DISSERTATION
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DOI:10.18136/PE.2026.991

Pál Péter Hanzelik

University of Pannonia
2026

Integrated methodologies for data-driven soft sensor enhancement

The thesis was prepared for the award of a doctoral degree (PhD) within the framework of the Chemical Engineering and Material Sciences Doctoral School at University of Pannonia

in the discipline of Bio-, Environmental and Chemical Engineering Sciences

written by: Pál Péter Hanzelik

Supervisors: Prof. Dr. habil. János Abonyi, Dr. Alex Kummer

I recommend the dissertation for acceptance: yes / no.

…….
Prof. Dr. habil. János Abonyi
(supervisor)

I recommend the dissertation for acceptance: yes / no.

…….
Dr. Alex Kummer
(supervisor)

I recommend the dissertation for peer review.

…….
chair of the DDHC

The PhD-candidate has achieved ……. % at the public debate.

The composition of the Final Examination Committee:

chair:…….

reviewers:.……

members:…….

Veszprém, ……

…….
chair of the committee

Qualification of degree: ….....

Veszprém, ……

…….
chair of the UDHC

Pannon Egyetem

Doktori (PhD) Értekezés

Integrált módszertanok az adatvezérelt szoftver szenzorok fejlesztésére

Szerző:
Hanzelik Pál Péter

Témavezetők:
Prof. Dr. habil. Abonyi János
Dr. Kummer Alex

Értekezés doktori (PhD) fokozat elnyerése érdekében
a ***Pannon Egyetem***

Vegyészmérnöki és Anyagtudományok
Doktori Iskolájához tartozóan

Folyamatmérnöki Intézeti Tanszék

Pannon Egyetem
2026

University of Pannonia

Doctoral (PhD) Dissertation

Integrated methodologies for data-driven soft sensor enhancement

Author:
Pál Péter Hanzelik

Supervisors:
Prof. Dr. habil. János Abonyi
Dr. Alex Kummer

A dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Chemical Engineering and Material Sciences Doctoral School
of University of Pannonia

Department of Process Engineering

University of Pannonia

2026

"Senki sem tudja, mi a helyes algoritmus, de reményt ad, hogy ha sikerül valami durva közelítést felfedeznünk ennek az algoritmusnak a függvényében, és ha számítógépen implementáljuk, az sokat segíthet a fejlesztésben."

"Nagyon sok gyártóüzemben jártam. Még nem léptem be olyanba, ahol azt gondoltam, hogy az AI-megoldások ne tudnának segíteni."

Andrew Ng

"No one knowns what the right algorithm is, but it gives us hope that if we can discover some crude approximation of whatever this algoritm is and implement if on a computer, that can help us make a lot of progress."

"I have been to so many manufacturing plants. I've yet to walk into one where I did not think AI solutions wouldn't help."

Andrew Ng

PANNON EGYETEM

Kivonat

Mérnöki kar
Folyamatmérnöki Intézeti Tanszék

Philosophiæ Doctor

Integrált módszertanok az adatvezérelt szoftver szenzorok fejlesztésére

írta: Hanzelik Pál Péter

Az értekezés egy olyan átfogó keretrendszert mutat be, amely a soft szenzorok teljesítményének és ipari alkalmazhatóságának növelésével hiánypótló megoldást kínál a terület aktuális kutatási kérdéseire. A munka túlmutat az egyedi prediktív modellek fejlesztésén, az Ipar 4.0 elvárásaival összhangban a modellek robusztus és fenntartható gyakorlati implementációjára helyezi a fókuszt.

Az értekezés négy, egymással szorosan összefüggő módszertani pillérre épül. Elsőként egy új hierarchikus adatkiegyenlítési eljárást mutatok be, amely jelentősen növeli a modellek pontosságát és megbízhatóságát. Másodsorban kidolgoztam egy complex-level ensemble fusion (CLF) elnevezésű új adatfúziós technikát, amely teljesítményé-ben felülmúlja a hagyományos megoldásokat. Harmadrészt a disszertáció egy olyan mesterséges adatgenerálási módszert ismertet, amely robusztusabb modelltanítást tesz lehetővé. Végül javaslatot teszek egy szisztematikus életciklus-kezelési keretrendszerre, amely biztosítja a modellek hosszú távú fenntarthatóságát és relevanciáját. A kidolgozott módszertanokat valós ipari környezetből származó és benchmark adatokon alapuló esettanulmányok validálják. A munka kulcsfontosságú megállapítása, hogy e négy pillér egységes, automatizált rendszerbe foglalása koherens és adaptálható keretet kínál az okos gyártás és a modern folyamatirányítás következő generációja számára.

Az értekezésben bemutatott kutatás olyan módszertani vázat ad, amely a digitális átalakuláson áteső iparágakban közvetlenül hasznosítható. A munka az adatintegráció és a modellfenntarthatóság kritikus kihívásaira kínál megoldásokat, ezzel elősegítve a hatékonyabb és megbízhatóbb adatvezérelt működést a modern ipari szektorokban.

UNIVERSITY OF PANNONIA

Abstract

Faculty of Engineering
Department of Process Engineering

Doctor of Philosophy

Integrated methodologies for data-driven soft sensor enhancement

by Pál Péter Hanzelik

This dissertation presents a comprehensive framework for enhancing the performance and industrial applicability of soft sensors, addressing a critical gap in current research. The work goes beyond the development of individual predictive models to focus on their robust and sustainable implementation within the context of Industry 4.0.

The study is based on four interconnected methodological contributions. First, a novel method for hierarchical data reconciliation is introduced, which significantly increases the accuracy and reliability of models. Second, a new data fusion technique, complex-level ensemble fusion (CLF), was developed that surpasses traditional methods. Third, the dissertation also describes a method for artificial data generation, which enables more robust model training. Finally, a systematic lifecycle management framework is proposed, ensuring the long-term viability and relevance of the models. The presented methodologies are validated by case studies using real-world industrial and benchmark data. The work's primary conclusion is that integrating these four pillars into a unified, automated system provides a coherent and transferable blueprint for the next generation of smart manufacturing and process control.

The comprehensive research presented in this dissertation provides a practical structure that can be directly applied in industries undergoing digital transformation. The work offers solutions to the critical challenges of data integration and model sustainability, thereby promoting more efficient and reliable data-driven operations across all sectors of modern industry.

PANNONISCHE UNIVERSITÄT

Auszug

Fakultät für Ingenieurwissenschaften
Abteilung für Verfahrenstechnik

Doktor der Philosophie

Integrierte Methoden zur datengesteuerten Soft-Sensor-Verbesserung

von Pál Péter Hanzelik

Diese Dissertation präsentiert ein umfassendes Framework zur Leistungssteigerung und industriellen Anwendbarkeit von Soft-Sensoren, um eine kritische Forschungslücke zu schließen. Die Arbeit geht über die Entwicklung einzelner Vorhersagemodelle hinaus und konzentriert sich auf deren robuste und nachhaltige Implementierung im Kontext von Industrie 4.0.

Die Studie basiert auf vier miteinander verbundenen methodischen Beiträgen. Erstens wird eine neue Methode zur hierarchischen Datenabgleichung vorgestellt, die die Genauigkeit und Zuverlässigkeit von Modellen erheblich steigert. Zweitens wurde eine neue Datenfusionstechnik, complex-level ensemble fusion (CLF), entwickelt, die herkömmliche Methoden übertrifft. Drittens beschreibt die Dissertation auch eine Methode zur künstlichen Datengenerierung, die ein robusteres Modelltraining ermöglicht. Schließlich wird ein systematisches Lifecycle-Management-Framework vorgeschlagen, das die langfristige Lebensfähigkeit und Relevanz der Modelle sicherstellt. Die vorgestellten Methoden werden durch Fallstudien mit realen Industrie- und Benchmark-Daten validiert. Die zentrale Schlussfolgerung der Arbeit ist, dass die Integration dieser vier Säulen in ein einheitliches, automatisiertes System einen kohärenten und übertragbaren Bauplan für die nächste Generation von Smart Manufacturing und Prozesssteuerung darstellt.

Die in dieser Dissertation vorgestellte umfassende Forschung bietet eine praktische Struktur, die direkt in Unternehmen, die sich im digitalen Wandel befinden, angewendet werden kann. Die Arbeit liefert Lösungen für die kritischen Herausforderungen der Datenintegration und Modellnachhaltigkeit und ebnet so den Weg für effizientere und zuverlässigere datengesteuerte Abläufe in allen Sektoren der modernen Industrie.

Acknowledgements

I extend my sincerest gratitude to my supervisors, Professor Dr. habil. János Abonyi and Dr. Alex Kummer. Their guidance and support throughout my doctoral studies have been invaluable. They provided immeasurable scientific mentorship and demonstrated immense patience, even when I made the journey challenging for all of us. Their exemplary conduct, human values, and behavior within the research community have taught me more than any textbook ever could. I am profoundly thankful for the opportunity to have worked with them; these years will not be forgotten.

My heartfelt thanks also go to my family and friends, whose unconditional love and support carried me through countless hours of research and writing. I am especially grateful to my wife and children, and to my parents for their unwavering patience and support. Their understanding greatly contributed to my completion of my studies and the writing of my dissertation.

Last but not least, I would like to thank Gergely Szilveszter, who helped me develop my statistical and computational knowledge through consultations, László Győry, with whom we developed a new methodology for building neural networks and generating artificial data, my managers and colleagues at MOL Group, my professors, fellow PhD students, and the professional community around me for their continuous encouragement. The discussions and support I received from them were incredibly useful throughout my academic career.

Contents

Chapter 1

Introduction

The development of industrial digitalization and machine learning models has been of prime importance in the past period in terms of competitiveness and environmental protection. Implementing Industry 4.0 solutions plays a vital role in the development of processes and has many advantages compared to traditional methods. Companies using these solutions can gain a competitive advantage in the market, as they can operate more efficiently and flexibly. By using digital solutions, machines and systems in smart factories can communicate with each other, increasing efficiency. Thanks to data-driven decision making, product quality improves as regulation becomes faster and simpler. In addition, the solutions enable for a rapid response to changing market needs, so they will be more flexible in tackling new challenges.

Industry 4.0 solutions can help reduce energy consumption and environmental impact, and due to automation and robotics, physical work is reduced, thus reducing health, safety, and environmental (HSE) protection risks. In my research, the most highlighted area related to Industry 4.0 solutions is data-driven decision making, which enables companies to make more accurate forecasts by analyzing large amounts of data, with which they can make better decisions with. The application of Industry 4.0 solutions in process development is crucial to preserve and increase the competitiveness of companies. Thanks to new technologies, companies can operate more efficiently, flexibly and sustainably. In the context of Industry 4.0, data analysis is primarily driven by three key areas: advanced analytics of large datasets, real-time trend recognition, and informed decision-making support. These areas are underpinned by the development of artificial intelligence

for process optimization and predictive maintenance. A crucial enabler is the seamless networking of IoT sensors and devices, which facilitates the continuous collection and analysis of data. The introduction of automated production processes, which increases efficiency and quality and reduces HSE risk, as well as 3D printing, enables the rapid and cost-effective production of individual products. With industry 4.0 solutions, we can operate chemical industry units in an optimized manner, and one of the main topics of the thesis is that we can speed up quality assurance processes. With these solutions, for example, the qualification of the product or raw material of continuous or batch-based plants can be performed in real time. Real-time data analysis provides immediate insight into processes, so you can react to changes and problems more quickly, and promotes optimized decision-making, which can increase productivity. With continuous analysis of the data and the use of models, it is possible to continuously monitor the condition of the equipment and predict failures, thus avoiding unexpected shutdowns. For real-time data analysis, we need sensors that can provide information on the material flows of raw materials, intermediate products, or final products in real-time.

In large-scale chemical plants, soft sensors are computational models that provide real-time estimations of key process parameters which are difficult, expensive, or impossible to measure directly with physical hardware. They operate by using easily measured process variables (e.g., temperature, pressure, flow rate, or spectroscopic data) as inputs to infer the value of the desired, unmeasured variable (e.g., product composition, quality, or viscosity). Essentially, a soft sensor is a data-driven model that acts as a virtual instrument, offering a cost-effective and efficient alternative to physical sensors, thereby enabling better process monitoring, control, and optimization. The integration of these virtual instruments into existing industrial control systems is a cornerstone of the digital transformation in the process industries. Beyond their cost-effectiveness, soft sensors provide high-frequency data that are essential for advanced process control and predictive maintenance strategies. However, their successful deployment requires a deep understanding of both the underlying physical processes and the statistical nature of the input data. In modern refinery environments, these models must also be resilient to sensor drift and changing operational conditions to maintain their reliability. Consequently, the development of robust and adaptable soft sensors is no longer just a technical advantage but a necessity for sustainable and safe industrial operations.

The application of machine learning algorithms plays a major role in industrial data-driven decision-making. In recent years, the use of ML algorithms has undergone rapid development and plays an increasingly important role in the industry every year. With ML models, companies can make their processes more efficient, develop new products, or test their existing processes under extreme conditions. In the case of data-driven decision-making, ML models can process and analyze huge amounts of data, thus enabling efficient and real-time decision-making. Many tasks can be automated using ML, such as quality control, predictive maintenance, process optimization and consumption forecasts. Furthermore, ML models can be used to provide personalized products and services to customers, thereby increasing their loyalty, and supporting customer services with chatbots and virtual assistants. ML can also be used for fraud detection, for example in financial transactions. In the energy industry, for example, ML algorithms can be used to predict energy consumption and increase energy efficiency. By increasing the rate of automation, errors are reduced, efficiency is increased, and manpower is freed up for other tasks with higher added value. Machine learning is revolutionizing the industry, enabling companies to operate more efficiently, develop new products and gain an operational excellence.

In my thesis I deal with data analysis, artificial intelligence and IoT. The quality assurance techniques that I developed and the measurements that I performed in my research do not contain dangerous substances, are fast non-destructive, cheap and, last but not least, can be installed in soft sensors. The methodology used in the thesis is summarized in Figure 1.1 which shows the logical relationship between traditional and machine learning-based solutions. The direct and traditional method is depicted on the left-hand side of the triangle [1].

Machine learning models used in industry have many advantages and can be applied in many fields. The applicability of machine learning models is contingent upon several critical criteria, including the overarching data management strategy governing data collection, storage, and processing. Furthermore, the availability and integrity of data quality are paramount. Finally, ensuring the long-term viability of these models necessitates a commitment to their continuous training, refinement, and development. A significant part of the mentioned aspects is dealt with by the machine learning operations (MLOps) technique, which shows in detail what is needed for a developed ML model to produce benefits in the long term.

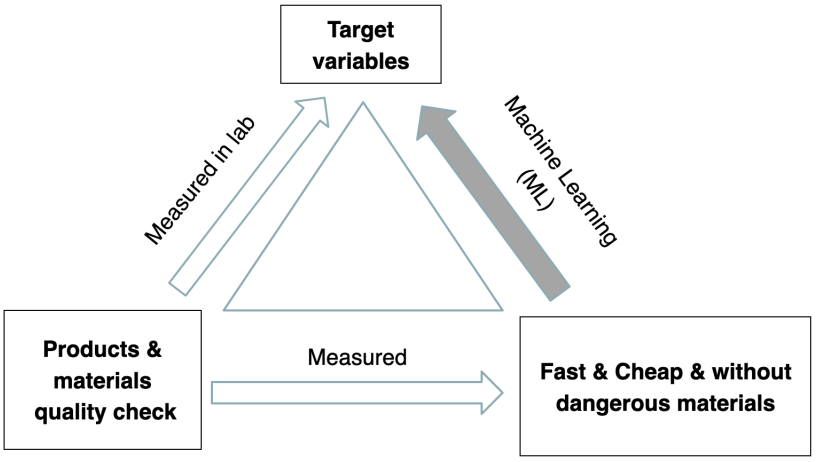


FIGURE 1.1: Schematic diagram of the machine learning development in quality assurance.

MLOps is a technique that encompasses the entire life cycle of machine learning algorithms, from development to production and continuous maintenance. The goal of MLOps is to get machine learning models into production faster and to make them work reliably in the long term. MLOps helps automate some of the steps in model development and deployment, so models can go into production faster, which can give you a competitive edge. MLOps can be used to ensure that the developed models perform well on real data and that their performance does not deteriorate over time. Furthermore, it allows models to be scaled to large amounts of data and users, and the technique allows machine learning projects to be done at lower costs, automating processes, and minimizing errors. By using MLOps, following the suggested techniques ensures that the models are reproducible. In this thesis, I will present a framework related to MLOps in detail.

Soft sensors can especially be used in data fusion, and during a process, it may happen that we cannot measure all signals and therefore sometimes reconciliation must be applied in the data to maintain the physical and chemical constraints. Data fusion and data reconciliation both help to build more robust and more accurate machine learning (ML) models. Another significant development opportunity for ML models is to augment the training dataset with artificial data, thereby increasing the number of data points. While each of these techniques aids in the use of models, their validation is of paramount importance. I will discuss both techniques in detail in the thesis.

This dissertation meticulously addresses the following pivotal thematic areas, each of which will be comprehensively elaborated upon in the subsequent chapters of this work. Importantly, each of these delineated areas represents a critical component in the development of sophisticated Industry 4.0 solutions for enhancing complex industrial processes:

- Chapter 2 serves as a foundational component, presenting a comprehensive literature review. This review meticulously explores soft sensors within the broader context of Industry 4.0 solutions, detailing their intricate development and showcasing their diverse industrial applications. It further offers an in-depth analysis of recent developmental trends in this field. The chapter consolidates this extensive research into a dedicated literature review table and a general diagram outlining soft sensor development, with several key publications highlighted for detailed discussion.

- In complex systems, models must be endowed with sufficient information to manage system complexity; consequently, my research focused on data re-conciliation in hierarchical time series. Chapter 3 presents the ensuring hierarchical consistency, particularly aggregation constraints, that is critical in system modeling, yet independent modeling at each level introduces inherent errors. To address this, an optimal data reconciliation technique, accounting for both measurement and modeling errors, is essential. This study investigated three distinct machine learning (ML) approaches: independent ML modeling without reconciliation, reconciliation incorporating measurement errors for ML development, and direct fine-tuning of ML predictions based on their errors. Through three case studies of varying complexity—mineral composition (9 elements), retail sales forecasting (14 elements), and waste deposition forecasting (>3000 elements)—the third method consistently demonstrated superior performance, enabling the development of more reliable ML models.

- A data fusion methodology was developed to improve the performance and robustness of the models. Complex-level ensemble fusion (CLF) is presented as a two-layer chemometric algorithm that jointly selects variables from concatenated mid-infrared (MIR) and Raman spectra with a genetic algorithm, projects them with partial least squares and stacks the latent variables into an XGBoost regressor, thereby capturing feature- and model-level complementarities in a single workflow. When benchmarked against single-source

models and classical low-, mid-, and high-level data-fusion schemes, the CLF technique consistently demonstrated significantly improved predictive accuracy. Evaluated on paired Mid-Infrared (MIR) and Raman datasets from industrial lubricant additives and RRUFF minerals, CLF robustly outperformed established methodologies by effectively leveraging complementary spectral information. Mid-level fusion yielded no improvement, underscoring the need for supervised integration. These results constitute the first evidence that a stacked, complex-level scheme can surpass all established fusion levels on real-world spectroscopic regressions comprising fewer than one hundred samples and provide a transferable recipe for building more accurate and resilient soft sensors in quality-control and geochemical applications (Chapter: 4).

- In the case of incomplete training datasets, a possible solution is to augment sections with missing data points or fewer data points with additional data. Research results in this area are presented in Chapter 5. The challenge of limited and unevenly distributed spectral data in industrial rock analysis is addressed by developing a method for generating artificial infrared spectra. By establishing a relationship between rock solubility and infrared spectra using Principal Component Analysis (PCA) and neural networks, we efficiently reproduced existing and generated new, constrained synthetic samples. The reliability of this method was confirmed through comparisons of original and artificial spectra. This framework provides a transferable solution for creating new, physically meaningful samples crucial for robust ML model development in data-scarce scenarios (Appendix: A.1).

- Additionally to the usability, performance, and artificially generated data of the models, the life cycle of the models is essential. The performance of the models can deteriorate over time, which is why monitoring and improving them in a short time is of utmost importance during Industry 4.0 processes. Chapter 6 presents a comprehensive framework for the full lifecycle management of machine learning (ML)-driven soft sensors in complex chemical processes. Leveraging Industry 4.0 technologies like ML, edge computing, and cloud services, the framework offers innovative solutions for difficult-to-measure laboratory variables. The primary goal is to ensure continuous product quality forecasting and stable plant conditions, supporting efficient and eco-friendly laboratory operations. Addressing existing challenges in

model maintenance and version control, the framework provides a structured methodology validated through real laboratory data. It enables continuous performance monitoring and data expansion, ensuring access for quality assurance engineers to highly accurate and updated data-driven models.

- The previous sections have detailed a range of state-of-the-art Industry 4.0 solutions. The Chapter 7 comprehensively summarises the research's objectives, detailing the initial aims and scope of the study. This comprehensive overview thus outlines both the potential applicability of Industry 4.0 integration and its tangible progress in different industrial environments.

- In Chapter 8 presents the key findings and significant results obtained, highlighting the primary goals and contributions of this work. Finally, potential next steps and future research directions are outlined, suggesting avenues for further exploration and development arising from these findings.

Chapter 2

Literature review on soft sensors in industrial applications

2.1 Introduction to soft sensors

In modern industrial systems, especially within the context of Industry 4.0 and smart manufacturing, the demand for real-time monitoring and control of key process variables has become more crucial than ever. However, direct physical measurement of certain variables (e.g., chemical composition, product purity, reaction kinetics) can often be technically challenging, economically unfeasible, or unsafe due to harsh environmental conditions. These limitations have led to the development and widespread application of *soft sensors*—mathematical constructs designed to estimate unmeasurable or difficult-to-measure variables using easily accessible measurements from physical sensors. Despite their growing prevalence over the past decade, a significant challenge remains in their economic operation and lifecycle management, limiting widespread adoption [2].

Developing methodologies for cost-, energy-, and resource-efficient soft sensor models is essential for continuous real-time monitoring [3, 4]. Current practices often rely on infrequent manual sampling and laboratory analysis, leading to insufficient data for effective process monitoring and control [5]. Furthermore, models built on small, statistically inadequate datasets can yield unsatisfactory accuracy, underscoring the necessity of thorough statistical exploration and analysis during initial modeling phases [6].

Soft sensors (also known as virtual sensors or software sensors) bridge the gap between theoretical modeling and empirical data-driven analytics. They utilize input variables that are available in real time—such as temperature, pressure, flow rate, or spectral signals—to infer latent process states or quality attributes. This capability is particularly significant in scenarios where traditional hardware-based instrumentation cannot operate effectively, whether due to physical constraints, high maintenance costs, or delays caused by laboratory analyses.

As industries progress toward digital transformation, soft sensors serve as essential enablers of process intensification, autonomous control, predictive maintenance, and quality assurance. They play a key role in reducing waste, enhancing safety, and driving energy-efficient operations, all of which align with the broader objectives of sustainability and competitive advantage. This chapter offers an extensive literature review on the development and application of soft sensors in industrial settings. It is structured to guide the reader through the conceptual foundations and historical evolution of soft sensors, followed by a detailed comparison of different modeling methodologies, including white-box, black-box, and grey-box approaches. The discussion then moves to the critical role of data fusion and its various classification levels. Illustrative industrial case studies from diverse sectors are presented to contextualize these concepts. Finally, the chapter identifies key technological and methodological challenges, along with future research directions and gaps that this dissertation aims to address.

