
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 2, pp. 788–795, DOI 10.12694/scpe.v26i2.3898

THE PREDICTION AND EVALUATION OF MANUFACTURING TECHNOLOGY INNOVATION BASED ON MACHINE LEARNING AND BIG DATA ANALYSIS

FANG YANG*

Abstract. A data anomaly detection method was designed based on chemical manufacturing and oil refining units. Massive data storage and calculation are used in the cloud computing framework for petrochemical enterprises, refineries, and other large enterprises. The massive data is segmented based on the modified time series method, and the anomaly analysis is carried out. Thus, the abnormal data of the chemical manufacturing and oil refining units can be monitored. The practice proves that the algorithm proposed in this paper is a feasible, simple and effective data correction scheme.

Key words: Chemical manufacturing; Data correction; Machine learning; Fault error detection; Modified timing method

1. Introduction. Process data in the petroleum refining industry are generally flow-related operating parameters, such as flow, concentration, temperature and other field operating parameters. Under normal circumstances, accurate process test data is an essential basis for operation analysis, improvement of production process control and improvement of factory production and management [1]. However, the test results obtained in practical applications often contain some randomness and errors, which are inconsistent with the characteristics of the manufacturing process itself. The error of process test data can be divided into two kinds: one is random error, and the other is fault error. The random deviation in the system is caused by the noise of the test signal and the random variation during operation. Fault error is caused by unforeseen circumstances such as instrument failure, inaccurate calibration or reference drift, equipment leakage, etc. Data correction aims to improve the reliability, accuracy and integrity of data in process production to provide high-quality data for the production and management of process enterprises [2]. Although there are relatively complete commercial product development applications, they are still based on conventional statistical testing and linear analysis methods. It mainly focuses on error finding, data correction and parameter estimation. There are significant defects in its practical application: first, there is no credible reference standard for error detection, which leads to weak recognition and easy-to-miss diagnosis. Second, the process's data correction and parameter estimation are too dependent on the process structure and spatial information, and the process history information is not effectively mined. Third, the algorithm takes too long, making it challenging to realize the real-time correction of the measured data.

In this paper, the problems of data classification, error correction, data correction, and so on are deeply studied, as well as their organic integration with conventional test data correction [3]. In this way, the shortcomings of routine test data correction are solved. This method is suitable for data correction in the manufacturing process of large petrochemical enterprises.

2. Data processing methods.

2.1. Classification of measurement data. The measurement data correction must be based on the redundancy of the process variables. Only the remaining measurement data type and the observation type's non-measurement variables are corrected [4]. Due to many measured variables and constraints in industrial production, it isn't easy to correct and estimate them, so it is necessary to divide the process parameters to reduce the scale of problem-solving. The existing test data division method based on the zero-matrix method is limited in the complex industrial production process. According to the basic theory of graph theory, some scholars have established the sorting method, which does not require matrix operation and saves a lot of storage

*School of Economics and Management, Weinan Normal University, Weinan 714099, China (wnuyangfang@163.com)

space [5]. It is suitable for classifying more complex processes. In the case of no observed data, the equilibrium constraint equation with no observed value is obtained to modify the data directly.

2.2. Fault error detection. The measurement data error can be divided into two categories: random error and negligent error. Before the correction, if the error information cannot be found and excluded, the correction and estimation results will not be able to reflect the actual situation [6]. Therefore, the error correction should be done before the correction of observation data and parameter estimation. In addition, the failure of measuring instruments or pipeline leakage and other factors resulted in human error. Therefore, the conclusion of error discovery can help the operator to maintain the measuring instrument better and can resolve the problem of the instrument operation in time [7]. There are three ways to investigate errors: theoretically analyzing all kinds of data that may cause errors, using various measurement methods to realize the measurement comparison of the same process parameters, and verifying based on the statistical properties of the test data.

The error detection method based on mathematical statistics has a high application value based on the statistical characteristics of measurement data [8]. However, there are significant limitations in practical application. The traditional investigation methods include global inspection, node inspection and measurement data inspection.

