Theses of the doctoral (PhD) dissertation

Integrated methodologies for data-driven soft sensor enhancement

Pál Péter Hanzelik

University of Pannonia

Doctoral School of
Chemical Engineering and Material Sciences

Supervisors:

János Abonyi, PhD, Prof.

Alex Kummer, PhD

Department of Process Engineering

Veszprém

2026

1 Introduction and the aim of the work

The contemporary industrial landscape is increasingly defined by digitalization and the integration of Industry 4.0 solutions. These technologies are no longer merely optional enhancements but critical prerequisites for maintaining market competitiveness and achieving environmental sustainability. Smart factories utilize networked machines and Internet of Things (IoT) sensors to facilitate data-driven decision-making, leading to enhanced operational flexibility, reduced energy consumption, and minimized Health, Safety, and Environmental (HSE) risks.

In the specific context of the chemical industry, the primary challenge lies in the real-time qualification of raw materials and products. Conventional laboratory-based quality assurance often introduces significant latency, hindering rapid response to process fluctuations. Consequently, the transition toward real-time data analysis and automated process control is essential for optimizing productivity and preventing unexpected shutdowns.

Soft Sensors and Predictive Modeling

A cornerstone of this research is the development and application of "soft sensors." These virtual instruments function as data-driven mathematical models that estimate critical process parameters—such as product composition, viscosity, or quality—which are otherwise difficult, hazardous, or cost-prohibitive to measure with physical hardware. By utilizing easily measurable variables (e.g., pressure, temperature, or spectroscopic data) as inputs, soft sensors provide high-frequency estimations that enable tighter process control and optimization.

The performance of these models is intrinsically linked to the efficacy of Machine Learning (ML) algorithms. While ML revolutionizes industrial forecasting, its long-term industrial utility depends on robust data management and lifecycle monitoring. This research emphasizes the Machine Learning Operations (MLOps) methodology,

which encompasses the entire lifecycle of an algorithm—from development and automated deployment to continuous performance tracking and scaling.

Research Objectives and Methodology

The overarching aim of this dissertation is to advance the reliability and robustness of soft sensors through integrated Industry 4.0 solutions. The research uses industrial data science to examine the consistency of hierarchical data, sensor fusion, and data scarcity, and proposes an industrial framework. To achieve this, the work is structured around the following key objectives:

- *Hierarchical timeseries data reconciliation:* System modeling at multiple organizational levels often introduces inherent errors and inconsistencies. This work investigates optimal reconciliation techniques—specifically comparing independent modeling with error-based fine-tuning—to ensure that ML predictions adhere to physical and chemical aggregation constraints.

- *Advanced data fusion:* To enhance model resilience, a Complex-level Ensemble Fusion (CLF) methodology was developed. By integrating spectroscopic data (MIR and Raman) through a two-layer stacking algorithm, the research aims to surpass traditional low- and mid-level fusion accuracy in quality control applications.

- *Synthetic data generation***:** Addressing the challenge of limited industrial datasets, this research proposes a solution for generating physically meaningful artificial infrared spectra using Principal Component Analysis (PCA) and neural networks, ensuring robust model training in data-scarce scenarios.

- *CRISP-ML & Models lifecycle:* The research proposes a methodology for successful data-driven projects. The primary goal is to create a comprehensive framework for managing the entire life cycle of machine learning-based software sensors. This includes solving problems related to version management,

model degradation, and real-time monitoring in complex chemical environments (MLOps).

Structure of the Dissertation

The dissertation is organized as follows: Chapter 2 provides a systematic literature review of soft sensor trends. Chapter 3 details data reconciliation in hierarchical time series. Chapter 4 presents the methodology for complex-level data fusion. Chapter 5 discusses artificial data generation for spectroscopic models. Chapter 6 introduces the MLOps framework for industrial soft sensors, followed by a comprehensive summary of findings and future research directions in Chapters 7 and 8.

2 Theses

Thesis #1. I have demonstrated that the appropriate development of machine learning models can be integrated into a hierarchical model framework with data reconciliation, which is characterised by high predictive power, practicality, and robustness to measurement errors.

- I developed a data reconciliation technique for hierarchical time series that improves the usability of machine learning models.
- I applied it to three different datasets: an industrial one, a benchmark dataset, and a comprehensive waste management dataset. The results demonstrate that the technique helps to improve the usability of machine learning models.
- Through the three case studies, I demonstrated the utility of the technique and its importance for application in complex systems.

Related publications: 1, 6

Thesis #2. I have demonstrated that the performance of a weak machine learning algorithm can be significantly improved by applying complex data fusion techniques that leverage complementary information from multiple data sources.