2.2 Historical evolution and classification of soft sensors

The genesis of soft sensors lies in control theory and state estimation techniques. Early work during the 1960s and 70s focused on state observers and Kalman filters, which allowed estimation of unmeasurable internal variables in dynamic systems. These *white-box models* relied heavily on accurate physico-chemical knowledge of the processes and often involved differential equations grounded in first-principles. While foundational, these methods were sensitive to model mismatches and parameter uncertainties — issues that limited their practical use in rapidly changing or poorly understood environments. [7]

In the late 1990s and early 2000s, driven by the growing availability of industrial data and computational resources, a shift toward empirical *data-driven* models occurred. Soft sensors based on statistical learning and machine learning methods, such as Partial Least Squares (PLS), Principal Component Regression (PCR), and Artificial Neural Networks (ANN), started to replace purely model-based solutions. These models required no prior knowledge of the underlying process mechanisms and instead learned patterns directly from historical datasets.

More recently, the emergence of hybrid models — often referred to as *grey-box models* — has brought about a new paradigm by combining mechanistic understanding with machine learning flexibility. These models aim to retain the interpretability and physical realism of white-box models while leveraging the robustness and adaptability of black-box approaches.

Soft sensors can be broadly categorized into three methodological paradigms:

- **Model-Driven (White-Box)**: These rely on first-principles and physical equations to estimate process states. While interpretable and theoretically sound, they require extensive domain knowledge and are often limited by modeling complexity and inaccuracies in system parameters.
- **Data-Driven (Black-Box)**: Employing statistical and machine learning methods, these models predict outputs based purely on correlations in historical process data. Their advantages include adaptability, minimal process knowledge requirements, and ease of implementation. However, they often suffer from limited extrapolation capability and are considered "black-box" in nature.
- **Hybrid (Grey-Box)**: These combine mechanistic and empirical components. A typical example involves modeling the residuals of a physics-based prediction using a data-driven method. Hybrid models have shown improved generalizability and robustness, especially in dynamic or nonlinear systems. [8]

Developing a robust and maintainable soft sensor is a multi-step process that integrates data engineering, algorithm selection, model training, and deployment strategies. Each phase requires a combination of domain knowledge, statistical insight, and software engineering capabilities.

The general methodology can be structured as follows:

1. **Data Acquisition and Integration:** Gathering relevant time-series data from SCADA systems, Distributed Control Systems (DCS), or cloud-based IoT devices. High-frequency, high-resolution data are critical for capturing process dynamics.

2. **Data Selection and Curation:** Selection of relevant time windows under steady-state or representative dynamic conditions. This step often requires collaboration with process engineers to understand operational boundaries and anomalies.

3. **Preprocessing and Feature Engineering:** Addressing issues like missing values, outliers, and multicollinearity. Techniques include normalization, smoothing, PCA, and feature transformation.

4. **Modeling and Validation:** Depending on the objective, suitable algorithms such as PLS, Random Forest, Gradient Boosted Trees (e.g., XGBoost), Support Vector Machines (SVM), or Neural Networks are selected. Validation is conducted using cross-validation, bootstrapping, or time-series holdout methods.

5. **Deployment and Monitoring:** Final models are deployed in real-time environments using cloud applications, edge computing platforms, or integrated directly into industrial control systems. Continuous performance monitoring and periodic retraining are critical for long-term viability.

Figure 2.1 provides a summary of the general methodology for soft sensor development, showing a schematic diagram of the relationships between the individual elements.

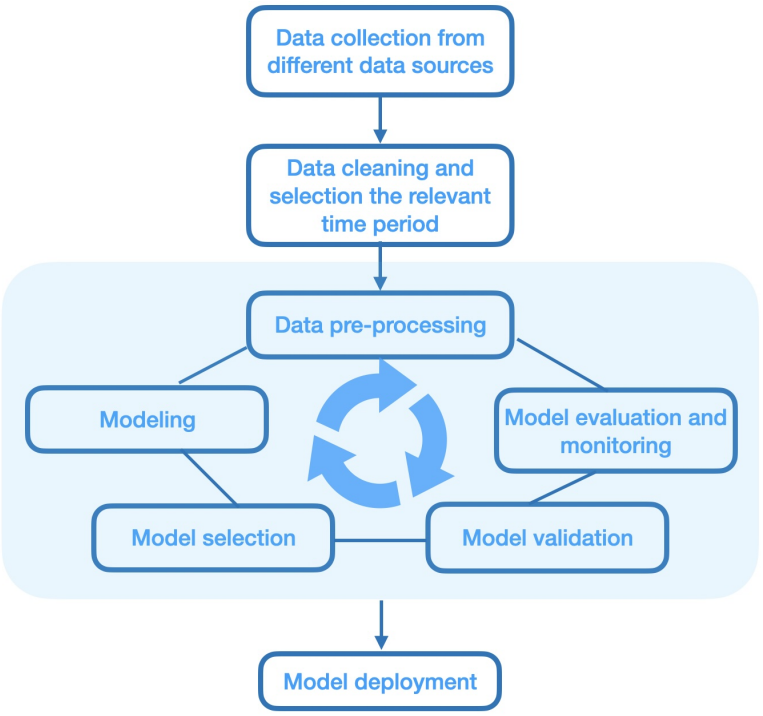


FIGURE 2.1: General methodology for soft sensor development

2.3 Overview of Industrial Soft Sensor Applications

My literature review, conducted on the Scopus database, utilized a keyword map to visually show the connections between soft sensors and key concepts in chemistry, chemometrics, and Industry 4.0. Distinct color clusters highlight these relationships: yellow for soft sensors and machine learning/IoT integration, red for mathematical modeling in bioreactors, green for real-time spectroscopic data and chemometrics, and blue for reproducibility in polymer science using soft sensors and chemometric methods. The literature, relevant publications concerning soft sensor development, are summarised regarding the industrial application in Table 2.1, where they are compared based on three criteria: research area, methodology, and the proposed solution's input data.

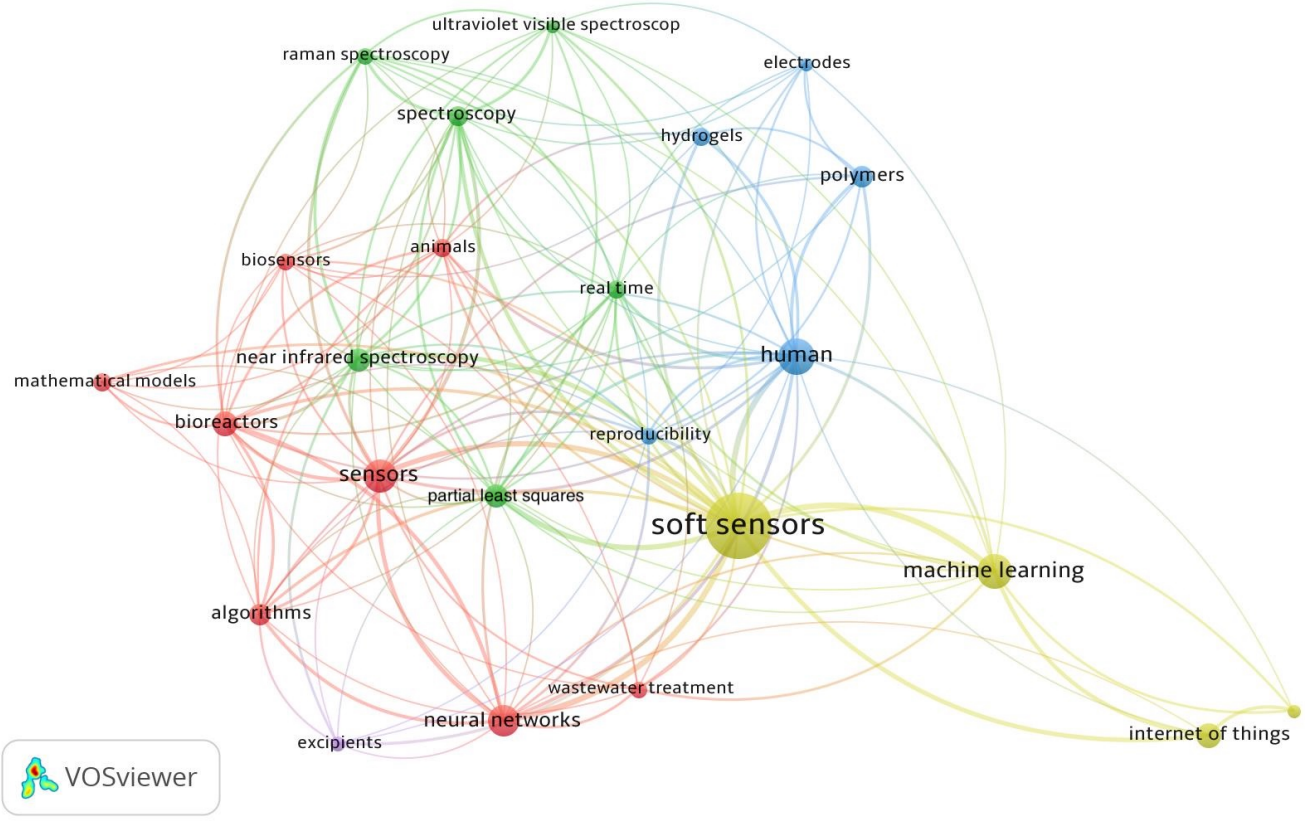


FIGURE 2.2: Network diagram of keywords soft-sensor or software sensor and chemistry or chemometrics or Industry 4.0 on Scopus.

TABLE 2.1: Comparison of literature studies about industrial application of soft sensors.

	Research area	**Method**	**Input and solution**	**Ref.**
1	Light naphta API and RVP parameter	ANFIS model	Iterative input selection minimizes squared error	[9]
3	PET viscosity	ANNSES	Sewage treatment and focusing on noisy and nonlinear data handling	[10]
4	Quality prediction of hydrocracking input and output	Review (PCR, PLS, ANN)	Spetroscopy data (NMR, UV-Vis, NIR, MIR, Raman)	[11]
5	FCC, VGH quality prediction with data driven	outlier detection method	sensor for product composition	[12]
6	Diesel key quality parameters (Flash point, FAME, CFPP ...)	AutoML and different PLS	Spectroscopy data (ATR-FTIR), flowrate, temperature and pressure	[13]
7	Kerosene stream from hydrocrack plant quality estimation	LWDA_SAE	Pressure difference of reactor, total amount water, temperature of hydrotreating reactor	[14]
8	Optimize the oil production in deep-water	EAND and RF	Process temperature, pressure at different locations	[15]
9	Real time prediction moisture and fat content of the olive	three-layer ANN	Flow, temperature, coadjuvants addition, water dilution level, NIR spectra exit of the decanter	[16]
10	Real time water quality prediction BOD	IBK	IoT sensors, propose a system architecture for soft sensor	[17]
11	SCGP process	FIR-CNN	Dynamic Soft Sensor model development	[18]
12	Crude oil fraction yield evaluation	Ensemble-deep-learning, NNR, CNN	1H NMR spectral data, increasing data redundancy and generating virtual samples	[19]
13	Manufacture of monoclonal antibodies in biotherapeutics	Review (PCA, PLS)	UV-vis, NIR, Raman, DLS, SLS, MALS, Fluorescence spectroscopy in PAT	[20]

This table effectively illustrates the diverse applications of soft sensors across various domains and the wide array of methodologies employed, ranging from simple to highly sophisticated or complex algorithms. The predominant solutions and input data involve spectral signals, which enable rapid and highly accurate measurements under industrial conditions. Beyond conventional sensors, the reviewed literature also highlights the integration of IoT devices, often fulfilling edge computing roles. Overall, Table 2.1 presents a representative subset of the comprehensive literature review, meticulously curated to include examples from a broad spectrum of industrial sectors.

ANN-based soft sensors have been successfully applied to real-time viscosity prediction in polymer plants, proving effective in handling noisy, nonlinear data [10]. Similarly, the Adaptive Neuro-Fuzzy Inference System (ANFIS) model has shown superior performance over traditional regression for predicting key light naphtha parameters like API gravity and RVP [9]. Spectroscopic data (IR, NIR, UV, Raman) are also vital inputs for soft sensor models in hydrocracking facilities, enabling the prediction of critical quality parameters for both feedstock and products. This highlights the need for a holistic approach that integrates automation, advanced soft sensors, and comprehensive diagnostics for efficient plant operations [11]. Mojo *et al.* developed soft sensors for oil refining units (FCC and VGH), using outlier detection to improve data quality. Their models achieved a 19% increase in efficiency, highlighting the value of soft sensors in complex industrial environments [12]. Within refineries, Partial Least Squares Regression (PLSR) is a widely used supervised learning method for soft sensors, while Principal Component Analysis (PCA) is a prominent unsupervised learning algorithm [10]. D.C.M. Souza *et al.* evaluated AutoML and PLS models using ATR-FTIR spectra for diesel analysis, employing Monte-Carlo double cross-validation to assess their predictive performance [13].

Based on the literature reviews, it is clear that soft sensors are becoming increasingly prevalent in chemical plants. However, these studies typically focus on the digitization of a single plant and build models using the available sensor signals, such as spectral data. What is missing from this literature is a framework that is usable in an industrial environment in near real-time, with extended robustness and usability of the machine learning models.

Chemometrics-driven models can significantly enhance quality assurance within Industry 4.0 advancements. However, literature reviews indicate that the lifespan

and robustness of these machine learning (ML) models can vary over time.

This dissertation addresses these limitations by introducing several mitigating elements. Chapter 3 presents a hierarchical data reconciliation technique designed to improve model utility by accounting for known model errors, supported by three case studies. Chapter 4 explores various data fusion techniques to enhance the accuracy of ML models, demonstrated through two case studies. Chapter 5 details the methodology and utility of artificially generated data, specifically spectra. Finally, Chapter 6 provides a comprehensive framework, based on CRISP-ML, to meticulously outline the applicability, lifecycle utility, and maintainability of ML models developed for industrial applications.

Chapter 3

Data reconciliation-based hierarchical fusion of machine learning models

3.1 Introduction

Digitalization offers significant opportunities through various machine learning (ML) algorithms that can replace conventional quality assurance methods. When developing ML models for processes, only certain parameters can typically be estimated with adequate precision. These ML model estimations carry varying degrees of error. Our modeling approach integrates engineering insights and additional data to ensure that the estimates meet specific constraints or conditions. The balance equations between hierarchical levels must be satisfied in the context of modeling hierarchical systems, and these constraints should be considered when training models.

Frequently, the systems to be modeled are naturally organized in hierarchical structures, mainly in forecasting problems; for instance, the demand for a product can be recorded on different hierarchy levels, like on a store, regional, or country level [21]. The measurements and observed values at each level will add up to the higher levels, which is called "coherence" [22]. In practical solutions, model development and prediction occur independently at each hierarchy level, resulting in incoherence in the results, so the predictions do not aggregate well.

This incoherence or balance error between the hierarchy levels was handled myopically with "bottom-up", "middle-out", and "top-down" methods. With these methods, it is not necessary to develop a model for each hierarchy level, but only one model is developed, and then we aggregate and/or disaggregate the predictions to other hierarchy levels [23]. For instance, the "bottom-up" technique works by developing models at the most granular level of the hierarchy and then summing lower-level predictions [24].

The fact that these myopic methods do not consider some useful information on other hierarchy levels is partly solved by combination approaches, which use statistically weighted information from all hierarchical levels [25]. However, none of the mentioned approaches result in optimal reconciliation among the predictions [26].

Optimal reconciliation is based on generating independent models for all elements of the hierarchy, where the prediction results are incoherent. However, the next step is to perform optimal reconciliation to adjust the independent predictions, which will lead to the prediction being consistent with the hierarchical structure [27]. In this case, reconciliation is formulated as a regression model which projects the elementary forecasts and predictions onto a subspace where the predictions adhere to the aggregation constraints [28]. Weighted squared-error optimal reconciliation was proposed to perform optimal reconciliation forecasts, where the base predictions are adjusted minimally due to the least-squares function [29, 30]. Generalized least squares and weighted least squares are also a solution that can be used to obtain the optimal reconciled forecasts and predictions, but it is really difficult to estimate the covariance matrices used in the solution. The easiest way to perform an ordinary least-squares estimate is to reconcile the base predictions [31]. These hierarchical forecasting methods worked well in different cases, like forecasting the electrical loads of a building [32], in supply chain forecasting [33] or in tourism forecasting [34].

The presented optimal reconciliation approaches only deal with the constraint due to the hierarchical system, but other system- and modeler-defined specifications and constraints can be imagined, too; for instance, the predictions must add up to a user-specified constraint, like the percentages must add up to 100 percent [35]. Data reconciliation (DR) is a general method used to implement and define both hierarchical constraints and any other necessary constraint [36]. DR enables the correction of measurements, reducing the associated uncertainty while satisfying different constraints [37].

The advantages of DR are also presented in various case studies, for instance, in analysis of measurement outcomes within the field of analytical chemistry, where the application of DR enhances predictions through real-time mass and element balancing. This strategy has been shown to decrease the standard error in predictions, eliminating the need for additional offline analyses [38]. The efficacy of the DR method in processing analytical chemistry data is clearly beneficial. In practical problems, maintaining variable consistency is crucial, and increased process automation requires rigorous monitoring. Errors are unavoidable in the measurement, processing, and transmission of signals. These errors can degrade the performance of the monitoring and control system and sometimes lead to process failures. Thus, minimizing these error effects is critical [39]. DR improves process measurement estimates by mitigating the impact of random errors [40]. The outcomes are independent and consistent, indicating that the value model at a higher hierarchical level produces results that differ from the aggregation of the lower-level models' results. In this case study, the ML models are structured hierarchically and subject to specific constraints.

In this study, a hierarchical modeling approach with optimal DR is presented that improves the performance of ML models to facilitate optimization with digital tools. With this approach, ML models can be developed that satisfy hierarchical and other user-specified constraints. Furthermore, with the expansion of software sensors in Industry 4.0 [41], the proposed method can be implemented in edge computing devices. Various case studies with different complexities were used to evaluate the proposed methodology. The main novelties of our work are summarized as follows:

- An approach to managing hierarchical constraints was developed to enhance the performance of ML models by accounting for the prediction errors in each model (see Section 3.2).
- In this study, a connection was established between the summation matrix utilized in hierarchical time series (HTS) forecasting and the incidence matrix employed in traditional DR methods (see Section 3.2).
- The developed methods exhibit strong performance in a range of case studies with different complexities. Our tests included a three-level scenario with 9 elements in rock composition estimation from spectral signals, a three-level

scenario with 14 elements in a distribution model (retail sales M5 competition), and a four-level waste deposition scenario involving more than 3000 elements (Hungarian counties, districts, and cities) (see Section 3.3).

3.2 Integrated correction of machine learning predictions using data reconciliation techniques

This section outlines the methods used in our approach to combine ML and DR techniques to improve the accuracy of ML predictions. By merely predicting each series separately, the hierarchical or grouping structure is ignored, resulting in forecasts that are not 'coherent', so they do not aggregate correctly and do not satisfy the hierarchical constraints.

3.2.1 Formulating the integration of machine learning and data reconciliation

The goal of this study is to correct ML predictions using optimal DR considering the modeling errors in hierarchical problems. Due to the hierarchical structure used, our method deals with multivariate modeling by considering multiple target variables ($\mathbf{y}$). The basic assumption is that all modeled variables are subject to error, so the model estimates for all variables need to or should be corrected. Consider hierarchical data, the target variables ($\mathbf{y}_t$) in the t^{th} sample instance can be written, in general, as follows:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_t \tag{3.1}$$

where $\mathbf{y}_t$ is the vector of the target variables, $\mathbf{x}_t$ represents the matrix of independent variables, $\mathbf{f}$ denotes the set of ML models, $\boldsymbol{\theta}$ represents the matrix of model parameters, and $\boldsymbol{\epsilon}_t$ is the prediction error vector. The model prediction ($\hat{\mathbf{y}}$) of the target variables can be written, in general, as follows:

$$\hat{\mathbf{y}}_t = \mathbf{f}(\mathbf{x}_t, \boldsymbol{\theta}) \tag{3.2}$$

Given that the measurements must adhere to the specified constraint, it is imperative that the model predictions comply as well. To achieve this, DR is used in this study. This process adjusts the predicted variables minimally to ensure compliance with a set of model constraints. It aims to reduce the discrepancy between predicted and reconciled values while considering the variance of these variables and guarantees that the reconciled parameters meet certain equality and inequality constraints [42]. Typically, the objective function for minimization is denoted as Equation (3.3). The general non-linear DR problem is outlined as follows:

$$\min_{\tilde{\mathbf{y}}_t} \left(\hat{\mathbf{y}}_t - \tilde{\mathbf{y}}_t\right) \mathbf{V}^{-1} \left(\hat{\mathbf{y}}_t - \tilde{\mathbf{y}}_t\right) \tag{3.3}$$

subject to

$$\mathbf{h}\left(\tilde{\mathbf{y}}_t\right) = \mathbf{0} \tag{3.4}$$

$$\mathbf{g}\left(\tilde{\mathbf{y}}_t\right) \leq \mathbf{0} \tag{3.5}$$

where $\mathbf{V}^{-1}$ is the inverse of the covariance matrix of errors, t is the sample instance, $\hat{\mathbf{y}}_t$ is a vector of model predictions, $\tilde{\mathbf{y}}_t$ is a vector of reconciled values for each target variable, $\mathbf{h}$ is a vector that describes the functional form of model equality constraints, and $\mathbf{g}$ is a vector that describes the functional form of model inequality constraints [36, 42].

Let us refer to the general HTS problem in Figure 3.1, where the relationship and subordination at the different hierarchy levels are illustrated [32]. For simplicity reasons, the measured and unmeasured variables were not labeled in the hierarchy. Every element (node) of the hierarchy (tree) is labeled as $y_{t,j}^{k,p}$, where $t = 1 \ldots \mathrm{n}$ denotes the sample instance, $k = 1 \ldots K$ means the level of the hierarchy, p presents the node's parent on the upper $(k-1)$ hierarchy level, and $j = 1 \ldots q_p^k$ is the number of child nodes at the k^{th} hierarchy level of the p^{th} parent.