Scholars mostly use the MT-NT combined test method to solve the defects of a single test. The idea is to combine the strengths of both. For example, principal component analysis can accurately judge the error orientation but often gives too much fault error. While NT does not spread the error throughout the system, the risk of error is more significant. So, the two can complement each other [9]. In the existing methods of correction of measured data, linear and nonlinear data correction are often treated separately. The standard correction method is linear correction, and nonlinear correction is used for flow rate, temperature and other data. In fact, according to the fundamental needs of data correction, the higher the redundancy of data, the better. If only the linear method is used to correct the data, the number of limiting equations required is limited, so the accuracy of the calculation results is not good. If only the nonlinear iterative method is used for correction, it will consume a lot of iterative operations. In this way, the real-time correction of the measured data of the equipment cannot be realized [10]. Due to the use of a separate limiting equation, no flow data is involved in the correction process, so the result is not accurate enough. This project intends to adopt two methods: linear correction and nonlinear correction. The velocity data after linear correction is used as the initial value, and then nonlinear iterative correction is carried out to minimize the number of iterations and calculation speed. In this way, the data can be corrected in real-time.

3. Time series analysis is applied to data correction. A prerequisite for revising process measurement data is to have some degree of redundancy. There's a lot of redundancy in process systems. This excess information can be divided into two types: one is caused by the presence of connections in the process, called "spatial redundancy," and the data obtained from multiple tests with the same precision instrument at the same measuring point is called "time redundancy." Time limits are domain limits. Equality limits are value limits. The above data correction algorithms are based on spatial redundancy [11]. A real-time database with high reliability is established for petrochemical enterprises using numerical control technology, which can collect and store the data. From the principle of data correction and effective use of information resources, we must consider spatiotemporal redundancy and spatiotemporal redundancy [12]. Therefore, this project intends to use the improved timing analysis method to correct the process data based on the timing characteristics and improve the accuracy of error discovery and data correction.

The improved time series analysis method is a smoothing algorithm based on robust local weighted regression used to analyze time series data. The time series $F = (K, U)$ is divided into three parts: trend component P, periodic component Z, and residual component S. Here $U = \{u_1, \ldots, u_n\}, u_n$ is the n time node; Where $U = \{u_1, \ldots, u_n\}, u_n$ is the data associated with time n.

$$F = P + Z + S$$

A prediction method based on the trend component is proposed. Periodic components can reflect periodic fluctuations in frequency. Residuals are the components that remain after removing the trend and cyclical

components [13]. The improved time series analysis method includes two aspects. It is divided into the outer cycle and the inner cycle. The direction component P and the periodic component Z are obtained by smoothing the timing of F in the inner cycle. The remaining components were collected in the outer cycle section. Assuming that the size of the point (t_l, u_l) in the time series is r_z, then the weight ω_j^l of t_j at any time is calculated in the range with u_i as the core and r_z as the interval:

$$\omega_j^l = \left(1 - \left(\frac{|u_j - u_j|}{u_{\text{farthot}} - u_i}\right)^3\right)^3$$

u_{Farthest} is the point in the region furthest from u_i. Take u_i as a linear regression at any time in this interval and find a smooth curve $f' = \alpha + \beta t$, then the smooth value at time point t_i is f_i. After a given interval length r_z, the timing of F can be decomposed to obtain the corresponding subsequence. After smoothing the subsequence, a periodic subsequence Z' can be obtained, and then the frequency component S is obtained by a low-pass filter, and the periodic component is expressed by $\mathrm{Z} = Z' - S$. And then keep going:

$$P' = F - Z$$

P' is smoothed at intervals of r_q, and a trend component P is obtained. And then, the remaining component is denoted by $S = F - P - Z$. Assuming that the initial data set corresponding to time series $K = \{t_1, \ldots, t_n\}$ is $U = \{u_1, \ldots, u_n\} = \{g(t_1), \ldots, g(t_n)\}$, and assuming that (t_i, u_i) data is missing, then:

$$W(t_i) = g(t_1) + (t_l - t_1)\, g[t_2, t_1] + (t_i - t_1)(t_i - t_2)$$
$$g[t_3, t_2, t_1] + \ldots + (t_l - t_1)(t_l - t_2)\ldots(t_i - t_{n-1})\, g[t_{n-1}, \ldots, t_2, t_1]$$
$$S(t_l) = (t_i - t_1)(t_i - t_2)\ldots(t_i - t_n)\, g[t_{n-1}, \ldots, t_2, t_1]$$
$$W(t_l) = g(t_1) + (t_i - t_1)\, g[t_2, t_1] + (t_i - t_1)(t_i - t_2)$$
$$g[t_3, t_2, t_1] + \ldots + (t_i - t_1)(t_i - t_2)\ldots(t_i - t_{n-1})\, g[t_{n-1}, \ldots, t_2, t_1]$$

