APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 2 | June 2025 | pp 157–172
https://doi.org/10.14743/apem2025.2.533

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Enhanced product defect forecasting using partitioned attributes and ensemble machine learning

Sun, Y.Y.[a], **Yang, J.H.**[a,*], **Zhai, L.Y.**[a], **Liu, N.**[a]

[a]School of Economics and Management, University of Science and Technology Beijing, Beijing, P.R. China

ABSTRACT

This study addresses a critical challenge in industrial big data analytics for smart manufacturing: conventional machine learning methods often fail to account for data discontinuities caused by scrapped defective intermediates in multi-stage production processes, inadvertently treating non-conforming products as qualified during model training. We propose a novel process-aware data analytics framework specifically designed for process industries, featuring: (1) intelligent attribute partitioning based on information flow discontinuity points, and (2) an ensemble modelling approach combining Random Forest and C5.0 Decision Tree algorithms to generate interpretable prediction rules with quantified feature importance rankings. Validated using real-world production data from a Chinese rail steel manufacturer, our methodology demonstrates superior performance by explicitly incorporating process-specific data correlations. The proposed solution effectively mitigates information distortion caused by scrapped intermediates while maintaining operational interpretability – a crucial requirement for industrial implementation. The research results increased the accuracy rate of the test set of the random forest experiment from 88.39 % to 92.69 %, and the accuracy rate of the test set of the decision tree experiment from 71.89 % to 79.15 %. Additionally, the experimental results verify that, compared with the traditional methods, our framework has better applicability in capturing product quality in the manufacturing industry when process attributes are considered.

ARTICLE INFO

Keywords:
Intelligent manufacturing;
Process industry;
Industrial data mining;
Defect prediction;
C5.0 decision tree;
Random forest;
Process-oriented analytics;
Machine learning

**Corresponding author:*
yangjh@ustb.edu.cn
(Yang, J.H.)

Article history:
Received 27 March 2025
Revised 28 May 2025
Accepted 10 June 2025

References

[1] Qin, S.J. (2014). Process data analytics in the era of big data, *AIChE Journal*, Vol. 60, No. 9, 3092-3100, doi: 10.1002/aic.14523.
[2] Topal, B., Sahin, H. (2018). The influence of information sharing in the supply chain process on business performance: An empirical study, *Studies in Informatics and Control*, Vol. 27, No. 2, 203-214, doi: 10.24846/v27i2y201808.
[3] Nagy, R., Horvát, F., Fischer, S. (2024). Innovative approaches in railway management: Leveraging big data and artificial intelligence for predictive maintenance of track geometry, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 4, 1268-1276, doi: 10.17559/TV-20240420001479.
[4] Nguyen, T.V., Zhou, L., Chong, A.Y.L., Li, B., Pu, X. (2019). Predicting customer demand for remanufactured products: A data-mining approach, *European Journal of Operational Research*, Vol. 281, No. 3, 543-558, doi: 10.1016/j.ejor.2019.08.015.
[5] Borchert, P., Coussement, K., De Weerdt, J., De Caigny, A. (2024). Industry-sensitive language modeling for business, *European Journal of Operational Research*, Vol. 315, No. 2, 691-702, doi: 10.1016/j.ejor.2024.01.023.
[6] Wu, Z., Shi, Y. (2024). Development and digitalization of cultural industry marketing based on big data, *Environmental Engineering and Management Journal*, Vol. 23, No. 5, 1097-1108, doi: 10.30638/eemj.2024.089.