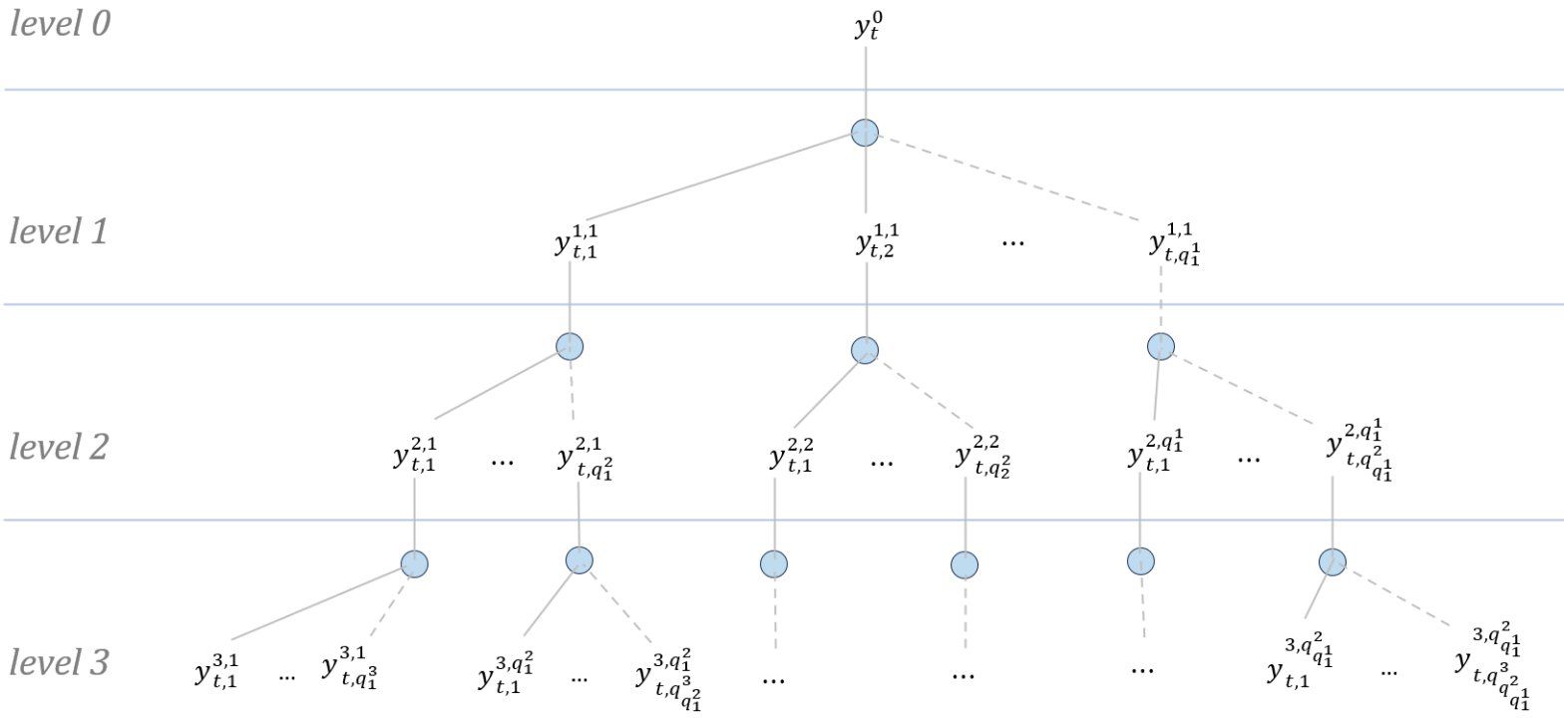


FIGURE 3.1: Schematic diagram of the three-level hierarchical system of the variable y. Level 3 contains all the target variable models; it is the level with the most detailed predictions, so the relationships of the third-level predictions are built into level 2, and the constraints of levels 2 and 3 are built into level 1. Even at level 0, the total prediction constraints determine the y value. Among the subscripts of y, the first, t, is the time, and the second, $j = 1 \ \ldots \ q_p^k$, is the number of child nodes at the given level. Among the superscripts, the first, $k = 1 \ \ldots \ K$, is the number of hierarchical levels, and the second, p, denotes the number of parent nodes.

This hierarchical structure means that the element at the root level (y_t^0) is the sum of the elements in the first level of the hierarchy, formally

$$y_t^0 = \sum_{j=1}^{q^{1,1}} y_{t,j}^{1,1} \tag{3.6}$$

where $q^{1,1}$ is the number of child nodes at the first hierarchy level.

A general element on the k^{th} level ($y_{t,j}^{k,p}$) is the sum of its child nodes, formally written as

$$y_{t,j}^{k,p} = \sum_{i=1}^{q_j^{k+1,j}} y_{t,i}^{k+1,j} \tag{3.7}$$

The number of elements in the k^{th} hierarchical level (q^k) can be formalized as the sum of the number of child nodes in the upper hierarchy level, as follows:

$$q^k = \sum_{p=1}^{q^{k-1}} q_p^k \tag{3.8}$$

The number of nodes in the hierarchy is obtained if the number of nodes in each hierarchy level is added together.

$$q = \sum_{k=0}^{K} q^k \tag{3.9}$$

By stacking all the tree elements in a vector ($\mathbf{y}_t$) based on Figure 3.1, the following is obtained:

$$\mathbf{y}_t = [y_t^0, \underbrace{y_{t,1}^{1,1} \dots y_{t,q_1^1}^{1,1}}_{\mathbf{y}_{t,1}^T}, \underbrace{y_{t,1}^{2,1}, y_{t,q_1^2}^{2,1} \dots y_{t,1}^{2,q_1^1}, y_{1,q_1^2}^{2,q_1^1}}_{\mathbf{y}_{t,2}^T}, \dots, \underbrace{\dots}_{\mathbf{y}_{t,K}^T}]^T \tag{3.10}$$

where $\mathbf{y}_{t,K}$ represents the elements on the bottom hierarchy level (K^{th} level).

Every element at each hierarchy level can be calculated using a summation matrix ($\mathbf{S}$) and the bottom-level elements ($\mathbf{y}_{t,K}$) as follows:

$$\mathbf{y}_t = \begin{bmatrix} y_t^0 \\ \mathbf{y}_{t,1} \\ \vdots \\ \mathbf{y}_{t,K} \end{bmatrix} = \mathbf{S}\mathbf{y}_{t,K} \tag{3.11}$$

The coherence requirements within the hierarchy can be defined with the help of a summing matrix ($\mathbf{S}$), which dictates the method in which the bottom-level series aggregate at higher hierarchy levels. The element of the summation matrix ($s_{i,j}$) refers to the tree node element i and tree leaf element j. This sets the matrix element of a given node i to 1 or 0 depending on if it is an ancestor of the leaf element j or not, respectively. The dimensions of $\mathbf{S}$ are $q \times q^K$, where q represents the total number of nodes in the hierarchy and q^K represents the number of nodes at the bottom level.

For the optimized reconciliation of the samples, a connection must be established between the summation matrix ($\mathbf{S}$) and the equality constraint of the DR technique at each hierarchical level (Equation (3.4)). Due to the hierarchical structure, this equality constraint can be formalized as $\mathbf{A}\tilde{\mathbf{y}}_t = 0$, where $\mathbf{A}$ is the incidence matrix. To create matrix $\mathbf{A}$, with dimensions $q \times q$, the [$\mathbf{0S}$] $q \times (q\text{-}q^K)$ null matrix and $q \times q^K$ summation matrix ($\mathbf{S}$) need to be subtracted from the $q \times q$ identity matrix ($\mathbf{I}$).

$$\mathbf{A} = \mathbf{I} - [\mathbf{0}\,\mathbf{S}] \tag{3.12}$$

Based on this, the following relation can be obtained to formalize the linear constraint for DR using the summation matrix:

$$\mathbf{A}\tilde{\mathbf{y}}_t = (\mathbf{I} - [\mathbf{0}\,\mathbf{S}])\tilde{\mathbf{y}}_t = 0 \tag{3.13}$$

We also consider additional linear equality constraints which need to be satisfied in addition to the hierarchical constraints, as shown in Equation (3.14):

$$\mathbf{A}^*\tilde{\mathbf{y}}_t = \mathbf{b}^* \tag{3.14}$$

Each linear constraint can be grouped, as shown in Equation (3.15):

$$\begin{bmatrix} \mathbf{A} \\ \mathbf{A}^* \end{bmatrix} \tilde{\mathbf{y}}_t = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}^* \end{bmatrix} \tag{3.15}$$

and, in a more compact form, we can rewrite this as follows:

$$\tilde{\mathbf{A}}\tilde{\mathbf{y}}_t = \tilde{\mathbf{b}} \tag{3.16}$$

The analytical solution of the DR (Equation (3.3)) with linear constraints (Equation (3.16)) is the following:

$$\tilde{\mathbf{y}}_t = (\mathbf{I} - \mathbf{V}^{-1}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T)^{-1}\tilde{\mathbf{A}})\hat{\mathbf{y}}_t + \mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T)^{-1}\tilde{\mathbf{b}} \tag{3.17}$$

where $\mathbf{I}$ is the identity matrix, $\tilde{\mathbf{A}}$ is the incidence matrix, and $\tilde{\mathbf{b}}$ are the constraint values. Equation (3.18) can be written in a more compact form as follows:

$$\tilde{\mathbf{y}}_t = \mathbf{P}\hat{\mathbf{y}}_t + \tilde{\tilde{\mathbf{b}}} \tag{3.18}$$

where $\mathbf{P}$ is the projection matrix and $\tilde{\tilde{\mathbf{b}}}$ is the correction term needed to satisfy the linear constraints.

$$\mathbf{P} = (\mathbf{I} - \mathbf{V}^{-1}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T)^{-1}\tilde{\mathbf{A}}) \tag{3.19}$$

$$\tilde{\tilde{\mathbf{b}}} = \mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{V}^{-\mathbf{1}}\tilde{\mathbf{A}}^T)^{-1}\tilde{\mathbf{b}} \tag{3.20}$$

In the case of linear regression models, Equation (3.21) becomes the following:

$$\tilde{\mathbf{y}}_t = \mathbf{P}\boldsymbol{\theta}\mathbf{x}_{e,t} + \tilde{\tilde{\mathbf{b}}} \tag{3.21}$$

where $\boldsymbol{\theta}$ represents the model parameters and $\mathbf{x}_{e,t}$ represents the extended input features at the t^{th} sample instance [36].

3.2.2 Methods for integrating machine learning and data reconciliation techniques

Based on the available information, ML models can be developed in three ways, as presented in Figure 3.2. The first route is to develop an ML model without any DR. The second route is that if there is information about the measurement uncertainty, then the inverse covariance matrix can be defined with the standard deviation of the measurement errors, and then DR can be performed on the measured dataset. Since the measurements are independent from each other, the covariance matrix for DR contains non-zero values only in the diagonal elements. After DR, ML models can be developed using the reconciled dataset.

The third route is when there is no information about the measurement uncertainty, but the prediction errors can be used to fine-tune the model predictions.

In this case, first, the ML models are trained on the raw dataset of the measurements, and then they are reconciled with the model predictions, where the standard deviation of the prediction errors is included in the inverse covariance matrix used to perform DR. In this case, we assumed that the modeling errors are independent of each other due to the fact that an independent ML model was developed for each node. Therefore, in this case, the developed covariance matrix for DR contains non-zero values only on the diagonals. Besides this assumption, there can be cases where the modeling errors are not independent, and hence the covariance matrix for DR should be developed with care in these cases.

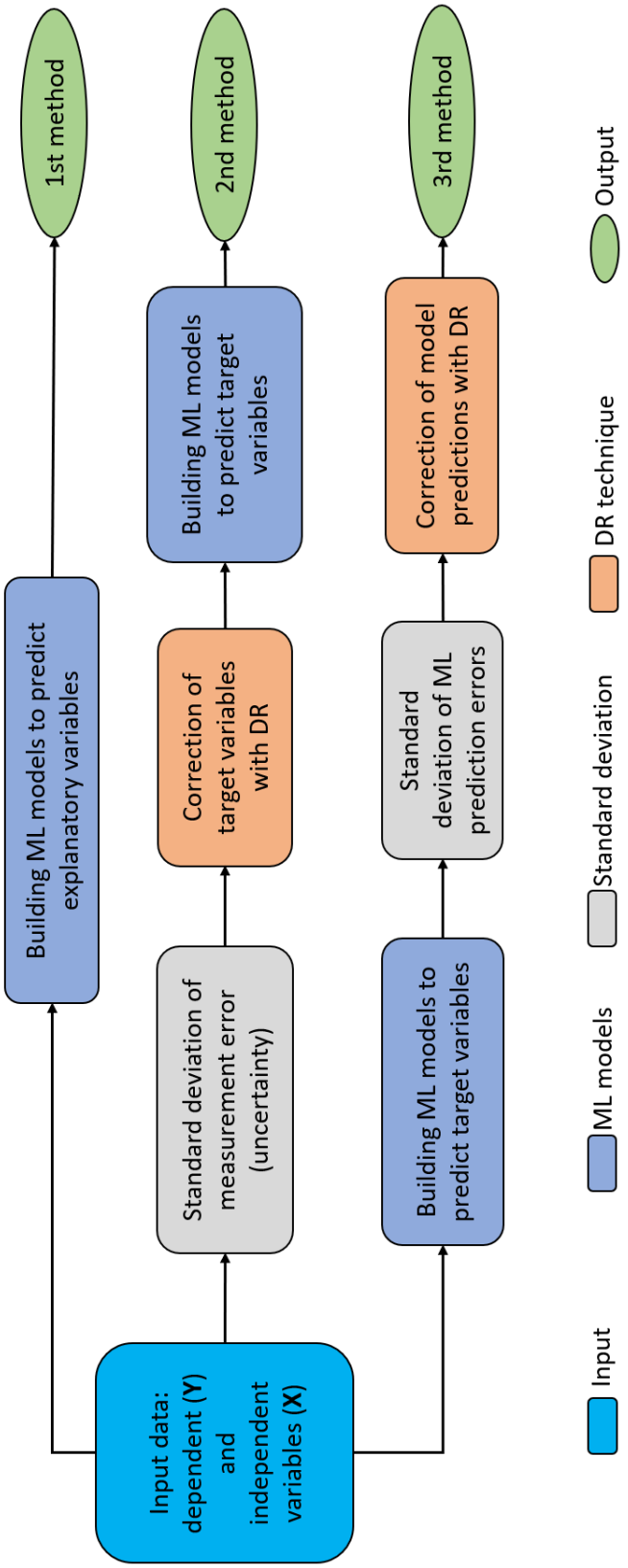


FIGURE 3.2: The three different applications of this methodology; the cyan box shows the input data, the blue boxes show the ML model predictions, the gray boxes show the uncertainty representation for DR, the orange boxes show the step of performing DR, and the green boxes contain the outputs of the methods.

3.3 Modeling results for cases of varying complexities

Hierarchical modeling with optimal DR was performed in three different case studies with different complexity levels. The first case study is about predicting the mineral composition of different rock samples from spectral data to support oil exploration. This problem includes three hierarchy levels with nine elements.

In the second case, a Walmart commercial dataset included in a Kaggle competition was investigated, where reconciliation was performed between time series predictions on three hierarchy levels with 14 elements. The third case also includes an HTS analysis dataset used to predict regional waste deposition in Hungary on four hierarchy levels with more than 3000 elements.

3.3.1 Mineral composition of the rock samples

The first case study uses the results of data generated during oil exploration and production as input data. The traditional approach to determining the mineral composition of rock samples has been to take field samples/drill cores collected during field investigations to the laboratory. Analytical data are prepared using traditional laboratory measurement methods, which are costly and time consuming, may include the use of hazardous substances, and are challenging due to the exploration of large areas [43]. From the point of view of technology development, the presented methodology is suitable for obtaining an accurate picture of the analysis during oil field drilling. This is a complex task for geologists in oil exploration and production. Geological analyses classify the different storage rocks into separate categories, which is particularly important from the point of view of geological interpretations [1]. Furthermore, knowing the detailed composition of the elements is also essential before executing the various efficiency-enhancing procedures of the wells that are already in operation.

Fourier-transform infrared spectroscopy (FTIR) measurements can also determine the composition and properties of rocks. In these cases, chemometric (ML) models are necessary to analyze spectra [1]. These models offer the possibility of installing software sensors in drill heads, which allows for a detailed analysis of the composition of rocks in the field [44].

The mineral composition of the rocks used in the modeling is represented by mass percentage values, which must total 100%. Our task involves a hierarchical regression process to predict the specific mineral composition of the rocks, ensuring their sum equals 100%. The aims of this methodology are to develop an algorithm that outperforms traditional models while meeting the initial conditions. In this study, ML models were constructed for determining the mineral composition of rock samples. The target (dependent) variables $\mathbf{y}$ of the model are derived from X-ray diffraction (XRD) measurements, while the independent variables $\mathbf{X}$ are derived from FTIR spectra.

In the first method, a partial least-squares (PLS) regression model was developed, creating a distinct model for each element's composition. In the second method, DR was applied considering the uncertainty associated with traditional XRD measurement errors. For the third method, DR was performed on the model predictions using the uncertainty of these predictions. The training dataset consisted of 618 samples and applied 10-fold cross-validation to fine-tune the PLS parameters. The refined models were then evaluated on 305 unknown samples.

The dependent variables are the mineral compositions obtained from different rocks. Despite these traditional laboratory measurements being performed with an XRD device, the total of the mineral compositions does not add up to 100, as expected. The DR method is also applicable to conventional laboratory measurements. To calculate the standard deviation of the XRD measurements, the measurement uncertainty associated with traditional measurements reported in the literature was taken into account. Since 2000, the Reynolds Cup has been held every two years, allowing various institutions and companies to participate in lap sample measurements [45]. From these measured data, an aggregated result is compiled in Table 3.1, which shows the standard deviation of the XRD measurements for the six mineral compositions [46].

TABLE 3.1: XRD measurement results [m/m %] from the 2006 Reynolds-cup [46].

	Actual	*Submitted**	*SD***
Quartz	29.9	30.67	0.77
K-feldpar & Plagioclase	8.6	8.0	0.6
Calcite	4.6	4.47	0.13
Kaolinite	15.0	15.7	0.7
Total clay without Kaolinite	20.2	19.17	1.03
Dolomite, Magnesite, Hematite Aragonite, Fluorite, Apatite ...	21.7	21.73	0.03

*The results contain the average result of the first three places of the 2006 Reynolds-cup.

**SD: standard deviation

The hierarchical levels related to the rock case study and the relationship between parent and child nodes are presented in Figure 3.3. This figure presents the hierarchical structure at the highest level (level 0), where the total of the compositions must always be equal to 100. At the subsequent level (level 1), $\mathbf{y}_{t,1}^{1,1}$ presents a silicate, and $\mathbf{y}_{t,2}^{1,1}$ presents carbonate rock types. The lowest level (level 2) includes the variables of the six mineral compositions. In this case study, siderite is represented by quartz, K-feldspar and plagioclase, calcite, and kaolinite. The total clay without kaolinite, along with dolomite, magnesite, and hematite, constitutes the carbonate group, and the t index in this case study is the number of samples, not a timestamp.

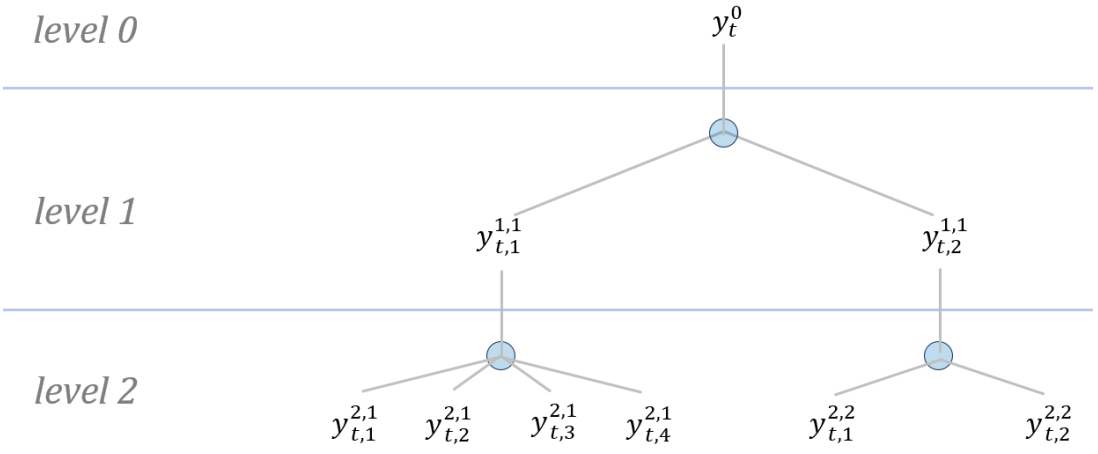


FIGURE 3.3: Schematic hierarchical diagram of the rock case study, and in this case study, the t index is the number of samples, not a timestamp.

The hierarchical relationship in the mineral composition is shown in detail in the summation matrix ($\mathbf{S}$).

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.22}$$

The incidence matrix ($\tilde{\mathbf{A}}$) is

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & -1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.23}$$

In this case study, the constraint that the sum of the silicate and carbonate content should be 100 is included in matrix $\tilde{\mathbf{b}}$.

$$\tilde{\mathbf{b}} = \begin{bmatrix} 0 \\ 0 \\ 100 \end{bmatrix} \tag{3.24}$$

$\mathbf{V}$ is a 8×8 matrix that includes the standard deviation of all variables in the diagonal of the matrix. So, the standard deviation of the measurements in the second method is taken from Table 3.1, and in the third method, it is the standard deviation of the ML prediction errors.

In the following, the results of the development of ML models for each method are presented through the average prediction errors and through the hierarchical balance error. The results of different hierarchical levels and different methods were compared with the root-mean-squared error of the prediction ($RMSE$) value, which indicates the accuracy of each estimate. At the first level, the third method performed the best for silicate ($RMSE = 4.577\%$), and the first method was the best ($RMSE = 4.154\%$) for carbonate, but here, the balance error was not met.

The most significant difference is seen between the kaolinite $RMSE$ values at the two levels, where the first method performed the best ($RMSE = 1.584\%$), but in this case, the balance errors were not met either.

The results for the rock samples are summarized in the bar plots in Figure 3.4.

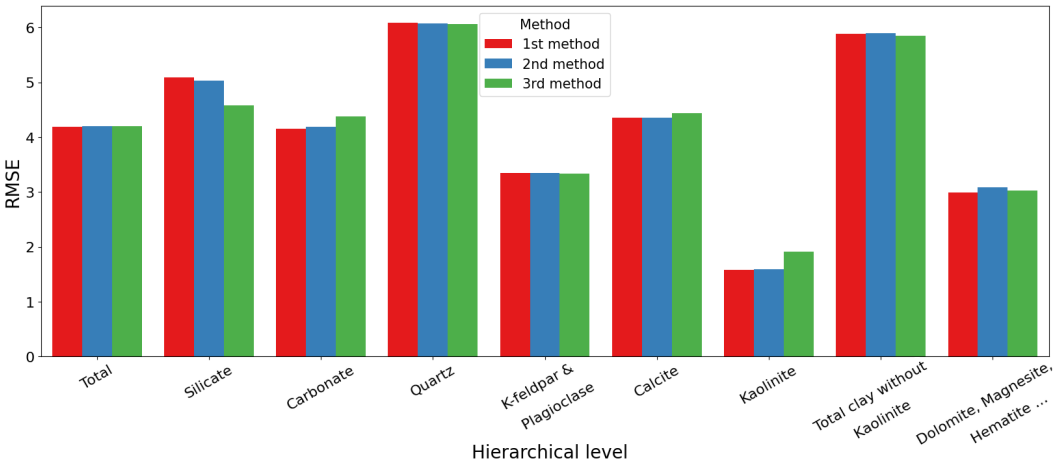


FIGURE 3.4: Modeling results of the three methods in predicting the mineral composition of rock samples evaluated on the test dataset for each element.

