M.; Givigi, S.N. Analyzing and predicting Overall Equipment Effectiveness in manufacturing industries using machine learning. In Proceedings of the 2022 IEEE International Systems Conference (SysCon), Montreal, QC, Canada, 25–28 April 2022; pp. 1–8. [CrossRef]
25. EL Mazgualdi, C.; Masrour, T.; El Hassani, I.; Khdoudi, A. Machine learning for KPIs prediction: A case study of the overall equipment effectiveness within automotive industry. *Soft Comput.* **2021**, *25*, 2891–2909. [CrossRef]
26. Imane, M.; Aoula, E.S.; Achouyab, E.H. Using Bayesian Ridge Regression to predict the Overall Equipment Effectiveness performance. In Proceedings of the 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, 3–4 March 2022; pp. 1–4. [CrossRef]
27. Imane, M.; Aoula, E.S.; Achouyab, E.H. Support Vector Regression to predict the Overall Equipment Effectiveness indicator. In Proceedings of the 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 18–22 May 2022; pp. 1–5. [CrossRef]
28. Bonada, F.; Echeverria, L.; Domingo, X.; Anzaldi, G. AI for improving the Overall Equipment Efficiency in manufacturing industry. In *New Trends in the Use of Artificial Intelligence for the Industry 4.0*; Martínez, L.R., Rios, R.A.O., Prieto, M.D., Eds.; IntechOpen: London, UK, 2020. [CrossRef]
29. Khoudi, A.; Masrour, T. Prediction of industrial process parameters using artificial intelligence algorithms. *arXiv* **2019**, arXiv:1812.10537v2. [CrossRef]
30. Khdoudi, A.; Masrour, T.; El Mazgualdi, C. Using machine learning algorithms for the prediction of industrial process parameters based on product design. *Adv. Intell. Syst. Sustain. Dev.* **2019**, *3*, 728–749. [CrossRef]
31. El Mazgualdi, C.; Masrour, T.; Barka, N.; El Hassani, I. A learning-based decision tool towards smart energy optimization in the manufacturing process. *Systems* **2022**, *10*, 180. [CrossRef]
32. Khdoudi, A.; Barka, N.; Masrour, T.; El Hassani, I.; El Mazgualdi, C. Online prediction of automotive tempered glass quality using machine learning. *Res. Sq.* **2022**. [CrossRef]
33. Zouhri, W.; Rostami, H.; Homri, L.; Dantan, J.Y. A genetic-based SVM approach for quality data classification. In *Artificial Intelligence and Industrial Applications*; Springer International Publishing: Cham, Switzerland, 2021. [CrossRef]
34. Dobra, P.; Jósvai, J. Overall Equipment Effectiveness (OEE) complexity for semi-automatic automotive assembly lines. *Acta Polytech. Hung.* **2023**, *20*, 63–82. [CrossRef]

35. Kardos, K.; Jósvai, J. *Gyártási Folyamatok Tervezése*; István Széchenyi University: Győr, Hungary, 2005. (In Hungarian)
36. Igual, L.; Segui, S. *Introduction to Data Science, a Python Approach to Concepts, Techniques and Applications*; Springer International Publishing: Cham, Switzerland, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.