- I developed a data fusion technique that is capable of creating higher-performing machine learning models by fusing data from different measurement spectra.
- Based on the two datasets I presented, the developed complex-level ensemble model demonstrated superior performance compared to other data fusion techniques.
- I tested data fusion techniques based on MIR and Raman spectra in datasets where the accuracy of the baseline models, built solely on MIR or Raman data, was poor (less than 90%).

Related publications: 2, 7

Thesis #3. I have demonstrated that synthetic infrared spectra can be purposefully generated for samples with complex matrices. The developed methodology was validated using independent real-world spectra and is applicable for augmenting incomplete datasets.

- I have developed a technique for generating artificial spectra that can efficiently generate data of appropriate quality, especially for sparse or incomplete data sets.
- I created the artificial infrared spectra using the developed methodology, PCA and ANN techniques.
- I tested the authenticity of the artificial spectra in detail and presented the methodology for their validation.

Related publications: 3, 4

Thesis #4. I have demonstrated that the architecture built on the CRISP-ML methodology effectively supports the operation of machine learning models with continuously changing performance.

- I developed an industrial-grade framework that uses cloud and edge-based computing to predict laboratory data in near-real time.
- I developed a framework that incorporates all elements of the CRISP-ML methodology, which allows for the continuous monitoring of model performance and, when necessary, retraining.
- I tested the framework on an industrial network, and its accuracy was demonstrated through two case studies involving machine learning models.

Related publications: 5, 8, 9

Utilization of the results and future aims

Industrial Utilization of Research Findings

The research presented in this dissertation bridges the gap between theoretical machine learning and the rigorous demands of the chemical and process industries. The methodologies developed were designed as integrated components of a broader Industry 4.0 ecosystem, offering multi-faceted advantages in quality assurance and operational efficiency.

Implementation of Robust Soft Sensors

The primary utilization of the results lies in the deployment of "soft sensors" as virtual instruments. By adopting the hybrid modeling approach discussed, industrial plants can reconcile high-level system constraints with low-level sensor data. This ensures that estimated parameters, such as product composition, remain physically consistent with mass and energy balances. For quality assurance, this provides real-time, reliable data that significantly reduces the frequency of time-consuming and destructive laboratory sampling.

Enhanced Decision Support through Data Fusion

The Complex-Level Ensemble Fusion (CLF) methodology provides a transferable framework for spectroscopic analysis. In environments where multiple sensors (e.g., Raman and MIR) are available, CLF exploits complementary spectral information to drastically reduce prediction errors. Its modular nature allows for the integration of new analytical tools without re-architecting the system, thereby future-proofing industrial quality control infrastructures.

Mitigating Targeted Data Gaps through Data Generation

A common bottleneck in industrial AI is the lack of comprehensive training data for rare process conditions. The artificial data generation framework addresses this by enriching the parameter space with physically meaningful synthetic samples. This allows for the development of robust models even before extensive historical data is available, shortening the time-to-market for new products and reducing the risks associated with data-driven systems.

CRISP-ML Framework and Model Lifecycle Management

The long-term usefulness of this work is embodied in the proposed CRISP-ML lifecycle. By applying the outlined systematic monitoring protocols, organizations can transition from static model deployment to a dynamic, self-correcting lifecycle. The framework uses edge and cloud computing technologies to keep soft sensors synchronized with the current state of the factory, ensuring stable production, minimizing off-spec products and chemical waste.

Future Aims and Strategic Directions

The goal is to use these methods to integrate them into an autonomous, process-oriented intelligence system, if not completely, then at least in a way that is appropriate for the business problem at hand.

- To enhance flexibility, data reconciliation rules—such as physical and chemical constraints—will be made customizable through a user-friendly interface. This will allow operators to adjust the underlying of the model without expert programming knowledge.
- Autonomous systems will be implemented in the future. The system's ML models predict the most accurate data possible every day through automatic retraining to ensure operational quality assurance.
- In the case of data generation, the future goal is to thoroughly test and quantitatively demonstrate the improvement in the accuracy of models using artificial data. Successful data generation makes the application of convolutional neural networks (CNN) feasible.

Final Synthesis

In conclusion, this dissertation proposes several methodologies for next-generation intelligent manufacturing. Through the synergy of data harmonization, sensor fusion, artificial data generation, and lifecycle management, the research proposes methods for industry that will enable more accurate, flexible, and sustainable industrial decision-making in the future.