Table 3.2 presents a chosen subset from the dataset of 305 rock samples, illustrating the actual composition of the samples and the outcomes achieved by each method. At level 0, the DR error is 3.24 for method 1 and 2.88 for method 2, while the third method exhibits no DR error. At level 1, method 1 has an error of 3.24, method 2 has an error of 2.88, and method 3 shows no error. At the lowest level (level 2), method 1 has a DR error of 2.11, method 2 has an error of 2.01, and method 3 has no error. These results clearly demonstrate the advantage of the third method, as the initial conditions are also satisfied, considering the errors.

The developed method helps to better predict the rock composition of oil fields. It can also be used to balance the material flows of products produced during chemical processes. It is also suitable for monitoring and controlling chemical and separation technology processes. Incorporating the correlations between predicted values facilitates the embedding of ML models into an industrial environment. In addition, the DR technique increases the usability of the models.

TABLE 3.2: Prediction results of the first, second and third methods in case of rock use case.

level 0	**Total**					
y_real	100					
$\hat{y}$*_1st*	103.24					
$\tilde{y}$*_2nd*	102.88					
$\tilde{y}$*_3rd*	100					
level 1	**Silicate**				**Carbonate**	
y_real	79				21	
$\hat{y}$*_1st*	66.55				36.69	
$\tilde{y}$*_2nd*	66.20				36.68	
$\tilde{y}$*_3rd*	65.22				34.78	
level 2	**Quartz**	**K-feldp. & Plagio.**	**Kaolin.**	**Clay without Kaolin.**	**Calcite**	**Dolom. Magne. Hemat. ...**
y_real	51	14	2	12	11	10
$\hat{y}$*_1st*	37.57	12.81	1.67	12.19	19.32	14.33
$\tilde{y}$*_2nd*	37.56	12.73	1.60	12.72	19.36	14.02
$\tilde{y}$*_3rd*	37.70	13.04	2.16	12.32	19.78	15.00

3.3.2 Retail sales forecasting

The second case study discussed in this chapter involves an HTS analysis conducted on data from the Walmart shopping chain, part of a data analysis competition announced in 2020 as M5. The competition aimed to improve the accuracy of forecasting and empirically compare various forecasting methods [47]. The input data consist of unit sales for 3,049 products in the US.

These products are sold in 10 shops across three states (CA, TX, WI).

Locations such as states and shops were selected carefully, representing different waiting habits, purchasing dynamics, durable consumer goods, and fast- and slow-moving products. The M5 dataset is an excellent choice due to its meaningful hierarchies and cross-sectional levels. The daily data span from 29.01.2011 to 19.06.2016 (1969 days), with the training set data spanning from 29.01.2011 to 24.04.2016 (1913 days), a 27-day validation period spanning from 25.04.2016 to 22.05.2016, and a 28-day test period spanning from 23.05.2016 to 19.06.2016. Explanatory variables include calendar information, sales prices, and promotional activities [48]. Figure 3.5 illustrates a hierarchical diagram of the M5 case study, and the related incidence matrix is shown in Equation (3.26).

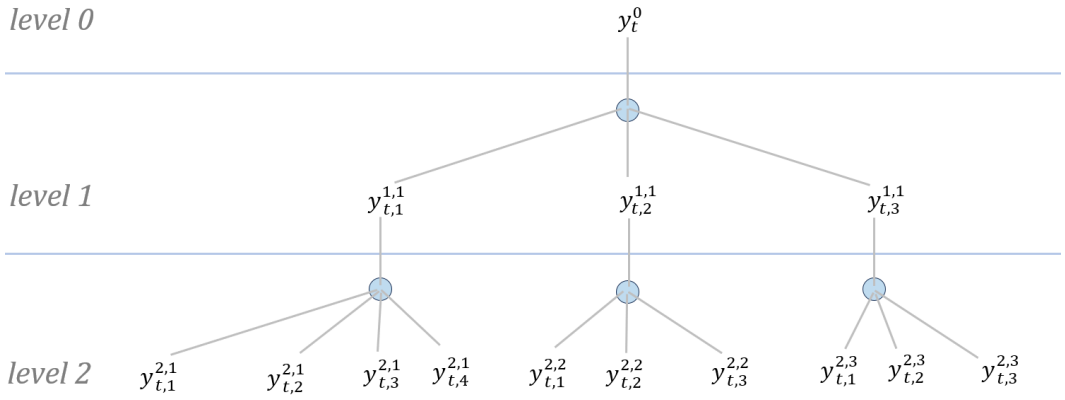


FIGURE 3.5: Schematic diagram of the two-level hierarchical structure's target variable ($\mathbf{y}$) in the M5 forecast—an accuracy competition. In the hierarchical levels, level 0 is the country level, level 1 is the state level, and the level 2 is the shop level.

In this case, the hierarchy of levels 0, 1, and 2 has ten elements at the lowest level (level 2: shop), three elements in a 4-3-3 split at the middle level (level 1: state) above, and one element at the top (level 0: country) in $\mathbf{S}$.

The hierarchical relationship in the sales dataset is shown in detail in the summation matrix ($\mathbf{S}$).

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.25}$$

The incidence matrix ($\tilde{\mathbf{A}}$) is

$$\tilde{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 \end{bmatrix} \tag{3.26}$$

In this case study, the sum of the lower levels should be output to predict the upper level. Matrix $\tilde{\mathbf{b}}$ can be written in the following form:

$$\tilde{\mathbf{b}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.27}$$

Figure 3.6 illustrates the performance of several forecasting models along with their hierarchical results. The metric used is the $RMSE$, displayed on the y axis. Different DR techniques are shown on the x axis, with colors indicating the hierarchical levels. In general, the middle-out method has the highest $RMSE$ (2.421), while the bottom-up method has the lowest (2.241). Despite the small overall differences in the $RMSE$, the bottom-up method is the best performer. At the country level, the top-down method has the highest $RMSE$ (5.408), whereas the bottom-up method, with a significantly lower $RMSE$ (4.822), is more efficient and suitable for country-level forecasting. At the state level, the top-down method has the highest $RMSE$ (2.911), while the bottom-up method has the lowest (2.487). In this category, the bottom-up method is the most effective. At the shop level, there are slight differences among all four methods. The top-down method has the highest $RMSE$ (1.656), while the bottom-up method has the lowest (1.569).

With the introduction and development of the analytical DR solution in this study, the overall $RMSE$ from the sales data is 2.305, which is in the mid-range compared to the other methods. At the country level, the analytical DR solution's value is 4.848, ranking second after the bottom-up method, indicating strong performance. At the state level, the DR analytical solution's value is 2.839, also showing good performance relative to the other methods. At the shop level, the DR analytical solution's value is 1.633, which is significantly lower than that of the other methods, though the RMSE of the bottom-up method is even lower. In

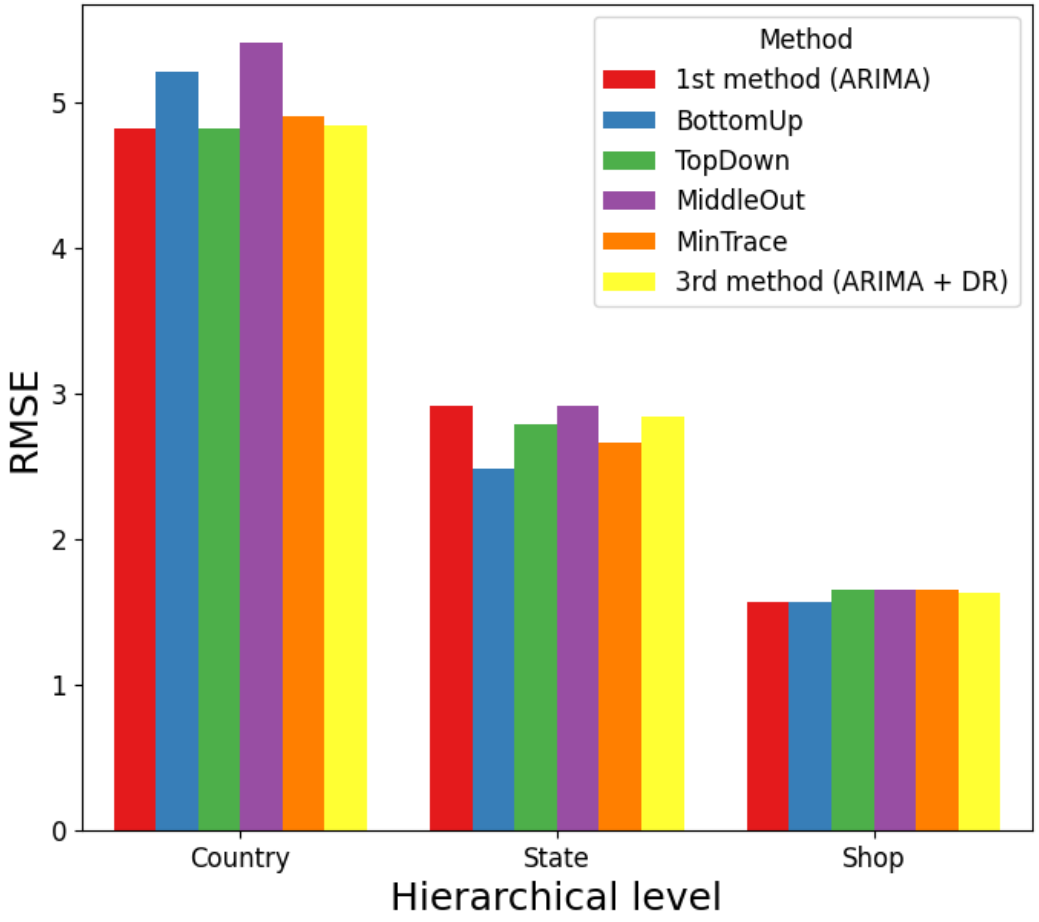


FIGURE 3.6: Results of the M5 competition and our own DR analytical solution compared based on RMSE values at different hierarchical levels.

this case study, the initial dataset satisfies the hierarchical balance and does not require reconciliation. Consequently, the second method is less pertinent. Only the first and third methods are suitable in this scenario. By emphasizing one of the forecasted outcomes on 18.05.2016, the results of the first and third methods are summarized in Table 3.3. The results show that the third method gives similar results to the first method, but the third method involves comparing the data. At the national level, the third method provides the best results, but bottom-up is the second worst method. At the state level, the bottom-up method gives the best result, with the third method in the middle. There are slight differences at the shop level, with the lowest error obtained by the DR techniques given by the bottom-up method and the second lowest by the third method.

When comparing the results tested on the data of 18 May 2016, the analytical solution of the HTS analysis using the DR technique (third method) showed little

TABLE 3.3: Prediction results of the first and third methods in case of sales use case.

level 0	**Country**									
y_real	22.0									
$\hat{y}\_1st$	14.102									
$\tilde{y}\_3rd$	13.938									
level 1	**CA**				**TX**			**WI**		
y_real	11.0				1.0			10.0		
$\hat{y}\_1st$	6.034				1.569			4.416		
$\tilde{y}\_3rd$	6.278				2.124			5.536		
level 2	**CA_1**	**CA_2**	**CA_3**	**CA_4**	**TX_1**	**TX_2**	**TX_3**	**WI_1**	**WI_2**	**WI_3**
y_real	2.0	4.0	1.0	4.0	0.0	1.0	0.0	1.0	3.0	6.0
$\hat{y}\_1st$	1.656	1.314	1.685	0.432	0.244	0.052	0.656	2.627	1.872	1.885
$\tilde{y}\_3rd$	1,729	1.376	1.879	1.294	0.458	0.544	1.122	2.442	1.501	1.593

difference from the results of the first methodology. Although the third methodology was among the top performers at the country level, it ranked in the lower half for the state and shop cases.

3.3.3 Waste management hierarchical time series prediction with data reconciliation

Waste management is a critical issue in Hungary due to the increasing amount of solid waste generated yearly. Effective waste management strategies require the accurate forecasting of waste generation, which can be achieved through time series analysis. Much of the literature focuses on predicting waste quantities as accurately as possible because precise solid waste forecasts are crucial for the circular economy, aiming to maximize recycling and enhance energy efficiency [49]. An integral part of these forecasts involves predicting waste amounts at collection points, which is vital from a workload perspective [50]. Eryganov *et al.* employed reconciliation in HTS forecasting, which is known to improve the quality of initial forecasts [51]. Moreover, HTS and DR technologies were applied in analyzing hazardous waste, utilizing a tree structure that mirrors the organizational layout of a region (including regions, micro-regions, and their sections), with the autoregressive integrated moving average (ARIMA) algorithm being used for predictions [52]. However, waste data are often incomplete and inconsistent, which results in unreliable predictions. HTS analysis has proven to be a powerful method for forecasting when dealing with multiple time series at various levels of aggregation. Additionally, DR techniques can enhance accuracy by resolving data inconsistencies. In this case study, HTS analysis and DR techniques were investigated to forecast solid waste generation within Hungary's hierarchical waste management system.

The solid waste data for Hungary were sourced from the Hungarian Central Statistical Office database. This comprehensive dataset includes the volume of solid waste at the national level for all 19 counties, including Budapest (the capital city), 175 districts, and 3,155 settlements. Consequently, the HTS analysis incorporates levels 0, 1, 2, and 3.

The modeling process used the Holt–Winters exponential smoothing (HWES) algorithm in this research, with annual data aggregation [53]. Data on solid waste quantities were collected from 2010 to 2022. The HWES models were trained from 2010 to 2019 and forecasted until 2022, and the measured data are available for 2020, 2021, and 2022. A settlement is selected in Table 3.4, and the forecast obtained by the first ML method and the third ML + DR method is displayed. The second method is not considered because in the first case, the balance error is met, so the DR of the input data of the ML models is not relevant in this case study. The reliability of the methods was examined for the years 2020, 2021, and 2022. The selected settlement is a randomly selected small town in Hungary, located in Veszprém County.

TABLE 3.4: Summary of the forecast results of the first and third routes of a selected settlement, its district and its county levels in the case of of amount of solid waste on Hungary in 2020, 2021, 2022 (tons).

	2020	**2021**	**2022**
level 0	**Total**		
y_real	3 301 482.3	3 350 245.0	3 215 457.4
$\hat{y}\_1st$	3 254 760.8	3 256 817.8	3 258 874.8
$\tilde{y}\_3rd$	3 254 503.5	3 256 560.6	3 258 617.7
level 1	**County**		
y_real	119 844.3	120 933.9	116 180.1
$\hat{y}\_1st$	116 495.1	117 577.4	118 659.6
$\tilde{y}\_3rd$	121 780.4	122 861.6	123 942.9
level 2	**District**		
y_real	15 558.7	16 903.2	15 751.4
$\hat{y}\_1st$	14 640.5	15 013.5	15 386.6
$\tilde{y}\_3rd$	14 767.7	15 140.1	15 512.5
level 3	**Settlement**		
y_real	1 341.7	1 427.4	1 355.0
$\hat{y}\_1st$	1 037.7	1 031.7	1 025.7
$\tilde{y}\_3rd$	1 038.3	1 032.3	1 026.3

Figure 3.7 shows the RMSE value of the model errors, taking into account the national level, including all counties, all districts, and all settlements, in 2020, 2021 and 2022. Figure 3.7 clearly shows that the biggest errors are at the county level, and these errors are lower if DR is applied after ML. Taking into account the years, the highest error is seen in 2021, and the lowest is seen in 2022.

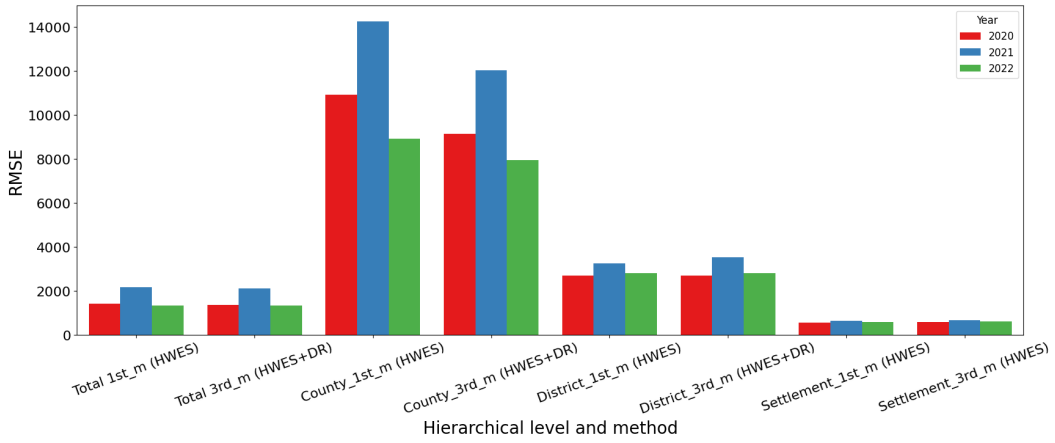


FIGURE 3.7: Results of ML prediction and our own DR analytical solution using data on the amount of solid waste in Hungary, compared based on $RMSE$ values at different hierarchical levels.

3.4 Chapter summary

This study introduces a novel approach that combines data reconciliation techniques with machine learning. In addition to the already widespread and established methods, this research formalizes a unique analytical approach. In the case of hierarchical systems, machine learning models do not satisfy various constraints arising from hierarchical structures, and the balance equations between hierarchical levels must be satisfied when modeling hierarchical systems. In this study, three different machine learning model development methods were investigated and compared: (1) model building without data reconciliation, (2) model building with reconciled measurement data based on the measurement uncertainty, and (3) model building with direct reconciling of the machine learning model predictions based on the modeling errors of each model.

In this research, each method was examined through three case studies of different complexities. The results show that the fine-tuning of model predictions with

optimal data reconciliation helps to improve the accuracy and reliability of model predictions, and all necessary constraints can be satisfied with this technique.

The presented method enables data integration from different measurement techniques, incorporating errors from different methods into the machine learning prediction process. In subsequent developments, this approach will allow us to create specific algorithms that take estimation errors into account and manage them efficiently. By remedying the imbalance, the goal was to develop more accurate, flexible, scalable, and usable models.

The limitations of the method presented in this study are that the hierarchical structure of the system, the quality of the data, and the reliability of the sources must be known precisely, and all the essential limitations of the system must be determined. If the hierarchical structure or constraints of the system change, the model structure must also be updated.

Chapter 4

Data fusion of spectroscopic data for enhancing machine learning model performance

4.1 Introduction

Our research presents and investigates various data fusion (DF) techniques to integrate results from different spectroscopic measurements. Chemometric models typically rely on laboratory analyses to predict a critical qualitative or quantitative parameter based on spectroscopic data. This study subjected samples to multiple spectroscopic techniques, and the resulting data sets were subsequently fused. The primary objective was to develop more robust and accurate machine learning (ML) models using the complementary information inherent in these diverse measurements. This approach demonstrated that the joint treatment of information from different sources can significantly enhance the performance of even relatively weak individual ML models.

Rapid, non-destructive spectroscopic techniques such as mid-infrared (MIR) absorption and Raman scattering have become indispensable in Process Analytical Technology (PAT) or soft sensor development for online quality control and real-time release testing. Each modality probes complementary molecular vibrations — MIR is highly sensitive to polar functional groups, whereas Raman is selective for symmetric, non-polar bonds — so combining them offers a richer chemical

fingerprint than either technique alone. Therefore, data fusion (DF) strategies promise substantial gains in predictive accuracy for soft sensors, especially when sample throughput or regulations limit the amount of reference chemistry that can be performed. Research shows that even small improvements in ML models can lead to significant gains in reducing production downtime, chemical usage, and even CO_2 emissions and reduce the random errors [54, 55].

DF techniques began to spread in the second half of 2017, after which they gained more and more emphasis in the work of quality assurance laboratories. Most scientific literature on DF techniques has been published in the food and pharmaceutical industries. The growing number of high-throughput multidimensional analytical instruments and chemical measuring devices that produce data sets of different dimensions provides a considerable amount of analytical information about complex samples and, at the same time, poses a challenge during data evaluation. The use of ML tools helps to make efficient use of these data volumes. Furthermore, the development of chemometrics revolutionized the steps in the interpretation of analytical processes and contributed to the solution of more complicated analytical problems. The DF techniques are necessary to examine the complexity of an analysis problem from several different perspectives. The goal is to create the most comprehensive picture possible of the reviewed materials. It is also essential for us to obtain as much information quickly and nondestructively using a small number of samples [1]. Quality assurance can be solved by combining the results of different instruments (such as MIR, NIR, Raman, XRF, NMR, and chromatographic analyses) with a broader characterisation of the given samples using Industry 4.0 technologies. The DF enables the simultaneous extraction of meaningful and valuable information from various analytical sources and it is also essential for us to obtain as much information quickly and nondestructively using a small number of samples.

Different analytical techniques provide various information about the same sample, highlighting some information in the case of Raman, MIR and near-infrared (NIR) spectroscopy (see the Appendix A.2) [56]. The different analytical techniques measure different bonds in molecules. In the case of Raman, homonuclear bonds, in the case of MIR and NIR, are polar bonds, with the addition that in the case of NIR, they are H-containing. The absorption bands in the case of Raman are scattered radiation and, in the case of MIR and NIR, absorbed radiation. The signal intensity is different in the case of Raman, and MIR/NIR is poor in the

former case and sound in the latter case. The absorption is strong in the case of Raman and weak in the case of MIR/NIR, which also shows that the methods can complement each other. Furthermore, interfering effects can be eliminated using the Raman and MIR/NIR methods because the Raman broad fluorescence baseline causes the interference, in the case of MIR/NIR is water. The application cases are offline and online; inline is good in the case of Raman and NIR and poor in the case of MIR.

Table 4.1 provides an overview of significant studies utilizing DF across various industries to tackle different challenges. This table details the usage of DF in various sectors, specifies the types of datasets employed, the ML techniques adopted, the measurements that helped create the top models, the DF methodologies executed to boost model performance, and the degree of improvement in model accuracy. During the table's compilation, emphasis was placed on ensuring that DF techniques were based on datasets originating from a wide range of sources and data derived from different laboratory experiments. While DF is commonly utilised in classification (qualification) and regression (quantification) techniques. The popular measurement techniques are the high-throughput spectroscopic tools, such as MIR, NIR, Raman, and the classic separation techniques, as chromatographic measurements. The literature in this table presents outcomes both with and without DF application. In many situations, only a single DF level (low, medium, or high) is used and compared to the traditional (non-DF) technique. Out of the fourteen cases summarized, just one showed a better outcome from the model without DF — predicting the Li content of rocks with a 2% accuracy improvement; in all other instances, DF application enhanced single model results. The caprolactam case notably demonstrated a 64.7% improvement in accuracy with high-level DF compared to the NIR based single model.

TABLE 4.1: Comparison of literature studies about improving the performance of data fusion models.

