Revision of doctoral (PhD) dissertation

Name of the candidate: Pál Péter Hanzelik

Title: Integrated Methodologies for Data-Driven Soft Sensor Enhancement

Reviewer: Dr. András P. Borosy, Associate Principal Scientist, dsm-firmenich, 1242 Satigny, Switzerland

The dissertation presents a broad and ambitious framework for enhancing soft sensor performance in industrial settings. The integration of hierarchical data reconciliation, advanced data fusion, artificial data generation, and lifecycle management demonstrates a holistic approach. The work is grounded in state-of-the-art machine learning, chemometrics, and process engineering. The author employs rigorous methodologies, including genetic algorithms for feature selection, ensemble learning, and neural networks for artificial data generation. The combination of these techniques is novel and well-justified.

Chapter 1: Introduction

Strengths:

- Clearly establishes the context and importance of Industry 4.0, digitalization, and soft sensors.
- Outlines the main thematic areas and methodological pillars, providing a roadmap for the dissertation.
- Introduces MLOps which is a way to ensure **continuous performance monitoring**, **model updating**, and **long-term sustainability** for soft sensors in industrial environments. It enables the integration of machine learning models with edge and cloud computing, and supports the lifecycle management of predictive models
- Supported by schematic diagrams summarizing ML development and deployment.

Chapter 2: Literature Review of Soft Sensors in Industrial Applications

Strengths:

- Provides an extensive and structured review of soft sensor development and application.
- Discusses the significance for process intensification, predictive maintenance, and quality assurance.
- Presents a general methodology for soft sensor development with clear schematics.
- Includes a comprehensive literature review table and network diagrams.
- Identifies technological and methodological challenges and future research directions.

Weaknesses:

- The review mostly summarizes rather than critically analyzes previous works.
- The literature review table does not discuss weaknesses or gaps in cited studies.

Chapter 3: Data Reconciliation-Based Hierarchical Fusion of Machine Learning Models

Strengths:

- Presents a rigorous, mathematically detailed approach to integrating data reconciliation (DR) with machine learning (ML) for hierarchical systems.
- Explains the use of summation and incidence matrices, and the formulation of constraints in hierarchical modeling.
- Demonstrates the method on three case studies of increasing complexity: mineral composition prediction, retail sales forecasting (Walmart M5), and waste management in Hungary.
- Provides detailed results, including RMSE values and hierarchical balance errors, and shows the superiority of the third (ML+DR) method in maintaining constraints and improving accuracy.

Weaknesses:

- Assumes precise knowledge of hierarchical structure and constraints, which may not always be available.

- Requires high-quality, reliable data; impact of noisy or missing data is not deeply discussed.

Chapter 4: Data Fusion of Spectroscopic Data for Enhancing Machine Learning Model Performance

Strengths

- Provides a thorough review of data fusion (DF) techniques in chemometrics, including low-, mid-, high-, and complex-level fusion.
- Introduces the novel Complex-Level Ensemble Fusion (CLF) method, which combines genetic algorithm-based feature selection, PLS projection, and XGBoost stacking.
- Benchmarks CLF against classical fusion schemes and single-source models, showing significant improvements in predictive accuracy, especially in small-sample scenarios.
- Applies the method to both industrial lubricant additives and the RRUFF mineral dataset, demonstrating generalizability.
- Discusses the practical advantages and caveats of CLF, such as the need for rigorous preprocessing and feature selection.

Weaknesses:

- Case studies are limited in sample size; impact on model robustness and transferability is not critically examined.
- Does not discuss scaling to larger, more diverse datasets or multi-site environments.

Chapter 5: Generating Realistic Infrared Spectra Using Artificial Neural Networks

Strengths:

- Addresses limited and uneven spectral data by generating artificial spectra using PCA and neural networks.
- Methodology is clearly illustrated and validated statistically and visually.
- Artificial spectra fill gaps in parameter space, improving training data distribution and model robustness.

- Quantitative comparison to real spectra using multiple statistical measures.
- Method is transferable, with additional case studies in the appendix.
- Discusses impact of principal component number and outlines future directions.

Weaknesses:

- Dependence on the number of principal components is acknowledged but not systematically analyzed.
- Validated only on rock samples; applicability to other spectra or industrial data is not discussed.
- Future plans (e.g., CNNs) are briefly mentioned without detail on challenges.

Chapter 6: Edge Computing and Machine Learning-Based Framework for Software Sensor Development

Strengths

- Presents a comprehensive framework for the lifecycle management of ML-driven soft sensors, leveraging edge and cloud computing.
- Reviews related work and patents, using the PRISMA methodology to systematically analyze the literature.
- Proposes the CRISP-ML methodology for sustainable model development, deployment, monitoring, and retraining.
- Discusses the integration of IoT, ERP, and laboratory data, and the use of predictive model markup language (PMML) for model portability.
- Includes a case study comparing ML models for nitrogen and quartz content, with detailed performance metrics and lessons learned.
- Addresses practical aspects such as model maintenance, robustness, and the need for continuous monitoring.

Weaknesses:

- Does not address scalability issues or edge device limitations.
- Risks and challenges of maintaining models in dynamic environments are not critically analyzed.

Chapter 7: Conclusions

Strengths:

- Thorough summary of research objectives, scope, and progress in Industry 4.0 integration.
- Highlights main contributions of each chapter, linking them to overall goals.
- Discusses practical applicability and potential for real-world implementation.
- Outlines future directions, including integration of methodologies for smart manufacturing.

Weaknesses:

- Little discussion of risks, uncertainties, or areas where methodologies may not perform as expected.

Chapter 8: Thesis Findings

Strengths:

- Clearly presents key findings and significant results, highlighting primary goals and contributions.
- Outlines potential next steps and future research directions.
- Demonstrates empirical utility through detailed case studies.
- Emphasizes practical, industry-ready digital solutions leveraging Industry 4.0.

Appendices, Acronyms, Bibliography, Further Publications

Strengths

- Appendices provide additional data, visualizations, and case studies that support the main text.
- The bibliography is comprehensive and up to date, reflecting a solid grounding in both classical and contemporary literature.

Questions to the Author

How would the proposed framework perform in settings where system hierarchies are not well-defined or where data quality is poor?

What measures can be taken to ensure the reproducibility and accessibility of the framework for practitioners without advanced machine learning expertise?

Have you considered the long-term impact of artificial data generation on model interpretability and robustness in production environments?

Are there specific industries or data types where you anticipate the framework would not be applicable or would require significant adaptation?

Summary Proposal

In summary, the dissertation presents a significant and original contribution to the field of data-driven soft sensor development. The work is methodologically sound, well-structured, and demonstrates practical relevance through case studies. Following the successful final defense I recommend that the candidate be awarded the PhD degree.

Satigny, April 6, 2026 A.P. Borosy