$g[t_i, t_j]$ is the first-order differential quotient of $g(t)$ at point t_i, t_j. Where $W(t_i)$ is the Newton interpolation approximation. $S(t_i)$ is a residual function [14]. Define the data set that has been populated with lost values as $U' = \{u_1', \ldots, u_n\}$. The maximum value is $u_{\max}$ and the minimum value is $u_{\min}$. The method of Inormalization is used so that all data values fall within the range of $[0, 1]$. For example:

$$f_i = \frac{u_i^i - u_{\min}}{u_{\max} - u_{\min}}$$

The data set $F = \{f_1, \ldots, f_n\}$ obtained at the end of the pre-processing. The improved time series analysis method eliminates the trend and periodicity components of the sequence. This makes it easier to find outliers and reduce problems such as error alarms caused by outliers [15]. A modified timing method obtained residuals S for the pre-processed data set. S Perform electrostatic protection tests. Here's how it works:

1. Calculate the middle-value M of the residual series data S.
2. Find the deviation of M from the median value.
3. Calculate statistics for each data point in S:

$$R_i = \frac{S_i - \bar{S}}{\text{mad}}$$

   $\bar{S}$ is the sample average.
4. The maximum value in R is statistically treated, and if the value exceeds the critical value ε, it is regarded as an outlier and removed from the time series.
5. Repeat the process (1) to (4).

4. Application of time series analysis method in practice. The improved time series method was tested and evaluated in a refinery's atmospheric and vacuum plant, focusing on the ability of error detection and the influence of time domain value and error size on error correction.

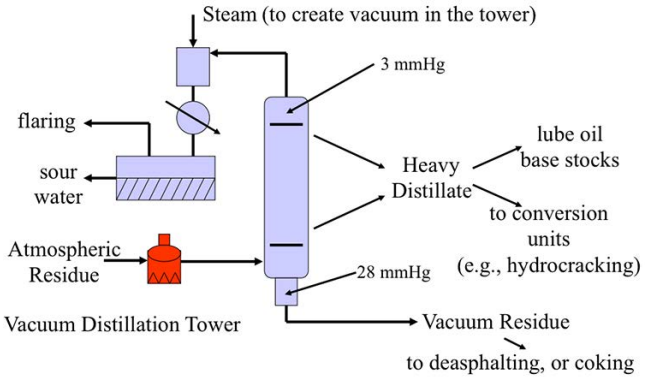

Fig. 4.1: Flow chart of petroleum atmospheric and vacuum device.

4.1. Process Overview. This type of atmospheric and vacuum equipment from a refinery produces oil from an oil depot and sends it to the equipment before heating it through an exchange process and then feeding it to an electric desalination tank. It is desalted and dehydrated in the process. In this process, it is also given a simple flash [16]. The flash-top gas is exchanged with the flash-bottom oil in the atmospheric tower and then heated into the tower for flash. Usually, line, two-line and three-line products are produced from the sideline of the atmospheric tower. The gas is discharged from the standard gas compressor and then through the heat exchange at the bottom of the stabilizing tower to obtain stable gas, liquefied gas, and naphtha. The conventional bottom fraction is divided into four parallel sections and fed into the pressure reducer for heating. It is then fed into a vacuum fractionator for classification. The products of reducing the first, second, third, and fourth lines are produced by the sideline of the decompression tower. The decompression tower extracts the oil at the top of the decompression tower, while the bottom reduction residue is discharged by heat transfer. A simple schematic diagram of the flow structure is obtained through the positioning analysis of each logistic measuring instrument in the equipment (Figure 4.1).

The process data in Fig. 4.1 was classified by sorting rule classification and matrix method. In this way, the simplified process structure of the standard pressure-reducing valve can be obtained [17]. There are 29 flow units and 6 nodes in the whole process.

4.2. Analysis of causes for error discovery. This project intends to establish 500 sets of 29,000 observation samples with the field calibration results of this equipment as actual values and add two 2.5% random deviations of positive and negative values. The random error is added to it to study its detection ability in various cases.