Dataset, Sample	**ML**	**Data source**	**Best DF type**	**Accuracy improvement**	**Ref.**
aglicultural, honey	classification	MIR, NIR, Raman	high-level	19%	[57]
aglicultural, honey	classification	NIR, HPLC	low-level	30%	[58]
mineral exploration, rock	regression	XRF	Li - single model, Zr - high-level	Li - 2% Zr - 17%	[59]
aglicultural, soothing herbs	classification	HPLC-UV, UV-Vis	mid-level	10-20%	[60]
oil, hydrocarbon	regression	Raman, NIR	mid-level	5.5%	[61]
mineral exploration, soil	regression	Vis-NIR, XRF	depends on target	10-15%	[62]
food, caprolactam	regression	MIR, NIR	high-level	64.7% for NIR, 0.1% for MIR	[63]
pharma, tablet	regression	MIR, Raman	high-level	25.9%	[64]
aglicultural, wheat flour	regression	MIR, NIR	mid-level	30-56%	[65]
food, beer bitterness	classification, regression	LC-MS pos. and neg. ionization	sustanable mid-level	18.2%	[66]
environmental protection, analyze aniline compounds	regression	UV-Vis multiple settings	high-level	40%	[67]
medicine, food	classification	NIR, HPLC	high-level	13.3% with SNV, 8.4% with MSC	[68]
aglicultural, soil	regression	Vis-NIR, MIR	mid-level (OPA)	4%	[69]
clay mineral, rock	classification	LIBS, Raman	low-level	3.6%	[70]

The DF techniques implemented across scientific disciplines can develop predictive models for quality control. From the literature review, three specific DF techniques have been pinpointed. These distinctions are mainly based on the stage at which data integration occurs within the processing pipeline:

- *Low-level fusion*: involves amalgamating raw data into one dataset, the foundation for feature selection and model development
- *Medium-level (Mid-level) fusion*: focuses on preprocessing separate datasets, merging the refined datasets, and constructing models based on the combined datasets
- *High-level fusion*: feature selection is executed independently on preprocessed datasets, which are subsequently unified, and models are developed using this integrated dataset

This hierarchical framework is schematically illustrated in Figure 4.1 [71]. It is important to note that the specific terminology and categorisation of DF techniques may vary across research fields, and some studies may utilise unique DF techniques.

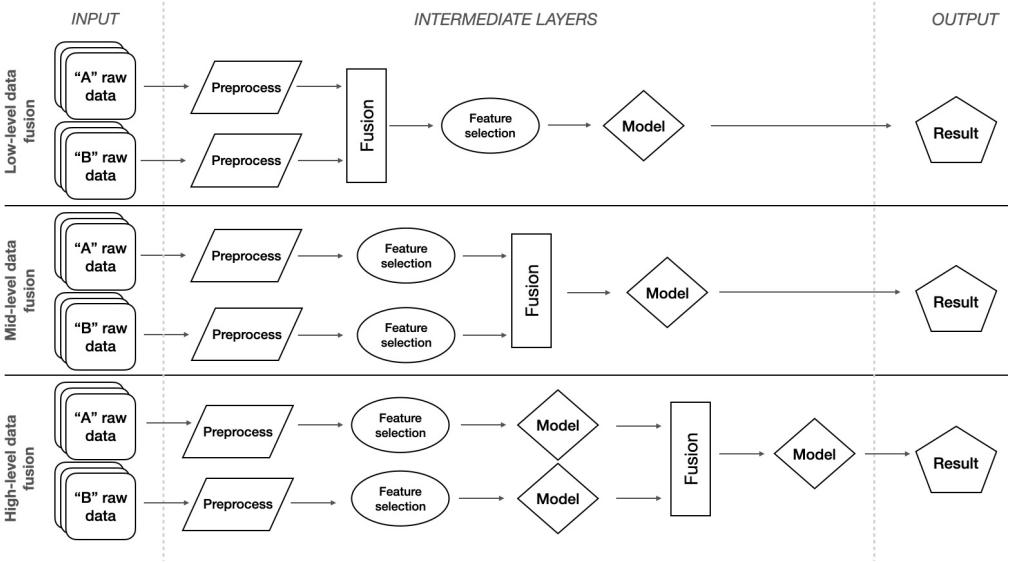


FIGURE 4.1: General schematic diagram of the three different techniques of data fusion [71].

The raw data sets "A" and "B" are preprocessed separately in each case. In the case of low-level DF, the fusion is the merging of the preprocessed data, followed by

feature selection and model building. In the case of medium-level, DF is performed after preprocessing and then model building. In the case of high-level, models are built separately after separate feature selection for the "A" and "B" datasets. Fusion is performed based on the results of the models, and then the second model is built on the fused dataset.

Previous chemometrics work has almost exclusively compared low, medium and high-level DF schemes for a single parameter of a type of sample set. In our case, we intentionally used two completely independent datasets to compare different data fusion techniques. No systematic study has evaluated a *complex-level ensemble fusion* method — in which feature-level concatenation, supervised latent-variable extraction, and model stacking are integrated into a single workflow — on paired MIR and Raman regressions containing fewer than one hundred samples. Moreover, compared to the single model, existing comparisons rarely quantify how much each additional fusion layer contributes to error reduction.

The DF has become a cornerstone of process analytical technology over the past two decades, enabling chemometric models to exploit complementary modalities such as MIR and Raman spectroscopy for rapid, non-destructive quality control. Despite this progress, systematic comparisons of DF levels — especially stacked, complex-level ensembles — remain scarce for small-sample industrial scenarios. In our investigation, we tested a range of DF techniques and analyzed their outcomes on two distinct case studies. We used a real-world industrial dataset and spectroscopic reference data from the publicly available benchmark (RRUFF) database, the RRUFF refers to the name of the project. This approach allowed us to compare the performance of machine learning models on both single-source and merged datasets, providing a critical evaluation of our methodologies.

The RRUFF Project is a joint scientific effort to build a complete, integrated, and publicly accessible database of high-quality scientific data for every known mineral on Earth [72]. Testing combinations of DF techniques and analysing the results were critical elements of our investigation, and both case studies gave special emphasis to these elements.

This study is the first to demonstrate that a *complex-level ensemble fusion* (CLF) strategy — where genetic-algorithm-guided variable selection on concatenated MIR and Raman spectra is followed by partial least-squares (PLS) projection and

final stacking with an eXtreme Gradient Boosting (XGBoost) regressor — can outperform all classical low-, mid- and high-level data-fusion schemes in real-world chemometric regressions comprising fewer than one hundred calibration samples.

The remainder of this chapter is organised as follows:

- The evaluation and comparison of data fusion techniques at different levels is done by processing spectroscopic data. The input data are MIR and Raman spectra, and the data fusion techniques are built on regression ML models in this study.
- Four different data fusion paths are presented in the study, as well as separate ML models built on the two spectroscopic data. A CLF data fusion technique was developed in the research, which outperforms the rest techniques. Here we introduce a CLF method that jointly selects variables from concatenated MIR + Raman spectra and stacks their latent variables into an XGBoost regressor (see Section 4.2).
- The developed methodology was applied to two case studies. One concerned the target variable influencing the quality of chemical additives, and the other concerned the result of the main mineral composition of a reference rock sample. In both case studies, the high and complex DF significantly improved the prediction accuracy of the regression models (see Section 4.3). Across two case studies CLF reduced validation RMSE and increased the R^2 relative to single-source models, outperforming classical low-, mid- and high-level DF.

4.2 Data fusion techniques to improve machine learning model

The purpose of this section is to give a coherent, self-contained description of the DF workflows used in this study, reproduce the individual processing steps, and clarify where this research adds value. We first explain why we used MIR and Raman spectra, then outline and summarise the evaluated five modeling paths (A–E).

From the perspective of developing chemometric soft sensors, better modeling results can be achieved by consciously expanding the measurement domain. MIR absorption probes polar functional groups, while Raman scattering is sensitive to non-polar or symmetrical bonds, and the fusion of the two domains provides more complete chemical information and consequently promises better prediction performance. The challenge lies in choosing an appropriate fusion level that maximises the information without reducing the performance and complexity of the model, especially under the small sample conditions typical of process analytical technology. Therefore, we compared five paths. Figure 4.2 illustrates the five model development techniques evaluated with or without employing various DF methods. The innovative aspect of the methodology presented in this study is the "E" approach, which generates a new feature set to construct the second ML model by utilizing the information garnered from the models and the predicted values at the first level. Compared to the three paths shown in Figure 4.1, there is no level at which a new set of features is created from the models. In this study, we only incorporated the model results from the feature selection following the fusion of the two datasets into this feature set. In other instances, information from different data sources can also be added to this feature set, such as temperature or pressure data in the case of online, inline soft sensors.

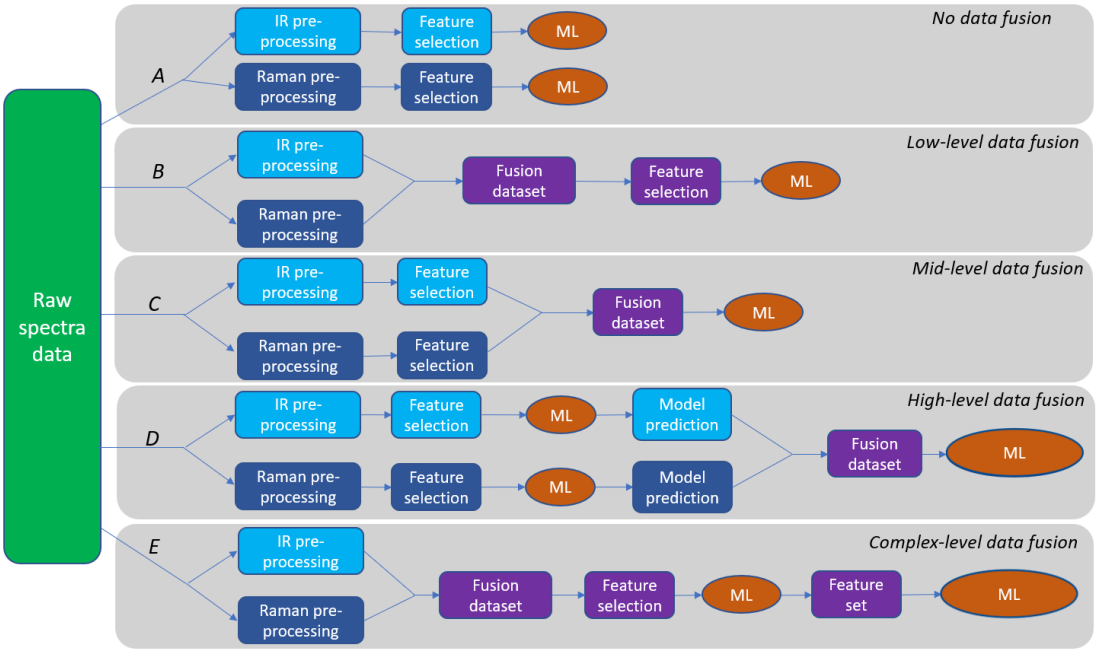


FIGURE 4.2: Architectural overview of data fusion methodologies (Paths B–E) compared to the baseline single-source approach (Path A).

No data fusion (**Path A**): When processing data from various measurement techniques, each type is handled independently since their raw results may show substantial variance. Techniques such as Raman, MIR, NIR, and X-ray necessitate distinct preprocessing approaches to interpret their measurement outputs. Due to differences in measurement methods, detector types, and the specific characteristics of the devices involved, these signals require individual preprocessing. Following preprocessing, feature selection is influenced by the target variable within each dataset. We use path "A" as a baseline to assess the effectiveness of different methodologies. The other four methods differ in terms of how extensively they implement DF methods, and path A offers an overarching view of the efficacy of DF.

Low-level data fusion (**Path B**): Here, raw datasets are consolidated and treated as one dataset at the low-level DF stage. The entire dataset undergoes feature selection, after which a regression ML model is constructed. This DF approach allows each dataset to undergo initial preprocessing (like standard scaling, baseline correction, centring, normalization), enabling the combined processing of signals on various scales. It is crucial to ensure proper preprocessing of features of each dataset and meticulous management of the resulting dataset before feature selection.

Mid-level data fusion (**Path C**): In the mid-level DF, key actual or latent features, extracted through feature selection from each dataset, are combined. Here, the primary task involves applying unsupervised learning to each separately preprocessed dataset to reduce crucial information. This approach is best used in cases where large datasets need to be processed together and signal processing needs to be done many times a day in a short time. The variable reduction helps to use only a reduced dataset after DF before modeling.

High-level data fusion (**Path D**): The high-level DF path builds separate models after separate preprocessing and provides their output and model components as input to the fused dataset. An important step is selecting and processing the main variables of the models. In cases where different models are used, the preprocessing of the model components (normalization, centring, etc.) is paramount. The second ML model can be considered a second layer that processes the information extracted from the first layer from different data sources.

Complex-level-ensemble data fusion (**Path E**): The CLF advanced DF method is built by preprocessing the different datasets and performing a joint feature selection, similar to low-level DF, which then serves as an input to the model at the first level. Later, using both the feature of the model set and first-level prediction outcomes, a unified dataset is developed, which can accept additional inputs from other sensor data, like temperature and pressure measures. A new ML model is then developed on this newly created database, corresponding to the second ML model shown in Figure 4.2.

The difference between path D and path E is that in path D, a model must be built separately for the two data sources in the first step, followed by data fusion, and then a model must be built on the output results in the second step. In path E, data fusion occurs at the level of preprocessed data sets from the two data sources in the first step, and then a unified model (similar to the low-level) is built in the second step.

It is essential to emphasize that in our analysis (A-E), each set of input data was independently preprocessed, and raw data was not utilized in any instance. The "ML" in the figure can be both supervised and unsupervised learning techniques. In our case studies, the first ML was always a Partial Least Squares (PLS) regression model, before which a genetic algorithm (GA) performed the feature selection step. The second ML model was an XGBoost regression model. The GA was used exclusively for feature selection in the PLS models. The GA is utilised solely for feature selection and is performed strictly as an offline process. Crucially, it is not employed during regular prediction or subsequent model retraining phases, which ensures that it does not interfere with the real-time operation of this framework. This is considered an offline step that can be repeated as needed before the models are retrained. For instance, the execution time for the 41 rock samples was only 25 seconds using the MATLAB 2020a PLS Toolbox. Given that this retraining phase for soft sensors is expected to be a relatively infrequent event—occurring approximately once a month, depending on the material—the computational cost of the GA is not a significant concern.

In our research, we tested different DF approaches on MIR and Raman spectroscopy spectra of oil industry samples to create models with sufficient accuracy for difficult-to-predict quality parameters. We present two case studies, one estimating the hydrocarbon/imide ratio of oil industry additives, and the other estimating

the quartz content of the RRUFF dataset. In both cases, the input was the MIR and Raman spectra.

Infrared (IR) spectroscopy analyzes molecular vibrations upon IR light absorption and subsequent photon release, detected as an absorbance versus wavenumber spectrum, useful for identifying functional groups (MIR region) and specific compounds (fingerprint region) [56]. Raman spectroscopy, conversely, uses inelastic light scattering from a laser to analyze shifted photon energies, revealing molecular vibrations and unique structural fingerprints for material identification [73]. Raw spectroscopic data often requires preprocessing, linking it to calibrations or mathematical techniques for meaningful interpretation and modeling in machine learning. Common IR preprocessing includes scatter correction and spectral derivatives, with baseline correction and scatter correlation being relevant for the MIR range [74]. This study employs signal detrending for baseline correction and Standard Normal Variate (SNV) for scatter correction on MIR spectra. SNV centers each spectrum by subtracting its mean and then scales it by its standard deviation, applied individually to each spectrum [75]. For MIR data, signal detrending was used as an additional pretreatment alongside SNV centring. For Raman spectra, the Automatic Whittaker Filter, similar to Weighted Least Squares (WLS), automatically removes baseline offsets using a piecewise approach. Key Whittaker filter settings include Λ, controlling baseline curvature (smaller allows more curvature), and P, governing the allowed asymmetry (larger permits more negative regions) [76]. In this Raman data, Λ was 10000 and P was 0.001. This method, performed via the WLS function with the Whittaker filter option, was followed by normalization and centring of the preprocessed Raman spectra.

Partial Least Squares (PLS) regression models the relationship between $\mathbf{X}$ and $\mathbf{Y}$ data blocks using score matrices. The PLS model consists of outer relations for each block ($\mathbf{X} = \mathbf{S}_X\mathbf{L}'_X + \mathbf{E}_X$ and $\mathbf{Y} = \mathbf{S}_Y\mathbf{L}'_Y + \mathbf{E}_Y$) and an inner relation linking their scores. The objective is to minimize the $\mathbf{Y}$ block's residual ($\mathbf{E}_Y$) while finding a useful connection between $\mathbf{X}$ and $\mathbf{Y}$ scores. This inner relation is conceptually based on the relationship between $\mathbf{S}_Y$ and $\mathbf{S}_X$ [77]. Boosting algorithms, including gradient boosting, improve prediction by combining weak learners. XGBoost, a gradient-boosting implementation, was tested with different parameters and 10-fold cross-validation [1].

Appropriate feature selection is crucial in machine learning modeling for multivariate data, as not all variables contribute equally, and excluding some can improve

prediction. In MIR and Raman spectroscopy, GA-s can identify the most relevant spectral sections by treating wavenumbers as potential features [78]. GA is an optimization method inspired by natural selection, managing a population of potential solutions (individuals). The GA iteratively evolves this population through initialization, evaluation (assigning fitness), selection of the fittest for reproduction, crossbreeding to create offspring, and mutation to ensure diversity. For spectral data, a GA individual is a binary vector where '1' indicates wavenumber inclusion. The population is a set of these binary vectors, and the fitness function evaluates each individual's predictive power (e.g., accuracy, R^2, $RMSE$). Selection favors individuals with higher fitness, and crossover combines traits of two parents to create offspring. Mutation introduces random changes, and a new generation is formed from the selected, crossed, and mutated individuals, with the process repeating until a satisfactory solution emerges [79].

We present two distinct case studies: one focusing on estimating the hydrocarbon/imide ratio in industry additives, and the other on determining the quartz content from the RRUFF database. In both scenarios, MIR and Raman spectra served as the primary input data. For our modeling, "ML" refers to PLS regression models, while second "ML" denotes XGBoost regression models, with GA employed for feature selection. Our research rigorously investigated various DF approaches applied to MIR and Raman spectroscopic data from industrial oil samples. This was done to develop models with sufficient accuracy for challenging-to-predict quality parameters.

4.3 Results and case studies

This section presents a detailed empirical validation of the proposed DF methodologies through two distinct case studies. These investigations were carefully selected to showcase the versatility and robustness of our novel DF paths across varied analytical challenges.

The first study involved an industrial dataset, where models were developed to predict the crucial hydrocarbon/imide ratio of additives in the oil industry. Concurrently, a second study utilized a benchmark rock dataset to estimate quartz content, a key geological parameter. In both instances, single models were initially built using only MIR and Raman spectroscopic datasets to establish a single model result.

These rigorous comparisons underscore the significant potential of our DF framework to enhance predictive accuracy and operational insights in complex industrial and scientific environments. The results collectively demonstrate the robustness of our approach, providing a clear pathway for its broader application.

4.3.1 Data fusion applied to the case of prediction of quality parameters of additives

Our goal is to accurately predict the quality of additives. For this, we need fast, easy-to-use, and accurate forecasting. We preferred fast, non-destructive techniques based on spectral data, but if only MIR or Raman techniques were used, the accuracy of the models needed to be improved. Therefore, the quantitative prediction must be improved, which is done using DF in this study. In our case, this is the hydrocarbon/imide ratio of the selected additive. During the production of additives, classification is based on chemical and physical parameters. Quality parameters are difficult to predict, so we have to use DF techniques. In the future, our goal is to integrate inline industrial Internet of Things (IIoT) sensors into our processes in addition to laboratory measurements, which can be used to predict the quality of raw materials, intermediate products and final products online in real time, as well as the different effects of quality on production [41]. Our chosen analytical measurements are spectroscopic measurements, including MIR and Raman spectroscopy. Both measurement techniques have industrial sensors available

on the market. During Raman inelastic light scattering, the monochromatic excitation laser beam hits the sample material, and the scattered light provides information about molecular vibrations and chemical structure. Raman detection detects vibrations along the covalent bonds, the advantage of which is that the possible water content of the sample does not interfere with the measurement, in contrast to IR spectroscopy. IR is based on the molecular absorption of irradiated IR light caused by vibrational and rotational transitions in covalent bonds. The IR range is the spectral section between 12800 and 10 cm^{-1}, within which we distinguish three different regions, and the MIR range (MIR: 4000-650 cm^{-1}) was used in this study. At the same time, IR detects the vibration of polar covalent bonds, while Raman detection detects the vibration of nonpolar covalent bonds. The advantage of IR spectroscopy is that it is not disturbed by fluorescence compared to the Raman technique [80]. Our target variable is the hydrocarbon/imide ratio, an essential parameter when qualifying the additive sample. Our goal was to build an accurate model for this parameter when building the model, so we compared the five paths based on these.

The number of additive samples was 99, whose MIR and Raman spectra were measured. The dependent variable $\mathbf{X}$ of the spectrum of the samples was preprocessed with Autoscale and Mean center settings in the PLS toolbox of MATLAB 2020a. The results were summarized based on the methodology presented in this section (4.3), and the models were compared based on $RMSE$ and Pearson R^2 key performance indicators. The models were checked using both calibrations, where all samples were used for model development and the Venetian blinds cross-validation techniques. In the case of the Venetian blinds cross-validation, each test set is determined by selecting every s-th object in the data set, starting at objects numbered 1 through s, which in our case was s=10 (Figure 4.3 and 4.4 and Table 4.2). The mutation rate is not a fixed value; rather, the toolbox adds a random number drawn from a Gaussian distribution to each individual. This mutation is controlled by the "Scale" and "Shrink" parameters, which have a default value of 1. The crossover rate was set to the default of 0.8, and the number of generations was 100 times the number of variables.

The results show that paths B, D and E gave better results than path A. Among the five methods, E gives the best results. Venetian blinds cross-validation reduces the model error ($RMSE$) by more than half, doubling the square of Pearson correlation (R^2). In path E, we processed the input spectra separately for MIR

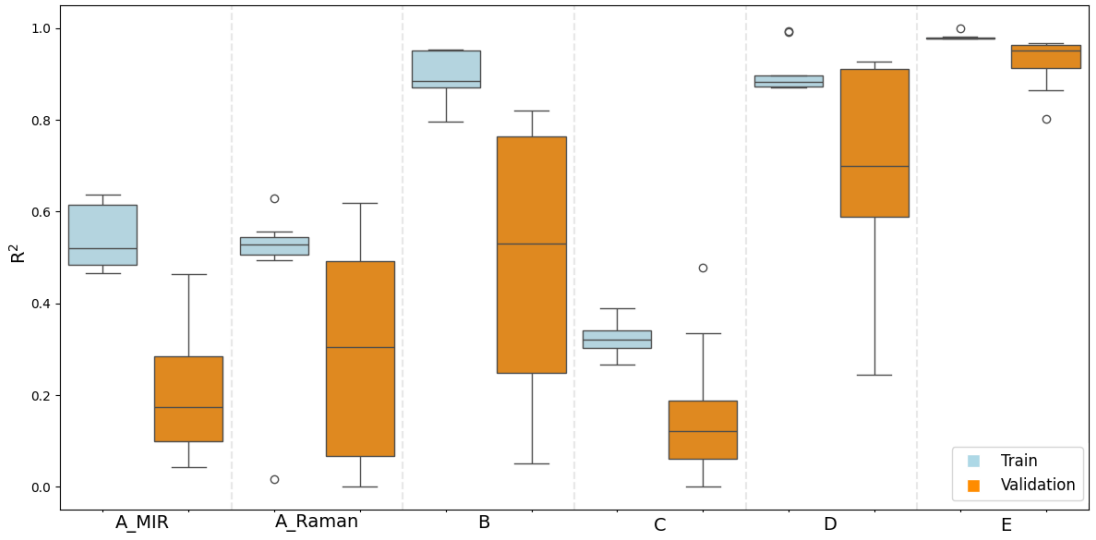


FIGURE 4.3: Additives dataset models R^2 results of train and validation.