[7] Kovacic, M., Zuperl, U., Gusel, L., Brezocnik, M. (2023). Reduction of surface defects by optimization of casting speed using genetic programming: An industrial case study, *Advances in Production Engineering & Management*, Vol. 18, No. 4, 501-511, doi: 10.14743/apem2023.4.488.
[8] Perzyk, M., Kochanski, A., Kozlowski, J., Soroczynski, A., Biernacki, R. (2014). Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis, *Information Sciences*, Vol. 259, 380-392, doi: 10.1016/j.ins.2013.10.019.
[9] Hsu, S.C., Chien, C.F. (2007). Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing, *International Journal of Production Economics*, Vol. 107, No. 1, 88-103, doi: 10.1016/j.ijpe.2006.05.015.
[10] Norrena, J., Louhenkilpi, S., Visuri, V.V., Alatarvas, T., Bogdanoff, A., Fabritius, T. (2023). Assessing the effects of steel composition on surface cracks in continuous casting with solidification simulations and phenomenological quality criteria for quality prediction applications, *Steel Research International*, Vol. 94, No. 5, Article No. 2200746, doi: 10.1002/srin.202200746.
[11] Hochbaum, D.S., Liu, S. (2018). Adjacency-clustering and its application for yield prediction in integrated circuit manufacturing, *Operations Research*, Vol. 66, No. 6, 1457-1759, doi: 10.1287/opre.2018.1741.
[12] Vukelic, D., Milosevic, A., Ivanov, V., Kocovic, V., Santosi, Z., Sokac, M., Simunovic, G. (2024). Modelling and optimization of dimensional accuracy and surface roughness in dry turning of Inconel 625 alloy, *Advances in Production Engineering & Management*, Vol. 19, No. 3, 371-385, doi: 10.14743/apem2024.3.513.
[13] Chongwatpol, J. (2015). Prognostic analysis of defects in manufacturing, *Industrial Management & Data Systems*, Vol. 115, No. 1, 64-87, doi: 10.1108/IMDS-05-2014-0158.
[14] Lee, C.K.H., Ho, G.T.S., Choy, K.L., Pang, G.K.H. (2013). A RFID-based recursive process mining system for quality assurance in the garment industry, *International Journal of Production Research*, Vol. 52, No. 14, 4216-4238, doi: 10.1080/00207543.2013.869632.
[15] Agarwal, S., Dandge, S.S., Chakraborty, S. (2019). Development of association rules to study the parametric influences in non-traditional machining processes, *Sadhana: Academy Proceedings in Engineering Sciences*, Vol. 44, Article No. 230, doi: 10.1007/s12046-019-1218-6.
[16] Breznikar, Z., Bojinovic, M., Brezocnik, M. (2024). Application of machine learning to reduce casting defects from bentonite sand mixture, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 634-643, doi: 10.2507/IJSIMM23-4-702.
[17] Ciarapica, F., Bevilacqua, M., Antomarioni, S. (2019). An approach based on association rules and social network analysis for managing environmental risk: A case study from a process industry, *Process Safety & Environmental Protection*, Vol. 128, 50-64, doi: 10.1016/j.psep.2019.05.037.
[18] Li, Q., Li, D., Cao, L. (2015). Modeling and optimum operating conditions for FCCU using artificial neural network, *Journal of Central South University*, Vol. 22, No. 4, 1342-1349, doi: 10.1007/s11771-015-2651-2.
[19] Irani, K.B., Cheng, J., Fayyad, U.M., Qian, Z. (1993). Applying machine learning to semiconductor manufacturing, *IEEE Expert*, Vol. 8, No. 1, 41-47, doi: 10.1109/64.193054.
[20] Keswani, M. (2024). Designing a fuzzy logic-based carbon emission cost-incorporated inventory model: A comparative analysis of different machine learning algorithms for demand forecasting with memory effects, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 4, 143-158, doi: 10.24818/18423264/58.4.24.20.
[21] Li, Z., Tie-Xin, C., Ying, M., Qi, L., Liu, M. (2016). Decision tree data mining model for welding parameters selection based on C5.0 improved algorithm and its application, *Chinese Journal of Management Science*, Vol. 2016, No. S1, 230-236.
[22] Grădinaru, G.-I., Manea, D.-I., Andreescu, F., Toma, D.-A., Paraschiv, L.-I. (2024). Identifying the main factors of elaborating "Smart City" strategy using machine learning: A comparative study among Romanian cities, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 3, 53-71, doi: 10.24818/18423264/58.3.24.04.
[23] Özbalci, O., Çakir, M., Oral, O., Doğan, A. (2023). Machine learning approach to predict the effect of metal foam heat sinks discretely placed in a cavity on surface temperature, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 6, 2003-2013, doi: 10.17559/TV-20240302001366.
[24] Wang, T., Wang, X., Ma, R., Li, X., Hu, X., Cahn, F.T.S. (2020). Random Forest-Bayesian optimization for product quality prediction with large-scale dimensions in process industrial cyber–physical systems, *IEEE Internet of Things Journal*, Vol. 7, No. 9, 8641-8653, doi: 10.1109/JIOT.2020.2992811.
[25] Esteve, M., Aparicio, J., Rodriguez-Sala, J.J., Zhu, J. (2022). Random forests and the measurement of super-efficiency in the context of free disposal hull, *European Journal of Operational Research*, Vol. 304, No. 2, 729-744, doi: 10.1016/j.ejor.2022.04.024.
[26] Han, J.H., Lee, J.Y. (2023). Genetic algorithm-based approach for makespan minimization in a flow shop with queue time limits and skipping jobs, *Advances in Production Engineering & Management*, Vol. 18, No. 2, 152-162, doi: 10.14743/apem2023.2.463.
[27] Stojic, N., Delic, M., Bojanic, T., Jokanovic, B., Tasic, N. (2024). Integrated model of risk management in business processes in industrial systems, *International Journal of Simulation Modelling*, Vol. 23, No. 3, 412-423, doi: 10.2507/IJSIMM23-3-689.
[28] Wei, Z.H., Yan, L., Yan, X. (2024). Optimizing production with deep reinforcement learning, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 692-703, doi: 10.2507/IJSIMM23-4-CO17.
[29] Yao, L., Ge, Z. (2018). Big data quality prediction in the process industry: A distributed parallel modeling framework, *Journal of Process Control*, Vol. 68, 1-13, doi: 10.1016/j.jprocont.2018.04.004.
[30] Negoiță, R.F., Borangiu, T. (2023). Robotic process automation of inventory demand with intelligent reservation, *Studies in Informatics and Control*, Vol. 32, No. 2, 5-14, doi: 10.24846/v32i2y202301.

[31] Novak, C., Pfahlsberger, L., Bala, S., Revoredo, K., Mendling, J. (2023). Enhancing decision-making of IT demand management with process mining, *Business Process Management Journal*, Vol. 29, No. 8, 230-259, doi: 10.1108/BPMJ-12-2022-0631.

[32] Ding, J., Liu, Y., Zhang, L., Wang, J., Liu, Y. (2016). An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model, *Applied Intelligence*, Vol. 44, No. 2, 340-361, doi: 10.1007/s10489-015-0713-7.

[33] Adnan, M.N., Islam, M.Z. (2016). Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm, *Knowledge-Based Systems*, Vol. 110, 86-97, doi: 10.1016/j.knosys.2016.07.016.