4.2.1. Influence of time domain value on error discovery rate. Select 12-30 errors in the time domain. Add fault errors of 20%, 40%, 60%, 80% and 100% in actual cases. The influence of time domain size on error detection ability is studied [18]. The evolution of the virtual detection rate of fault error over time is shown in Figure 4.2. At 16-20, the improved sequence method has a meager detection rate of fault errors, which can efficiently detect and eliminate fault errors and obtain more reasonable data correction.

4.2.2. Detection of errors of different sizes. As can be seen from the relative deviation between correction values and actual values in Table 4.3, the improved time series method has a good detection of errors of various sizes, and the results are the same as the previous examples.

4.2.3. Error identification in the communication of multiple orders of magnitude. The conventional error detection algorithm cannot locate the error accurately in the minor traffic flow, resulting in a significant deviation between the corrected result and the actual value. The test results show that the improved time series method can accurately detect and exclude. As can be seen from figures 4.4 and 4.5, the correction values are very close to the actual values.

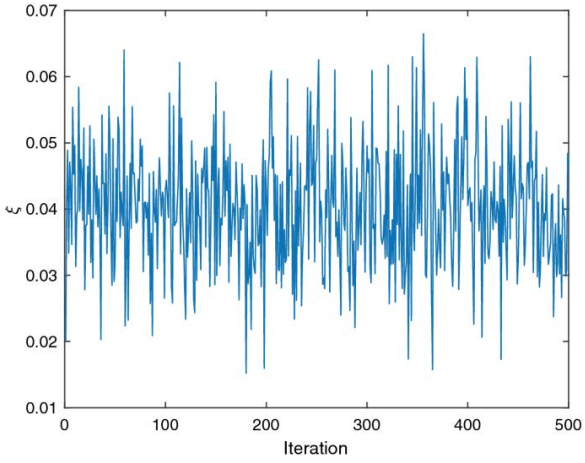

Fig. 4.2: Influence of time domain values on the detection capability of obsolete errors.

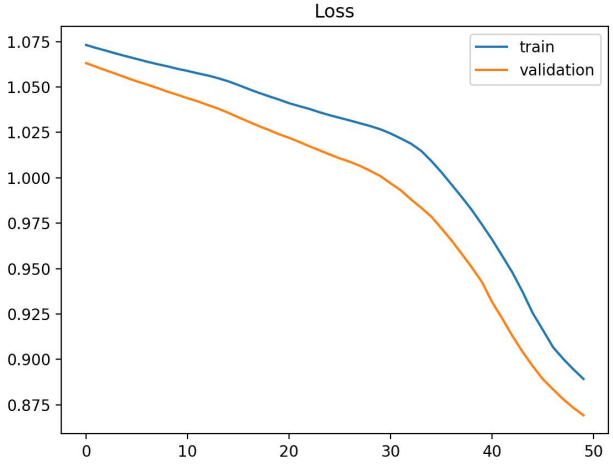

Fig. 4.3: Change of correction accuracy with error size.

4.3. Analysis of results after data modification. The time series method is used to modify 100 measured data. The results of the revised series are shown in Table 4.1.

As shown in Table 4.1, in the test data of the combination, 7,12, and 26 each carry 1 fault error. The average deviation between their correction and actual values is only 1%.

5. Conclusion.

1. According to the accurate division of the chemical production process, the defects of error detection are discussed, and a composite test method with process simulation as the core is established. This allows for better detection of errors.
2. Aiming at the linear and nonlinear problems existing in the system, the joint correction method is studied to increase the redundancy of the observation data. It also improves the speed and accuracy of data correction.
3. The process parameter correction method based on time series analysis is studied to use better the

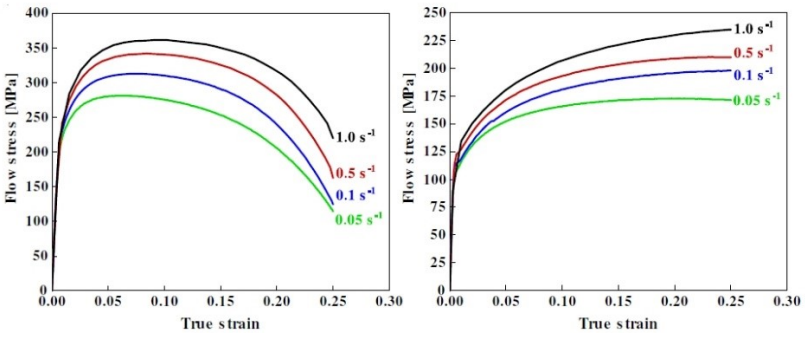

Fig. 4.4: Comparison of the results of a large order of magnitude traffic correction and actual value.