3 Publications related to theses

Articles in international journals

1. Pál Péter Hanzelik, Alex Kummer, and János Abonyi. „Edge-computing and machine-learning-based framework for soft-ware sensor development” *Sensors*, 22(11):4268, 2022.
   doi: 10.3390/s22114268
   Scimago Journal Ranking: Q1-Q2, Impact factor: 4.532

2. Pál Péter Hanzelik, Alex Kummer, and János Abonyi. „Data reconciliation-based hierarchical fusion of machine learning models” *Machine Learning and Knowledge Extraction*, 6(4):2601–2617, 2024.
   doi: 10.3390/make6040125
   Scimago Journal Ranking: Q1, Impact factor: 8.980

3. Pál Péter Hanzelik, Szilveszter Gergely, János Abonyi, and Alex Kummer. „Data fusion of spectroscopic data for enhancing machine learning model performance” *Digital Chemical Engineering*, page 100271, 2025.
   doi: 10.1016/j.dche.2025.100271
   Scimago Journal Ranking: Q1, Impact factor: 4.775

4. László Győry, Szilveszter Gergely, and Pál Péter Hanzelik. „Generating realistic infrared spectra using artificial neural networks” *Journal of Chemometrics*, 38(9):e3573, 2024.
   doi: 10.1002/cem.3198
   Scimago Journal Ranking: Q3, Impact factor: 2.189

5. Pál Péter Hanzelik, Szilveszter Gergely, Csaba Gáspár, and László Győry. „Machine learning methods to predict solubilities of rock samples” *Journal of Chemometrics*, 34(2):e3198, 2020.
   doi:10.1002/cem.3198
   Scimago Journal Ranking: Q3, Impact factor: 1.569

Articles in conference proceedings

6. Pál Péter Hanzelik, Alex Kummer, Ádám Ipkovich, and János Abonyi. „Fusion and integrated correction of chemometrics and machine learning models based on data reconciliation" In M.C. Georgiadis A. Kokossis and S. Pistikopoulos, editors, Computer Aided Chemical Engineering, volume 52, pages 1379–1384. Elsevier, 2023. Preceeding and presented at the *33rd European Symposium on Computer-Aided Process Engineering (ESCAPE-33) conference*. Availability: link

7. Pál Péter Hanzelik, Alex Kummer, Márton Mócz, Szilveszter Gergely, Dorián L Galata, and János Abonyi. „Comparison of different data and information fusion methods to improve the performance of machine learning models" In F. Manenti and G.V.R. Reklaitis, editors, Computer Aided Chemical Engineering, volume 53, pages 3007–3012. Elsevier, 2024. Preceeding and presented at the *34th European Symposium on Computer Aided Process Engineering /15th International Symposium on Process Systems Engineering (ESCAPE-34/PSE2024) conference*. Availability: link

Conference abstracts & presentations

8. Pál Péter Hanzelik, Alex Kummer, and János Abonyi. „Development of the edge- and cloud computing framework in the analytical chemistry" Conference presentation, August 2022. at the *International Congress of Chemical and Process Engineering (CHISA) conference*, Availability: link

9. Pál Péter Hanzelik, Alex Kummer, Ádám Ipkovich, and János Abonyi. „Performance monitoring of machine learning models" Conference presentation, September 2022. at the *Process Control Systems (PCS) conference,* Availability: link

4 Publications not related to theses

Articles in international journals

- Lei Fu, Yanxiang Yu, Chicheng Xu, Michael Ashby, Andrew McDonald, Wen Pan, Tianqi Deng, István Szabó, Pál Péter Hanzelik, Csilla Kalmár, et al. „Well-log-based reservoir property estimation with machine learning: a contest summary" *Petrophysics*, 65(01):108–127, 2024.
  doi: 10.30632/PJV65N1-2024a6
  Scimago Journal Ranking: Q3, Impact factor: 0.885

- Norbert Péter Szabó, Károly Nehéz, Oliver Hornyák, Imre Piller, Csaba Deák, Pál Péter Hanzelik, Csaba Kutasi, and Károly Ott. „Cluster analysis of core measurements using heterogeneous data sources: An application to complex miocene reservoirs" *Journal of Petroleum Science and Engineering,* 178:575–585, 2019.
  doi: 10.1016/j.petrol.2019.03.067
  Scimago Journal Ranking: Q1, Impact factor: 5.168

Conference abstracts & presentations

- Márton Mócz, Pál Péter Hanzelik, János Slezsák, and Gergely Szilveszter. „Impact of different model transfer algorithms on dilution series and oil samples" Presentation, September 2023. Poster presented at the *Conferencia Chemometrica 2023 conference.*

- Puskas, S., Vago, A., Toro, M., Ordog, T., Kalman, G., Hanzelik, P. P., Bihari, Zs., Blaho, J., Tabajdi, R., Dekany, I., Dudas, J., Nagy, R., Bartha, L., and I. Lakatos. „Surfactant-Polymer EOR from Laboratory to the Pilot" *Paper presented at the SPE EOR Conference at Oil and Gas West Asia*, Muscat, Oman, March 2018. doi: 10.2118/190369-MS