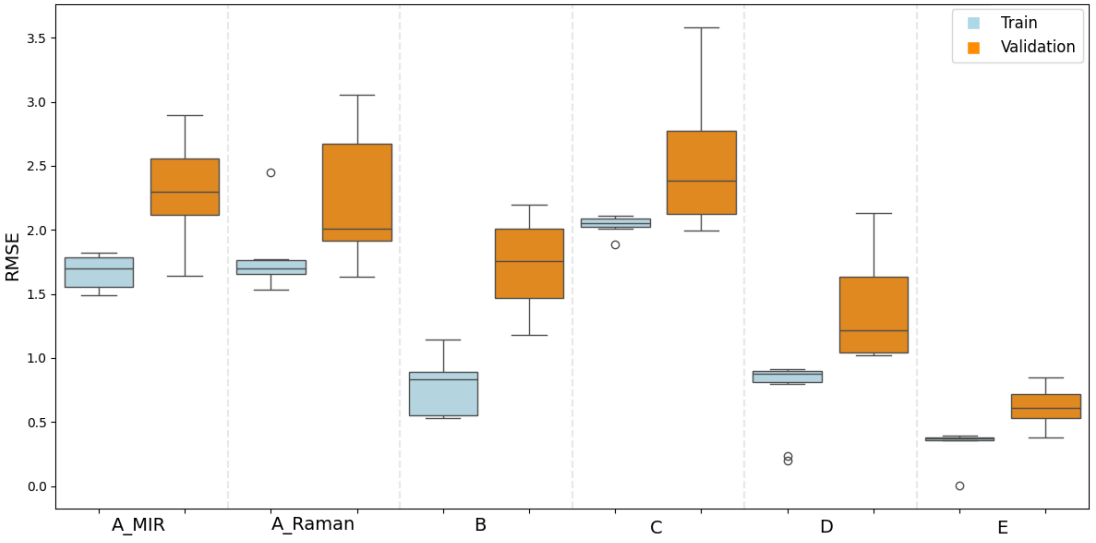


FIGURE 4.4: Additives dataset models $RMSE$ results of train and validation.

and separately for Raman, then pooled the entire spectral data set and ran PLS regression. After the PLS model, we built a second ML model, which was an XGBoost regression model. The latent variables of the first model served as input, which were selected by the GA. This technique automatically selects the variables that lead to models with lower values of $RMSE$. GA runs were performed with a window width of 20 and a population size of 64.5 replicate runs. Each time, the initially included variables are randomly chosen, so the results can vary. In path C, feature selection was based on PCA. The spectra were represented by the values of the first ten principal components, which explain more than 90% of the total variance for both Raman and MIR. The ML model refers to PLS models

Table 4.2: Summary table of model results for different DF paths in the additive dataset.

	A_MIR	**A_Raman**	**B**	**C**	**D**	**E**
R^2_{train}	0.544 ±0.065	0.486 ±0.160	0.892 ±0.057	0.324 ±0.035	0.904 ±0.046	0.981 ±0.007
R^2_{val}	0.206 ±0.137	0.295 ±0.236	0.494 ±0.278	0.160 ±0.140	0.694 ±0.221	0.928 ±0.052
$RMSE_{train}$	1.673 ±0.125	1.764 ±0.238	0.787 ±0.222	2.043 ±0.062	0.746 ±0.267	0.335 ±0.112
$RMSE_{val}$	2.313 ±0.380	2.232 ±0.459	1.727 ±0.352	2.504 ±0.461	1.369 ±0.372	0.623 ±0.137

combined with GA. The weakest results were given by path C, with medium-level DF. The results were even lower than in path A, where we did not use DF. Mid-level DF (path C) performs feature selection on MIR and Raman data sets. After selecting the function, we compiled the reduced data set produced by the PCA analysis of the MIR and Raman data sets separately. We then built a PLS regression model for the 10-10 main principal components. Paths D and E in the ensemble learning technique have different versions depending on the method used for learning; however, in this study, we focused in detail on the boosting technique. We selected the XGBoost technique for this investigation because the algorithm can dynamically determine the depth of the decision trees used as weak learners, adding penalty parameters to prevent overly deep trees, which can mitigate the risk of overfitting and enhance the performance of the model [81]. The results of paths B, C, D, and E prove that selecting variables is essential when using DF techniques.

Path E gives the best result, and the standard deviation of the predicted results is also the smallest on the path E.

The findings indicate that models developed using the hydrocarbon/imide ratio and the low-, high- and complex DF techniques outperformed those relying solely on MIR or Raman data single DF. The results also show that the intermediate technique result was worse than the original one when we did not use the DF technique. In all scenarios (B, C, D, and E), choosing variables is crucial and can be managed manually or automatically. On the paths A, B, D, and E, in our case, we employed a Genetic Algorithm (GA) for feature selection; in the case of path C, PCA was used. Among these, path E, utilizing the CLF approach,

provided the most favourable results. For the CLF (E), the Pearson correlation coefficient (R^2_{train}) attained 0.9805, with a training $RMSE_{train}$ of 0.3354, while for validation (via Venetian blinds cross-validation), R^2_{val} was 0.9279 and $RMSE_{val}$ was 0.6229. In contrast, the mid-level DF applied to path C did not outperform separate models for each dataset. The first ten principal components account for over 70% of the total variance for both Raman and MIR in the rock data set. This weaker performance in path C is attributed to the necessity of jointly managing feature selection when employing DF techniques.

Feature selection is crucial for spectrum in the case of the DF application, and this technique was applied across various paths (A & D, B & E and C) on the industrial dataset. The provided spectrum represents the MIR and Raman data of a single sample (Figures 4.5 and 4.6). Green bands indicate spectral regions selected by Genetic Algorithms (GA) for path A, while blue bands show those selected by GA for path B. Red bands highlight spectral regions with loading values exceeding + or – 0.05 for the top 10 principal components in path C (as shown in Figures 4.5 and 4.6). These figures also depict the feature selection outcomes for paths D and E, where features A correspond to the chosen spectral details of path D, and features B correspond to those of path E.

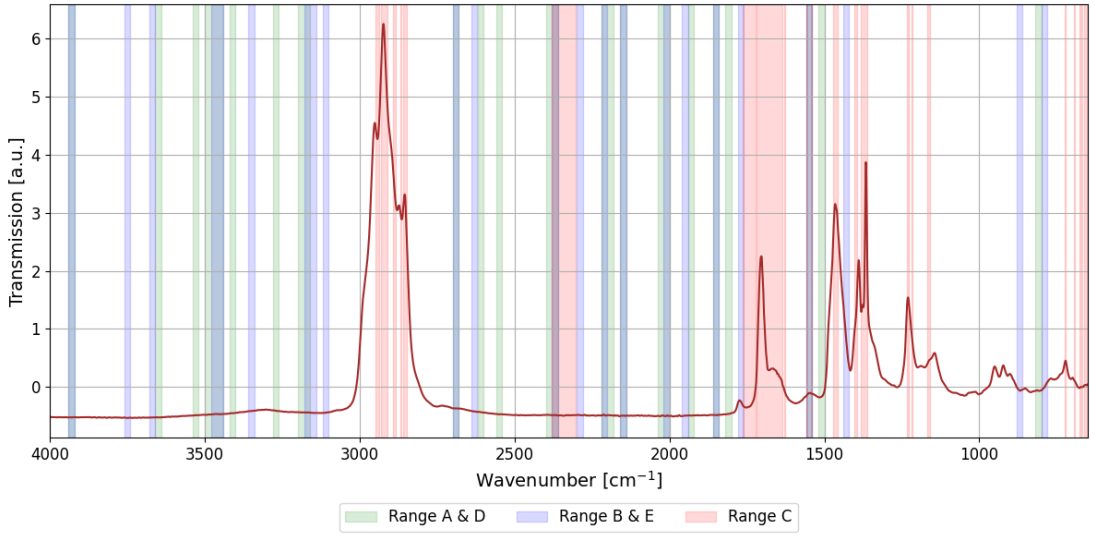


FIGURE 4.5: Visualization of the feature selection in A & D, B & E and C methods in case of MIR spectra. The spectrum example from the industrial dataset is a spectrum of one sample.

The additive dataset exhibited a scenario where CLF significantly enhanced prediction performance, particularly when individual spectral modalities (either MIR or Raman) showed inherently weak predictive capabilities on their own. In this

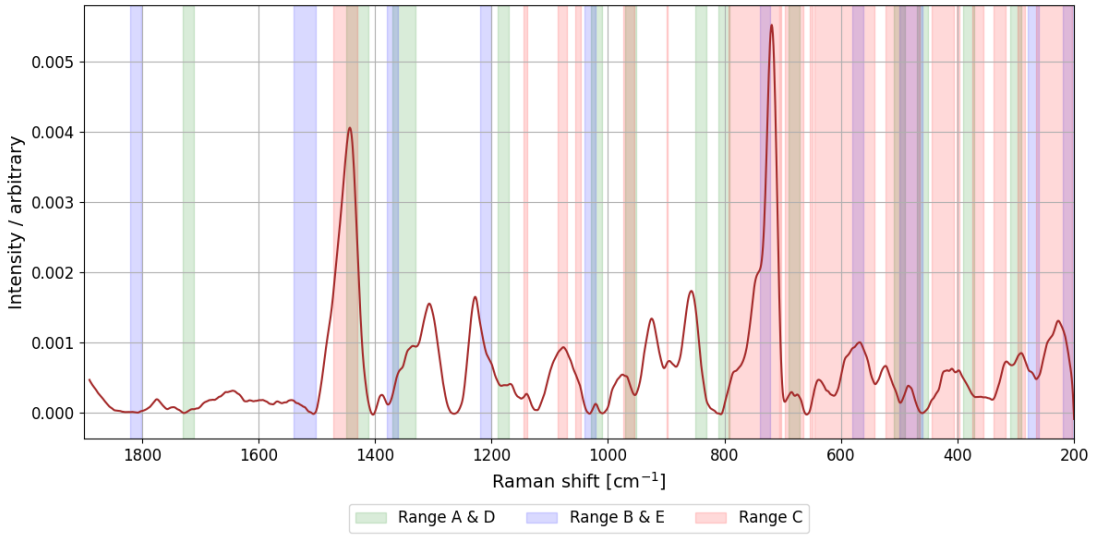


FIGURE 4.6: Visualization of the feature selection in A & D, B & E and C methods in case of Raman spectra. The spectrum example from the industrial dataset is a spectrum of one sample.

case, CLF effectively augmented the weak modalities by successfully integrating the additional information provided by the stronger counterpart.

4.3.2 Data fusion applied to the case of rock dataset and feature selection

The second data set contains the international benchmark collected in an extensive data set at the RRUFF project. The rock database is a digital library of detailed information about rock and mineral samples, with a focus on their spectroscopic properties. The RRUFF project is one of the largest and most comprehensive mineral research studies ever undertaken [82]. This data set is available online and contains many different minerals and in this study, reference data from many different minerals are collected, and MIR, Raman, and X-ray measurements are performed. In the second case study, a dataset of 41 rock samples was used to evaluate the performance of five distinct DF methods. The results from both the training and validation phases of this evaluation are presented in Figure 4.7 and 4.8 and Table 4.3, corresponding to the methodologies outlined in Figure 4.2.

The results show that the models based on the quartz content of the studied rocks gave better results with high- (D) and CLF (E) techniques than the models built on only MIR or Raman data. When applying the methodology, variable selection

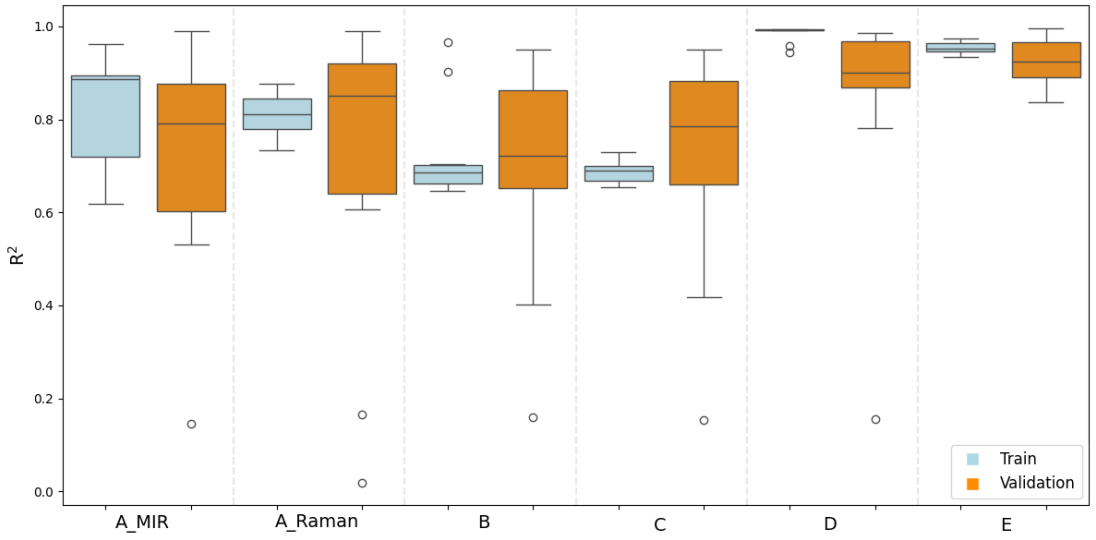


FIGURE 4.7: Rocks dataset models R^2 results of train and validation.

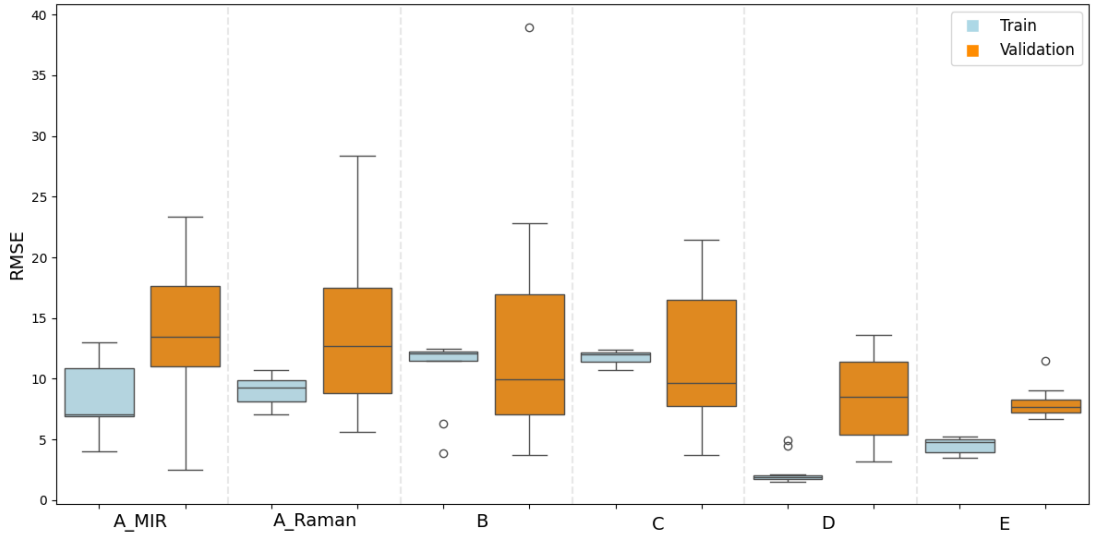


FIGURE 4.8: Rocks dataset models $RMSE$ results of train and validation.

is essential in all paths B, C, D and E. This selection can be performed manually or automatically, and in our use case, we select the features using the GA. Considering the different model results, the results of the paths D and E were good, and of these two, path E gives the best result. Using CLF (E), Pearson R^2_{train}= 0.9544, $RMSE_{train}$= 4.516 from the train set, Pearson R^2_{val}= 0.9215, $RMSE_{val}$= 8.019 from the validation set. The DF techniques B and C performed similarly to the single paths. In case C, the PLS model selected only the MIR data from the 10-10 MIR and Raman spectra. It is also important to note that the rock dataset was selected from the aforementioned benchmark dataset, so that each sample had both MIR and Raman spectra and did not form separate groups. Nevertheless,

Table 4.3: Summary table of model results for different DF paths in the rock dataset.

	A_MIR	**A_Raman**	**B**	**C**	**D**	**E**
R^2_{train}	0.825 ±0.123	0.811 ±0.046	0.727 ±0.106	0.686 ±0.023	0.985 ±0,017	0.954 ±0.012
R^2_{val}	0.720 ±0.238	0.704 ±0.327	0.691 ±0.235	0.709 ±0.238	0.838 ±0,236	0.922 ±0.053
$RMSE_{train}$	8.267 ±2.844	9.071 ±1.127	10.665 ±2.867	11.756 ±0.547	2.400 ±1.186	4.516 ±0.600
$RMSE_{val}$	13.808 ±5.413	14.700 ±7.375	14.042 ±9.988	11.666 ±5.439	8.438 ±3.496	8.019 ±1.325

the 10-fold variance of the prediction results is large compared to the first dataset.

In the case of the rock dataset, the path E proved to be the best, considering the R^2 and $RMSE$ results. The path E results are also good. Figure 4.7 clearly shows that the benchmark RRUFF dataset data is inhomogeneous and was selected from several rock types.

The feature selection is an essential step in spectrum DF, and this technique for different paths (A & D, B & E and C) is present on the RRUFF dataset. The spectrum depicted is the MIR and Raman spectrum of a well-known rock, calcite. The green ranges show the spectrum ranges selected with GA during path A, and the blue ranges show the spectrum ranges selected with GA during path B. The red range shows the spectrum ranges with a loading value greater than + or - 0.05 for the top 10 principal components during path C. The figures 4.9 and 4.10, as well as appendix A.2, A.3 show that no data from the Raman spectra were used for the path C, i.e. it was not necessary to use Raman data because they did not provide more information regarding the target variable.

Feature selection clearly shows which path selected which spectral details. This type of representation is impossible for paths D and E, as they process the results of PLS models.

In the rock database, CLF showed a lower contribution compared to the additive dataset. The MIR spectrum alone already provided strong predictive power, while the information content of the Raman spectrum was less relevant for predicting the target property. Consequently, CLF only provided a modest improvement. This result suggests that the integration benefit was limited, as one modality was

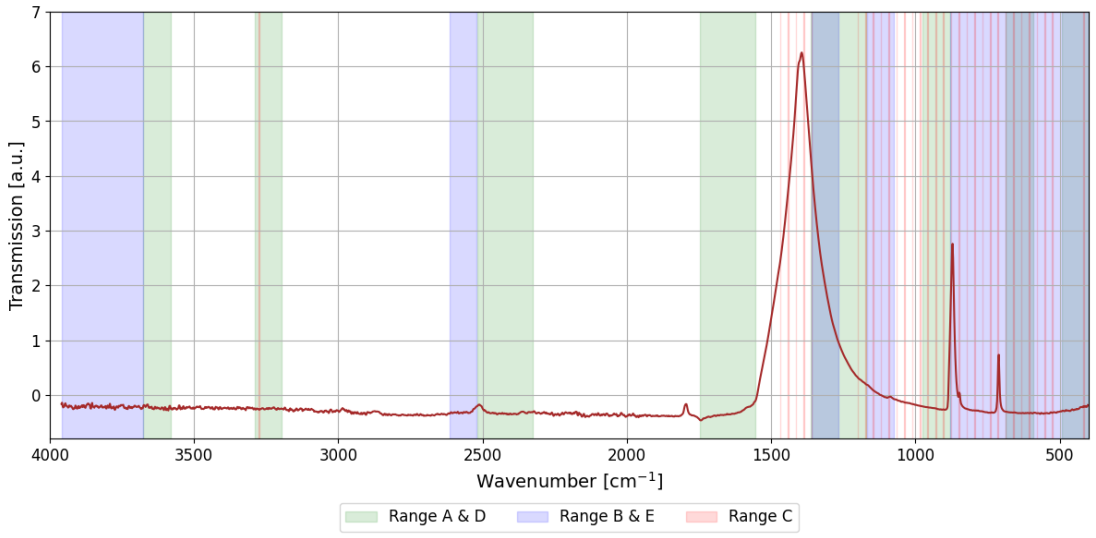


FIGURE 4.9: Visualization of the feature selection in A & D, B & E and C methods in case of MIR spectra. The spectrum example from the RRUFF dataset is a spectrum of calcite.

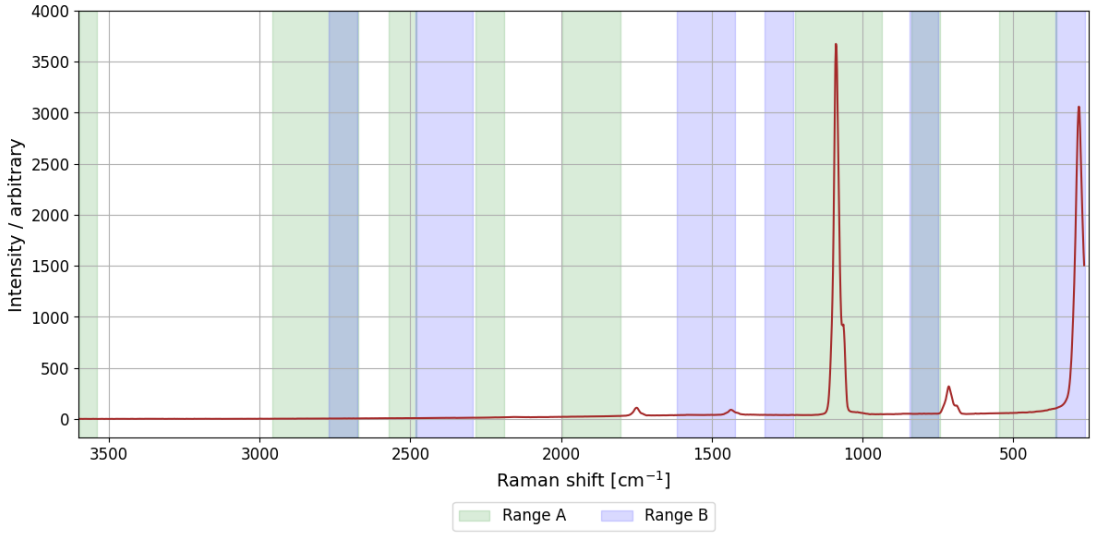


FIGURE 4.10: Visualization of the feature selection in A & D, B & E and C methods in case of Raman spectra. The spectrum example from the RRUFF dataset is a spectrum of calcite. C range was not significant in the case of the Raman spectrum, it can be seen in the accuracy of the C model that the accuracy of the MIR model.

already strongly correlated with the target property, while the second modality provided minimal additional predictive information.

4.4 Chapter summary

In this research, various frameworks of Data Fusion (DF) techniques built upon analytical chemical measurement spectra were rigorously tested. The accuracy of these DF models was subsequently benchmarked against the performance of single, un-fused models. The results encompassed comparisons across low-level, mid-level, and high-level DF methodologies, alongside the novel *complex-level ensemble fusion* (CLF) technique introduced in this study.