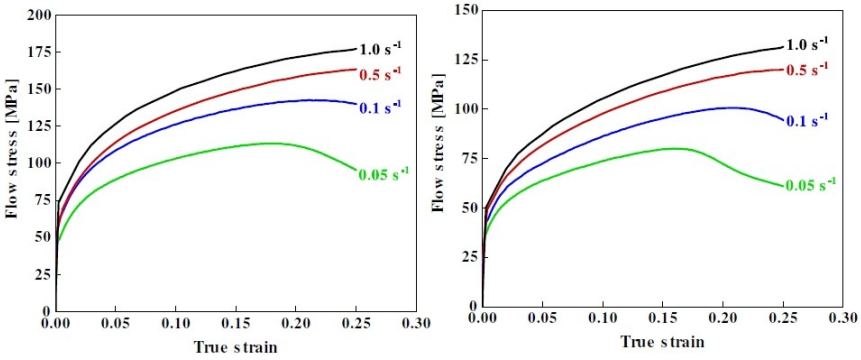

Fig. 4.5: Comparison of the results of a small order of magnitude flow correction and actual value.

massive historical data in the numerical control system. This project intends to use the improved sequence method to analyze observational data in the time domain. This reduces the influence of random error and the threshold of error discovery. The defects of conventional statistical testing methods are solved.

6. Acknowledgement. Shaanxi Provincial Education Department Project (21JK0627): Research on Mechanism and Path of Intellectual Property Protection to Promote Technological Innovation in Equipment Manufacturing Industry in Shaanxi Province.

REFERENCES

[1] Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. International Journal of Production Research, 59(16), 4773-4778.

[2] Jin, Z., Zhang, Z., Demir, K., & Gu, G. X. (2020). Machine learning for advanced additive manufacturing. Matter, 3(5), 1541-1556.

[3] Jiang, J., Xiong, Y., Zhang, Z., & Rosen, D. W. (2022). Machine learning integrated design for additive manufacturing. Journal of Intelligent Manufacturing, 33(4), 1073-1086.

[4] Alexopoulos, K., Nikolakis, N., & Chryssolouris, G. (2020). Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. International Journal of Computer Integrated Manufacturing, 33(5), 429-439.

[5] Solke, N. S., Shah, P., Sekhar, R., & Singh, T. P. (2022). Machine learning-based predictive modeling and control of lean manufacturing in automotive parts manufacturing industry. Global Journal of Flexible Systems Management, 23(1), 89-112.

Table 4.1: Results of time series correction.

	Flow rate /(t/h)				Temperature /K			
NO.	Truth value	Measured value	Corrected value	Calibration error /%	Truth value	Measured value	Corrected value	Calibration error /%
1	126.77	126.16	126.66	-0.08	117.13	116.82	117.09	-0.03
2	122.81	124.27	122.28	-0.44	115.01	117.61	114.98	-0.03
3	125.73	125.82	125.67	-0.08	117.11	117.15	117.08	-0.03
4	121.77	121.69	122.08	0.24	115.00	115.90	114.98	-0.03
5	130.42	131.04	130.70	0.20	218.07	217.85	218.15	0.03
6	110.83	111.96	110.43	-0.40	238.88	237.58	238.95	0.03
7	126.35	188.53	126.19	-0.14	209.82	207.42	209.88	0.03
8	129.48	129.47	129.38	-0.07	194.28	193.01	194.33	0.03
9	71.46	71.34	71.44	0.01	202.61	205.34	202.68	0.03
10	105.63	105.77	105.42	-0.20	288.48	288.07	288.57	0.03
11	105.63	105.19	105.71	0.09	288.48	288.58	288.57	0.03
12	108.85	56.23	108.52	-0.30	288.45	286.63	288.54	0.03
13	105.63	106.42	105.60	-0.01	288.48	291.23	288.57	0.03
14	425.63	425.77	425.25	-0.10	382.80	383.23	383.31	0.14
15	76.35	76.47	76.63	0.42	117.21	117.15	117.17	-0.03
16	30.31	30.23	30.33	0.14	164.52	165.52	164.51	0.00
17	76.56	76.42	76.84	0.43	227.79	233.00	227.77	-0.01
18	57.50	56.96	54.61	-5.27	335.84	334.25	335.80	-0.01
19	63.65	63.28	64.10	0.83	374.50	368.61	374.40	-0.03
20	64.58	65.19	65.13	0.83	374.48	369.59	374.36	-0.03
21	63.65	64.44	64.00	0.65	374.50	372.85	374.40	-0.03
22	64.58	65.16	65.03	0.68	374.48	376.29	374.36	-0.03
23	256.46	257.96	258.26	0.75	395.93	399.47	395.54	-0.10
24	15.63	15.50	15.61	0.07	156.27	156.33	156.27	0.00
25	65.00	65.17	65.48	0.83	260.70	257.08	260.75	0.02
26	48.75	71.96	48.90	0.34	316.99	313.24	317.06	0.02
27	0.00	0.00	0.00	0.16	386.46	383.29	386.46	0.00
28	126.15	126.77	127.26	0.95	387.80	391.03	388.07	0.07
29	1.01	1.02	1.01	0.04	83.33	82.96	83.33	0.00