We have shown that fusing complementary MIR and Raman spectra through a CLF workflow with GA feature selection on concatenated spectra, PLS projection, and XGBoost stacking consistently outperforms single-source models as well as the classical low-, mid- and high-level data-fusion schemes. On the additives data set (99 samples) CLF cut the cross-validated $RMSE$ by $\approx 72\%$ from 2.232 to 0.6229 and raised R^2 from 0.295 to 0.928, while on the RRUFF minerals (41 samples) CLF reduced the cross-validated $RMSE$ by $\approx 42\%$ from 13.808 to 8.019 and increased the R^2 from 0.720 to 0.922. These gains were achieved without enlarging the calibration set. This demonstrates that an information-efficient, stacked framework can recover synergistic spectral cues even under the small-sample conditions typical of PAT.

The degree of improvement achieved by CLF varied across the two case studies due to the relative predictive strength of each modality. In the industrial additives dataset, standalone models based on MIR or Raman spectra showed limited predictive accuracy, and CLF was able to leverage complementary features to significantly enhance performance. In contrast, the rock dataset (RRUFF) featured MIR spectra that already provided strong predictive power for quartz content, while Raman spectra contributed relatively little additional information. Consequently, CLF led to more modest improvements. These results suggest that CLF offers the greatest benefit in cases where (i) individual modalities are weak or noisy, and (ii) the modalities carry distinct, complementary information relevant to the prediction task. When one modality already dominates in predictive relevance, the relative gain from fusion may be smaller.

Beyond the quantitative improvements, CLF offers practical advantages: (i) it preserves the standard chemometric tool-chain, requiring only widely available GA and boosting libraries; (ii) its modular design allows the easy addition of further in situ sensors (e.g., NIR, temperature, pressure) at the stacking stage;

and (iii) the workflow is fully reproducible through a single-pass cross-validation protocol.

The study also reveals two caveats. First, mid-level fusion delivered no benefit, highlighting that unsupervised dimensionality reduction may discard task-relevant variance when highly heterogeneous spectra. Second, the magnitude of CLF's improvement depends on rigorous preprocessing and feature selection; inadequate tuning can mask the fusion benefit, as seen for the mid-level path.

In conclusion, it can be stated that data fusion technology has significant advantages in the performance of ML models built in industrial digitisation. In the two case studies we have presented, the two measurement techniques (MIR and Raman) complement each other. The advantage of DF is that it can improve the accuracy of inadequate models, but in any case, appropriate preprocessing and feature selection are essential. The disadvantage of using DF is the proper application of the two mentioned elements, which is a significant challenge. Our further experience is that DF techniques are worth considering in cases where the accuracy of individual models is low, and it is necessary to choose which measurement techniques to apply DF to carefully. As a further development opportunity, the goal is to develop an industrial sensor capable of predicting quality online in real-time. Future work will therefore focus on (a) extending the approach to classification tasks and to online, real-time soft sensors, and (b) validating the methodology on larger, multi-site industrial data sets to establish robustness across instruments and operators.

Together, these results establish CLF as a transferable, low-overhead route to markedly more accurate chemometric models and pave the way for its deployment in next-generation quality-control and geochemical screening workflows.

Chapter 5

Generating realistic infrared spectra using artificial neural networks

5.1 Introduction

Regrettably, there is often a scarcity of infrared spectra available for machine learning purposes in the case of authentic samples, the distribution of the input is not appropriate. In order to overcome its deficiencies, it is necessary to simulate spectra that complement these deficiencies. Suitable sampling is expensive, the sample is costly and there is little time available. The production of artificial data is obvious by simulation.

Based on the review of the literature on the topic of FTIR spectrum generation, the most common case is when the quantified uncertainties of measurements for individual wavenumbers are used to generate samples. Small intervals were used by Sales *et al.* when setting different intervals resulted in noisier spectra. The corresponding interval sizes were 24, 250, and 134 nanometers for MicoNIR^{TM}, and Fourier transform mid-infrared (FT-MIR), respectively. Similar virtual spectra can be produced to IR measured spectra with this methodology [83]. Some studies use linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), and soft independent modeling of class analogy (SIMCA) methodology, as well as simulated near-infrared spectra, for the cultivation medium for soft tissue

cells. Szabó *et al.* divided the spectra of the samples into different groups and created artificial spectra by taking into account the average spectrum of each group of samples and the variance of the wavelengths of the spectra. The approach is suitable for quality control and optimization of heat-treated medium powders for cells [84]. Scopus is a comprehensive, multidisciplinary, and trusted abstract and citation database that was used in two searches. In the first case, the keywords of the articles were - resampling, Monte Carlo simulation, bootstrap, simulated spectra, and analytical chemistry. The grouping of 198 results is shown in the network diagram (Figure: 5.1). The different areas of the topic appear in different colors on the network clearly showing that the relevant articles are divided into four groups. The red and yellow groups are the most obvious for us, as they include spectroscopy, spectrum generation, bootstrap, simulated spectra, and Monte Carlo simulation.

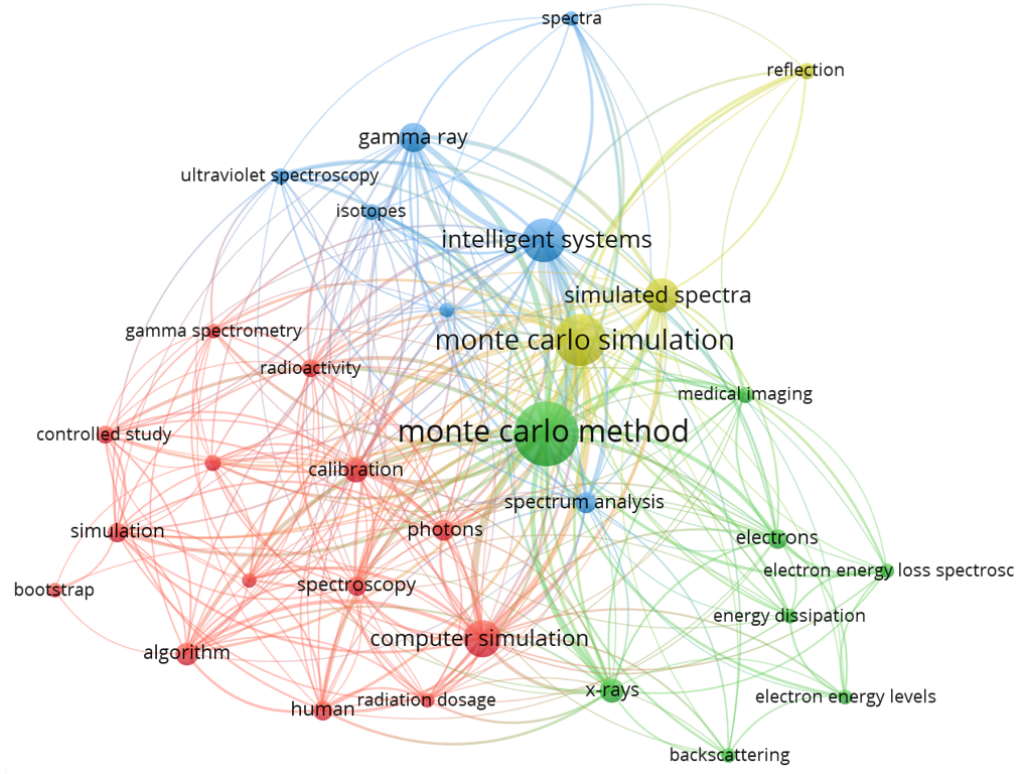


FIGURE 5.1: Network diagram of keywords *resampling*, *Monte Carlo simulation*, *bootstrap*, *simulated spectra*, and *analytical chemistry* on Scopus.

There are several options for generating artificial spectra, highlighting a few of them: algebraic method, bootstrap method, and Monte Carlo simulation. The algebraic method is the classical way and is based on the rules for mathematical error propagation. The error of the curve fitting method can be described

in algebraic form, and the error propagation rule is applied to each step of the relationship. The method assumes that the noise is normally distributed and of constant variance. These assumptions constitute the weakness of this method. The bootstrap method generates artificial data by taking random samples from the original data set and resampling with return . Some of the unselected samples may not be taken into account in the calculation. The deficiency of the Monte Carlo simulation is that it cannot generate a sample whose properties (reference value) are not known from the original dataset. One must know the variance of the random noise present in the data to apply a Monte Carlo simulation. The procedure involves computing the slope and intercept of each simulated data set containing random noise, performing this step multiple times with different noise sets [85]. Using Monte Carlo simulation, real gamma-ray spectra can be generated. Kwon *et al.* have demonstrated that the spectra generated by the model match well with real experimental data, under certain constraints [86]. The Monte Carlo method can be applied to the synthesis of early-time supernova spectra. Mazalli *et al.* reviewed the characteristics of supernovae, and the importance of spectrum synthesis, than they compared the simulation data with real data [87]. Furthermore, the Monte Carlo simulations can be used to generate the diffuse reflectance spectrum of the skin at different locations on the model and use the data to train artificial neural networks. The trained networks were used to predict the diffuse reflectance spectrum of the skin, which gave very accurate results. Tsui *et al.* suggest that this new method enables a quick and accurate prediction of the diffuse reflectance spectrum of the skin, which is important for understanding and diagnosing the health and condition of the skin [88].

In the second case the search keywords were "simulated" or "artificial", "spectra", "chemometrics", and "Fourier". The number of relevant articles was 121 (02.10.2023). Most of the most of the articles are more related to quantum chemistry and semi-empirical approaches and do not related closely to our practice, but we briefly summarized some interesting literature. Beć and Huck deal with the near-infrared (NIR) analysis of the components of the medical *Thymi herbarium* from a quantum mechanical perspective. The experimental part compares the NIR spectra of *T. herbarium* in solid and molten states and at different concentrations of CCl_4 solutions. The analysis reveals spectral regions with very different sensitivities than the sample state: 6000–5600 cm^{-1} and 4490–4000 cm^{-1}, these spectral details are also paramount in the PLS regression model, as they provide the bulk of the models [89].

We used various machine learning methods and compared their efficacy in predicting the rock solubilities in HCl, HF/HCl, and acetic acid (AcOH) [1] based on their infrared (IR) spectra. We had 888 infrared spectra and solubility data. The backpropagation artificial neural network (ANN) was one of the best-performing algorithm. The good performance can be attributed to the ANN's capability to establish a nonlinear relationship between the various peak intensities and the solubilities of the samples. The process consisted of the usual pre-processing steps such as standard normal variate (SNV), baseline correction, feature selection, after the pre-processing are built model with network optimization. Subjectivity was involved in the selection of the features, which brought ambiguities and distortions. Convolutional neural networks (CNN) are capable of identifying the most important aspects of two-dimensional data sets, which makes them well-suited to automatic feature selection without human intervention. Although the number of infrared spectra and the associated solubility data has increased recently, we still need a considerable number of samples exploit capabilities of CNNs. A CNN consisting of three convolutional and two neural layers has delivered a performance comparable to our ANN in ref [1]. without human intervention in feature selection ANN in the study [1]. The Partial Least Squares Regression (PLSR), eXtreme Gradient Boosting Regression (XGBR) and ANN regression performance indicator of the models are summarised in the table. Appendix A.4. Furthermore, the histogram of the ANN prediction values for the three target variables can be seen in the Appendix A.5. The number of samples are shown by the lengths of the bars and the red marks show the standard error of the prediction. The standard error of the forecast is high, if the sample number is small, this can be concluded from the Appendix A.5. Therefore, in this study we focused on reducing these errors and improving the distributions by producing targeted samples, taking into account the Kullback-Leibler and Jensen-Shannon deviations [90]. Another bottleneck of CNN models is the small number of training samples, which is supported by some literature [91, 92]. Comparing the performance of PLS and ANN models with CNN on the literature that if we have a large enough data set, the performance of CNN can be better [93, 94, 95]. In other words, it is necessary to generate as many artificial spectra as possible, which are as similar as possible to the real ones. Taking the example of sample multiplication by rotation or flipping over the pictures, we have decided to make artificial infrared spectra. Giving randomness to the spectra is insufficient because the spectral information and the solubilities of the samples correlate. We aimed to create artificial infrared spectra that correlate

with the associated solubility data as closely as the real infrared spectra do with the measured solubility data. The artificial spectra that we have prepared are well-supported by the empirical results of the specific physical and chemical properties of the samples. The main challenge was to generate appropriate spectra where real spectra were poorly or incompletely sampled. We elaborated a procedure to generate spectra that well represent the special data structure.

5.2 Materials and methods

The intention is to maintain the relationship between the spectra and the physical and chemical properties of the substances when simulating infrared spectra. Unlike the methods where new spectra are generated with the real spectrum and its noise. Artificial spectra are added to the existing collection of spectra, improving the areal distribution in the solubility parameters' space. The final goal of creating the artificial spectra is to apply them along with the infrared spectra taken in real-world samples in data-intensive applications, We want to predict the solubility of the samples based on the infrared spectrum, even if there is not enough data in the given range, supporting the acid job for the Exploration and Production (E&P).

5.2.1 Materials and instrumentation

Acid stimulation is used to restore wells' productivity, in the oil industry, during production on oil fields. Formation damage causes the reduction of fluid inflow into the wells, which can be eliminated by acid stimulation. Oil industry E&P professionals investigate the solubilities of rock samples to prepare and design stimulation, called acid job [96]. Based on the well-site geologist's decisions, the laboratory determines the solubility of a rock sample in either HCl and HF/HCl or HCl and acetic acid (AcOH). Two groups were formed from the data, and separate tests were performed in these two clusters. Different acid treatments are used to stimulate the wells [96], which are tested in laboratory conditions by the solubility of the rock samples to be tested in 15% hydrochloric acid (S1) and a mixture of acids containing 3% hydrogen fluoride and 12% hydrochloric acid, or if the sample is mainly carbonate, it is dissolved in 10% acetic acid (S2). Although the solubility in the American Society for Testing and Materials (ASTM) standard is a lengthy process and requires a lot of traditional laboratory work, it has a significant effect

in the quantitative and qualitative determination of acid and other parameters suitable for a specific acid job. Acids, too strong can damage the rock structure. Acids, too weak do not deliver the desired results. If an inappropriate acid is used, it may not achieve the desired stimulation effect in the surroundings of the well, leading to insufficient production increase. Moreover, the use of the wrong acid may also cause formation damage or clogging, further reducing the production. Therefore, it is crucial to carefully select the appropriate acid for each reservoir to ensure maximum production benefits. As discussed in ref. [1] the standard solubility test involves a 24-hour acid treatment followed by a minimum 6-hour soaking in tap water at room temperature, and subsequent drying takes another 8 hours. The process dramatically increases the response time of the laboratory [97]. In our database, most rock samples came from the Carpathian Basin and some from non-European oil fields. The rock samples to be tested came from drill cuttings sampled at the rig-site during drilling the well and sometimes were made from drill cores. Laboratory samples were subsampled while taking care of the representative sample size depending on the grain size of the sample. Bigger chunks were cut into smaller sizes with a drill and grinder and further powdered and homogenized in a mill and agate mortar. On average, the preparation of a sample takes 20 minutes, and the laboratory measurement of ATR FTIR takes half a minute. This technique can dramatically shorten the response time of the laboratory, resulting in shorter rig time and yet reasonable analytical accuracy. Spectral data were recorded using a Spectrum 400 (PerkinElmer Inc., Waltham, MA, USA) FTIR spectrometer with a universal attenuated total reflection (UATR) attachment containing a single reflective diamond/ZnSe composite crystal. The rock samples were scanned in transmission mode in the mid-infrared (MIR) range of electromagnetic radiation. Six scans with a resolution of 4 cm^{-1} were recorded for each spectrum in the wavenumber range between 4000 and 600 cm^{-1}. Spectral measurements and data acquisition were performed using Spectrum 10.5.4 software (PerkinElmer Inc., Waltham, MA, USA) [1].

5.2.2 Methodology

A short description of the method: create a link between the solubilities and the reduced number of spectral features, (wavenumbers from 4000 to 2520 and from 1620 to 600 cm^{-1}). The flowchart of Figure 5.2 summarises the main ideas and steps in producing artificial spectra.

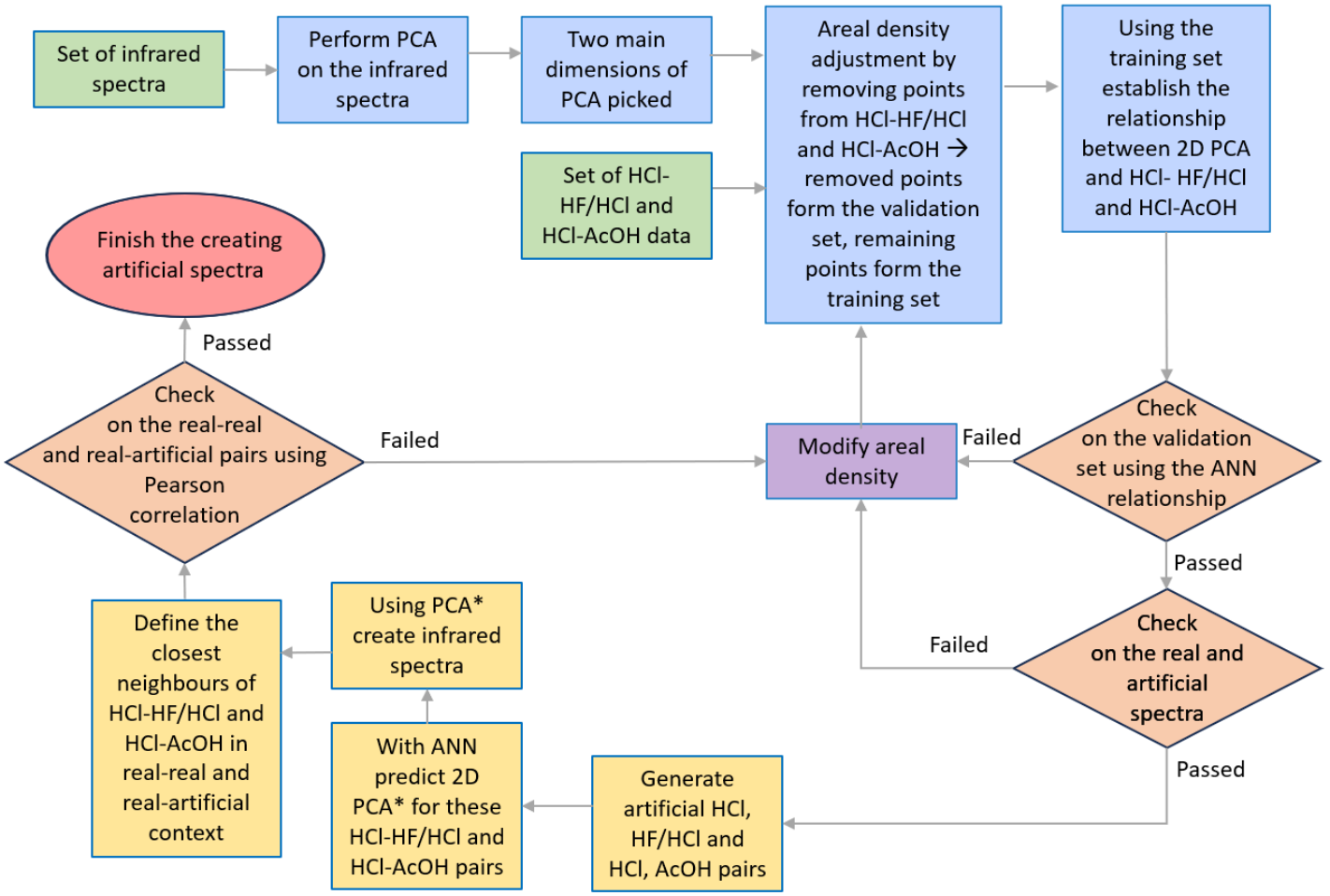


FIGURE 5.2: The flow chart of the methodology. The green color is the input data, blue is pre-processing before making artificial spectra, the orange is the decision point, yellow is the generation of artificial data, and red is the finish of the process.

The first step in making artificial infrared spectra is to do the usual spectral pre-processing, such as standard normal variate (SNV) and baseline correction. The second step is to investigate the solubility data. The samples come from hydrocarbon-bearing geological formations, and they do not evenly represent the variability of the solubility data. The data used in this study pertains to authentic drilling rock samples. The empirical correlation between these samples is presented in the form of Figure 5.3, which plots their solubility against two variables (HCl - HF/HCl and HCl - AcOH). The graph highlights distinct clusters, which can be attributed to the unique properties of actual rock samples. We add new samples to the investigated parameter space while maintaining the validity of the relationship between rock samples' solubility in acids and their infrared spectra. In Figure 5.3, the diagram on the left shows the S1 (solubility in HCl) and S2 (solubility in HF/HCl) solubility data pairs of sandstone samples. The diagram on the right shows the S1 (solubility in HCl), S2 (the solubility in acetic acid) solubility values of carbonate rock samples.

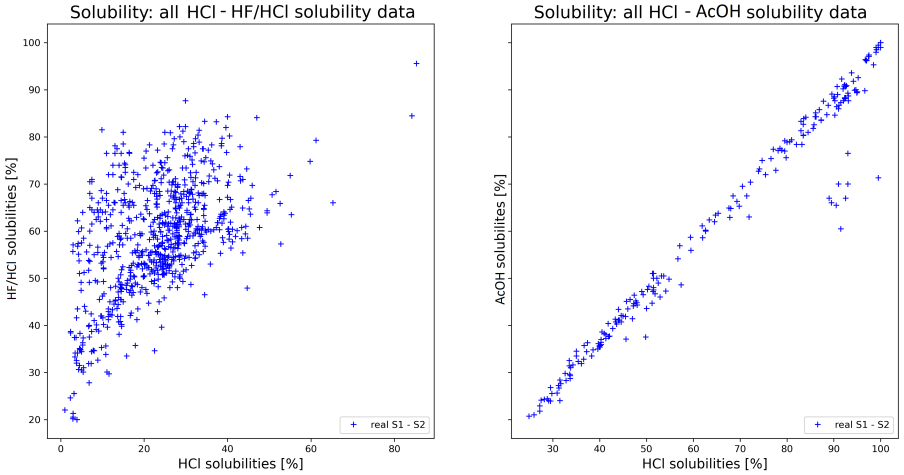


FIGURE 5.3: Solubility data of various rock samples in case of sandstone [HCl (S1) - HF/HCl (S2), *left*] and carbonate rock [HCl (S1) - acetic acid (S2), *right*] samples. The presented graph depicts the acidification outcomes of genuine drilling rock samples, wherein the irregular distribution is attributed to the authentic dataset. Our objective is to utilize the rule-based technique developed by our team to generate spectral data that accurately represents the target variable values for underrepresented or incomplete regions of the graph.