[6] Xia, C., Pan, Z., Polden, J., Li, H., Xu, Y., & Chen, S. (2022). Modelling and prediction of surface roughness in wire arc additive manufacturing using machine learning. Journal of Intelligent Manufacturing, 33(5), 1467-1482.

[7] Fernandes, M., Corchado, J. M., & Marreiros, G. (2022). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. Applied Intelligence, 52(12), 14246-14280.

[8] Ranjan, N., Kumar, R., Kumar, R., Kaur, R., & Singh, S. (2023). Investigation of fused filament fabrication-based manufacturing of ABS-Al composite structures: prediction by machine learning and optimization. Journal of Materials Engineering and Performance, 32(10), 4555-4574.

[9] Jiang, J. (2023). A survey of machine learning in additive manufacturing technologies. International Journal of Computer Integrated Manufacturing, 36(9), 1258-1280.

[10] Putnik, G. D., Manupati, V. K., Pabba, S. K., Varela, L., & Ferreira, F. (2021). Semi-Double-loop machine learning based CPS approach for predictive maintenance in manufacturing system based on machine status indications. CIRP Annals, 70(1), 365-368.

[11] Sing, S. L., Kuo, C. N., Shih, C. T., Ho, C. C., & Chua, C. K. (2021). Perspectives of using machine learning in laser powder bed fusion for metal additive manufacturing. Virtual and Physical Prototyping, 16(3), 372-386.

[12] Chen, L., Yao, X., Xu, P., Moon, S. K., & Bi, G. (2021). Rapid surface defect identification for additive manufacturing with in-situ point cloud processing and machine learning. Virtual and Physical Prototyping, 16(1), 50-67.

[13] Liu, Z., Rolston, N., Flick, A. C., Colburn, T. W., Ren, Z., Dauskardt, R. H., & Buonassisi, T. (2022). Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. Joule, 6(4), 834-849.

[14] Farbiz, F., Habibullah, M. S., Hamadicharef, B., Maszczyk, T., & Aggarwal, S. (2023). Knowledge-embedded machine learning and its applications in smart manufacturing. Journal of Intelligent Manufacturing, 34(7), 2889-2906.

[15] Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: a systematic review. Journal of Intelligent Manufacturing, 33(7), 1879-1905.
[16] Penumuru, D. P., Muthuswamy, S., & Karumbu, P. (2020). Identification and classification of materials using machine vision and machine learning in the context of industry 4.0. Journal of Intelligent Manufacturing, 31(5), 1229-1241.
[17] Barrionuevo, G. O., Sequeira-Almeida, P. M., Ríos, S., Ramos-Grez, J. A., & Williams, S. W. (2022). A machine learning approach for the prediction of melting efficiency in wire arc additive manufacturing. The International Journal of Advanced Manufacturing Technology, 120(5), 3123-3133.
[18] Thakur, V., Kumar, R., Kumar, R., Singh, R., & Kumar, V. (2024). Hybrid additive manufacturing of highly sustainable Polylactic acid-Carbon Fiber-Polylactic acid sandwiched composite structures: Optimization and machine learning. Journal of Thermoplastic Composite Materials, 37(2), 466-492.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: May 11, 2024
Accepted: Jun 20, 2024