This Figure demonstrates a varying area density of points. There are areas with more data and some other segments have only a few solubility data in the diagram for sandstone samples. Two groups of solubility data were created within the HCl - HF/HCl data frame. The first group is the training group, and the second is

the validation for the ANN. If the uneven distribution of data items prevails in some areas in the S1-S2 diagram (Figure 5.3, our model might be biased and suffer from over-fitting. To mitigate this issue, we employed a distance-based density algorithm to eliminate data points from the training set that are within an arbitrary threshold - in this particulate case it was 2% in terms of the Euclidean distance in the S1-S2 space. If we could not eliminate the issue, our goal was to minimize the error. When we removed such a point from the training group, we added these points to the validation group. Finally, the training set of the sandstone samples counted 199 data items, and the validation set did so with 579. We executed the same process for the HCl and acetic acid solubility data pairs. Figure 5.4 shows these data groups in separate diagrams. Given our objective to obtain a set of points with a specified threshold distance from any neighboring points, this served as our primary criterion. Consequently, we did not utilize the widely adopted Kennard-Stone method [98], as we sought to increase the number of samples for infrequent areas. Moreover, we generated spectra with comparable properties as well to the real samples to validate our methodology (Figure 5.4). Using Principal Component Analysis (PCA), we performed a projection of the feature count (consisting of wavenumbers-absorbance data pairs) from the original 2500-dimensional space to a 2-dimensional space, which explains 87% of the variance (sandstone samples with HCl and HF/HCl). In the fourth step of the process, we applied a neural network to establish a relationship between the S1 and S2 data pairs and the PC components of the infrared spectra. A multilayer perceptron type of neural network was used, which had the following structure: I2, H12, H16, and O2, where I and O stand for the input and the output layers, and the numbers behind them indicate the number of neurons. H stands for the hidden layer and the number after the letter also indicates the number of neurons in that specific layer. The activation function was ReLU (rectified linear unit). The latent variables (loadings, scores) are calculated during the PCA analysis. The main idea is that using these loadings and score values, we can recalculate the spectra backwards [99, 100]. We predicted the training and validation samples' PC-s and recreated their infrared spectra with fewer (first and second) predicted PC-s. We checked the validity of the relationship by calculating the root mean squared error (RMSE) for both the training and the validation sets. The RMSE for the training set of the sandstone samples was 0.1114, while for the validation set, it was 0.1020. We found these RMSE values satisfactory, and with this, we considered our neural network model validated. We then used this trained network to predict the PC-s based on the

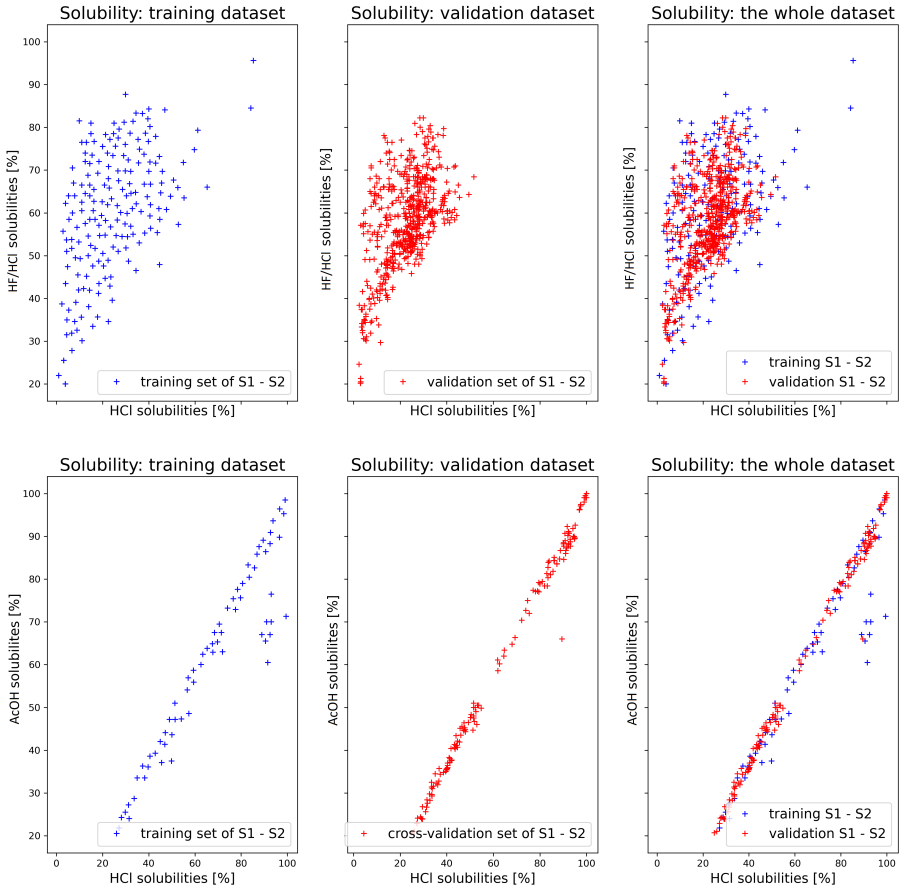
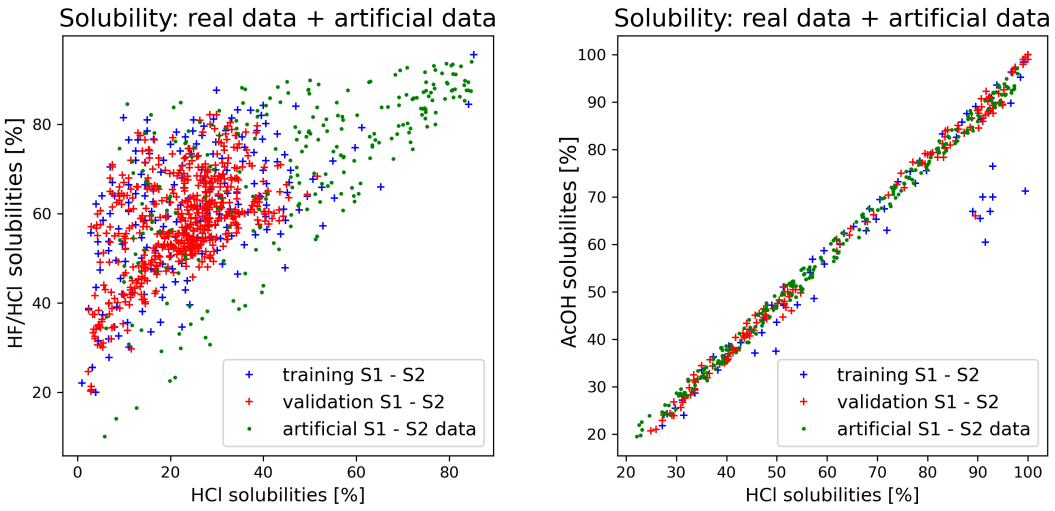


FIGURE 5.4: The solubility data of sandstone (upper row) and carbonate rock (lower row) samples grouped into training and validation sets (left and middle columns, respectively). The blue are the training samples, and the red are the validation samples.

artificial (randomly generated S1-S2 data pairs). We created randomly generated S1-S2 data pairs in the diagram only if they complied with a rule. This rule comes from the observed solubility data of the real-world rock samples. In the case of HCl and HF/HCl solubility, S2 > S1. It means that the solubility of sandstone samples in HF/HCl should always be higher than in HCl only. In the case of carbonate rock samples, it is reversed, the solubility in HCl is more significant than in acetic acid, S1 > S2. See the diagram in Figure 5.5 for the artificially produced solubility data of the sandstone samples.

We considered the suggestions of Guo *et al.*, focusing on adjusting the areal density



(A) Sandstone (B) Carbonate

FIGURE 5.5: Real and artificial solubility data for sandstone [sample size: training = 199, validation = 579, artificial = 250] (*left*) and carbonate rock [sample size: training = 66, validation = 149, artificial = 250] (*right*) samples.

of points in our case, as this significantly affected the correlation between solubility and PCA [101]. We used the neural network prediction based on the generated S1 and S2 data pairs. PC-s for creating the artificial infrared spectra for sandstone and carbonate samples, and depicted them in a separate diagram next to the original infrared spectra. Figures 5.6 and 5.7 show these two sets (sandstone and carbonate rock samples of infrared spectra.

5.3 Results

Throughout the process of sample selection, model construction, and result evaluation, we diligently monitored and avoided the most common errors, which Héberger *et al.* summarized [102]. We have checked the artificial infrared spectra in three ways.

1. Visual inspection (comparing the generated infrared spectra to the real ones) is the most straightforward way of checking.

2. In quantitative terms, how close are the measured and the artificial (in this case, the reproduced spectra) infrared spectra to each other provided their samples' solubility data are very similar.

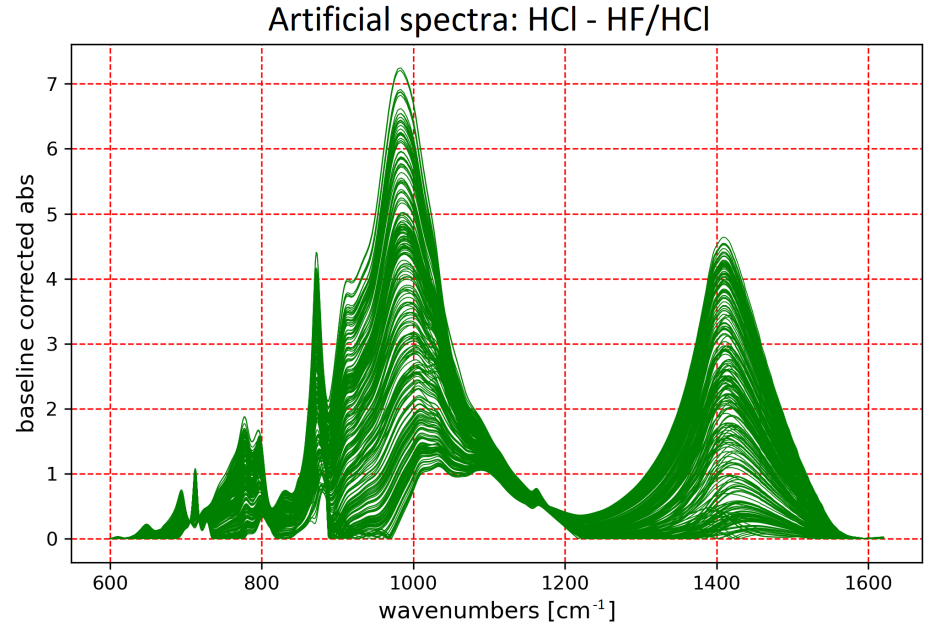


FIGURE 5.6: Artificial IR spectra of sandstone samples

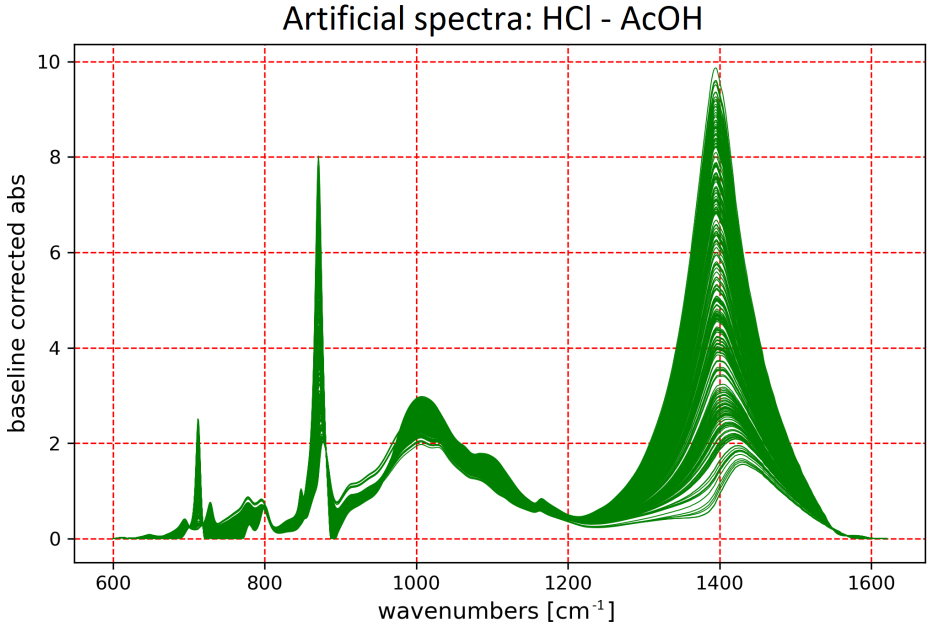


FIGURE 5.7: Artificial IR spectra of carbonate rock samples

3. We compared the average Pearson correlation coefficients between the real-real and real-artificial infrared spectra deriving from the two closest neighbors in the S1-S2 domain.

Visual inspection of Figure 5.8 shows the artificially created spectra for samples

dissolved in HCl and HF/HCl acids. The visual check compares the details of the infrared spectra: peak locations and intensities. Since we have depicted all the infrared spectra in one picture, the range of peak intensities is straightforward. In our methodology, we trained the score values of the principal component analysis using an artificial neural network and generated artificial spectra as a result of the training. So after analysing the actual spectra of the training set (blue) and comparing them to the reconstructed spectra (red), it can be inferred that the absorbance range of the reconstructed spectra is narrower, that is, the red spectra are a subset of the blue spectra. This can be attributed to the projection of the data onto a lower subspace via PCA, which resulted in the exclusion of smaller eigenvalue noise components. Of course, it can be refined by considering more principal components.

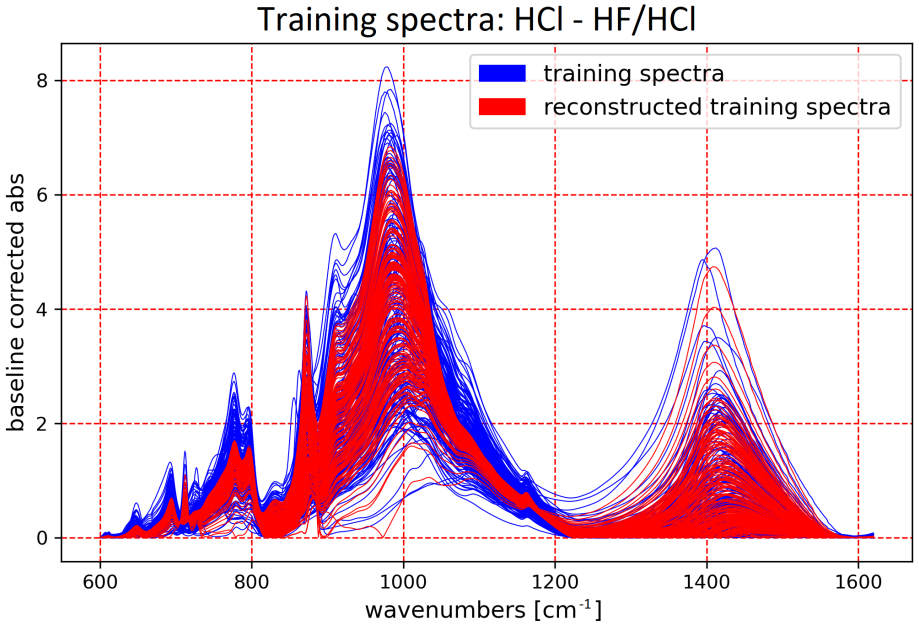


FIGURE 5.8: Real and artificial IR spectra of the training group samples

After training the neural network, we can compare the predicted PCA features to the original ones. If the model shows overfitting or low performance, this comparison reacted sensitively. While fine-tuning the ANN, we considered the model being overfitted if the RMSE value was much lower than that of the validation set. We continued the fine-tuning of the ANN until the RMSE values were close. For the training set, it was 0.1114, and for the validation set, it was 0.1020. In cases that showed an unsatisfactory fit, the calculated PCA values scattered in a narrower region, and they could not spread out as the real PCA values did in

Figure 5.9 or any other graphical depiction of these data. In Figure 5.9, the red dots depict the predicted features, which scatter in a narrower range for the model data than it does for the original values. This visual comparison of the PCA values played a role in the fine-tuning step of the ANN. The number of PCA features was two because we had only two parameters (S1 and S2) to predict the PCA features.

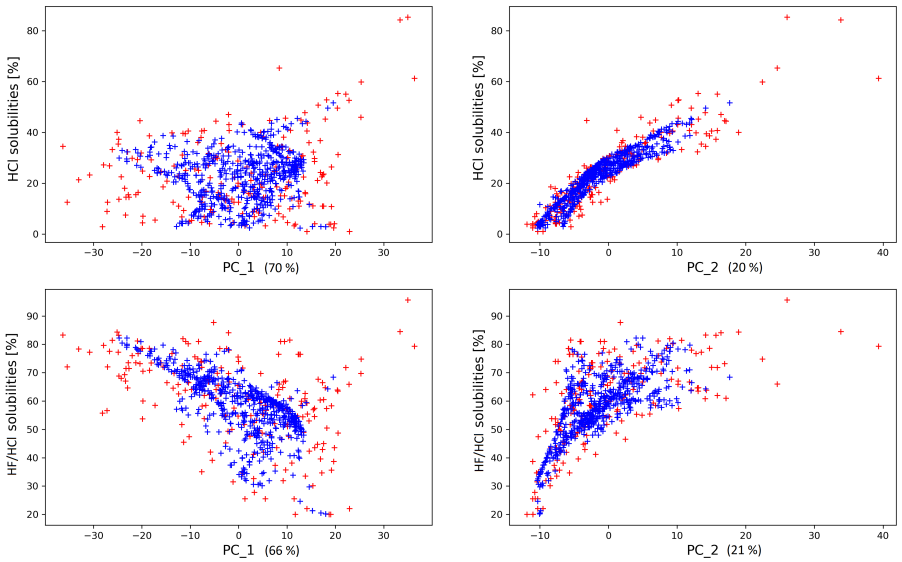


FIGURE 5.9: Real and artificial HCl solubility (S1) and HF/HCl solubility (S2) - PCA pairs

From the method's perspective, a quantitative comparison is more suitable for checking artificially created infrared spectra using quantitative measures. For this purpose, we applied Pearson correlation coefficients between the original and their artificial versions and the closest neighbors of real-real and real-artificial samples' spectra. The method consists of the following steps:

1. Create artificial spectra using the original S1-S2 data pairs.
2. Compare the original spectrum to its artificial counterpart in a pairwise way.
3. Calculate their Pearson correlation coefficient for the original and artificial pairs, both in the training and the validation groups.

This comparison assesses how closely the artificial spectra resemble their original counterparts, using only two PCA features to construct the artificial infrared spectra. In the case of the HF/HCl solubility samples the average Pearson coefficient in the training group was 0.9827, in the validation group the average Pearson coefficient was 0.9858. The lowest correlation coefficients in the training and validation groups were 0.7488 and 0.8254, respectively. The maximum values were 0.9995 and 0.9994. In the case of the HCl-AcOH solubility samples the correlation coefficients were as follows: 0.9704 (training), 0.9728 (validation), 0.8768 (training - minimum), 0.7867 (validation - minimum), 0.9980 (training - maximum), and 0.9978 (validation - maximum). See the histograms in Figures 5.10 and 5.11. We considered the artificially created spectra acceptable based on these correlation coefficient values [103]. The following points must be met for artificial S1-S2 data pairs and artificial spectra:

1. Create artificial S1-S2 data pairs that comply with the rule mentioned above.
2. Create artificial spectra of these samples using the PCA values and the trained neural network.
3. Identify the closest neighbors in the S1-S2 diagram using real-real and real-artificial neighbor investigations. These neighbors can be real-real and real-artificial data points.
4. Select their infrared spectra and calculate their Pearson correlation coefficient for the real-real and the real-artificial pairs.
5. Compare the Pearson coefficients for both groups of spectra (real-real and real-artificial).

We can run an algorithm to select the closest solubility neighbors and calculate the Pearson correlation coefficients between their infrared spectra to see how similar or different they are. We have done this for the real-real and the real-artificial infrared spectra of the closest solubility pairs. With this, we aimed to determine how closely the artificial infrared spectra resemble the real ones. Table 5.1 contains the Euclidean distance-related values, such as average Euclidean distance and minimum and maximum distance between the two closest neighbors computed for the whole dataset. In the case of the HCl-HF samples, the average distance

TABLE 5.1: Comparison of artificial and real spectra in the four investigated cases.

Comparison	**HCl - HF/HCl real-real**	**HCl - HF/HCl real-artificial**	**HCl - AcOH real-real**	**HCl - AcOH real-artificial**
number of samples	778	778	215	215
average Euclidean distance	0.868	2.458	0.954	1.545
min. dist.	0.100	0.047	0.100	0.008
max dist.	11.15	6.89	6.63	18.55
average Pearson coeff.	0.939	0.958	0.650	0.765
min Pearson coeff.	0.535	0.562	-0.135	0.237
max Pearson coeff.	0.9995	0.9991	0.999	0.992
average MSE [A^2]*	0.157	0.102	0.511	0.591
min MSE [A^2]	0.0006	0.0016	0.0005	0.0250
max MSE [A^2]	0.983	0.772	1.813	2.440
average $RMSE$ [A]	0.353	0.283	0.629	0.704
min $RMSE$ [A]	0.024	0.039	0.023	0.158
max $RMSE$ [A]	0.992	0.879	1.347	1.562

*A: absorbance

between the real and artificial samples is larger than between the real-real neighbors, which is due to the higher number of artificial S1-S2 data pairs in the region of scarcely populated real data. The average correlation between the real and artificial spectra is better (higher the Pearson correlation coefficient than in the real-real dataset. In the case of the carbonate rock samples, the same observation can be made.

The histograms of the Pearson correlation coefficients of the nearest neighbors can be seen in Figures 5.12 and 5.13. They tell us that the artificial infrared spectra (of those samples that have the closest solubility values in those two acids) are at least as visually similar to their nearest original counterparts as the real infrared spectra are to their nearest original ones. In the case of the HF/HCl samples, the real-artificial neighbors have a histogram with a higher maximum and steeper fall reaching the bottom sooner than the distribution of the Pearson coefficients does in the case of the real-real neighbors.

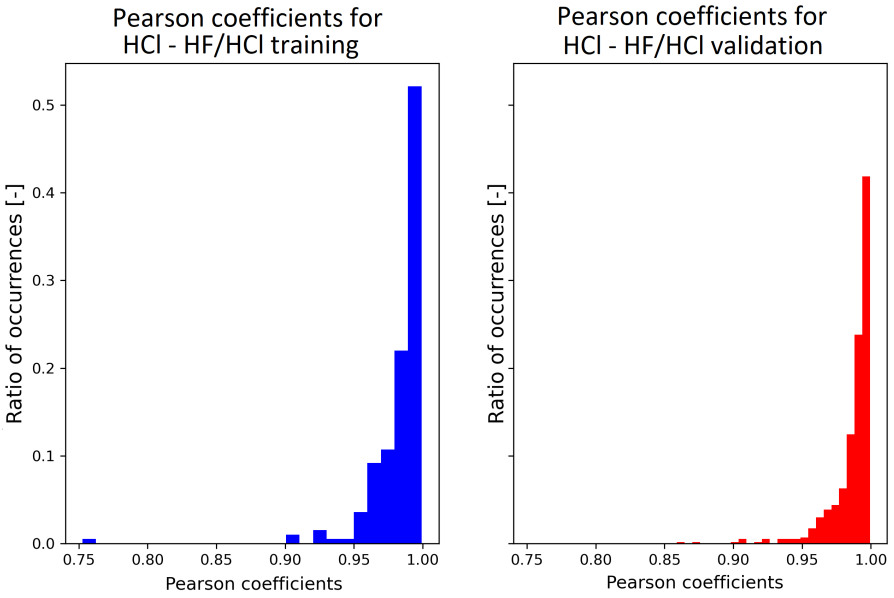


FIGURE 5.10: The histograms of Pearson correlation coefficients (sandstone samples)

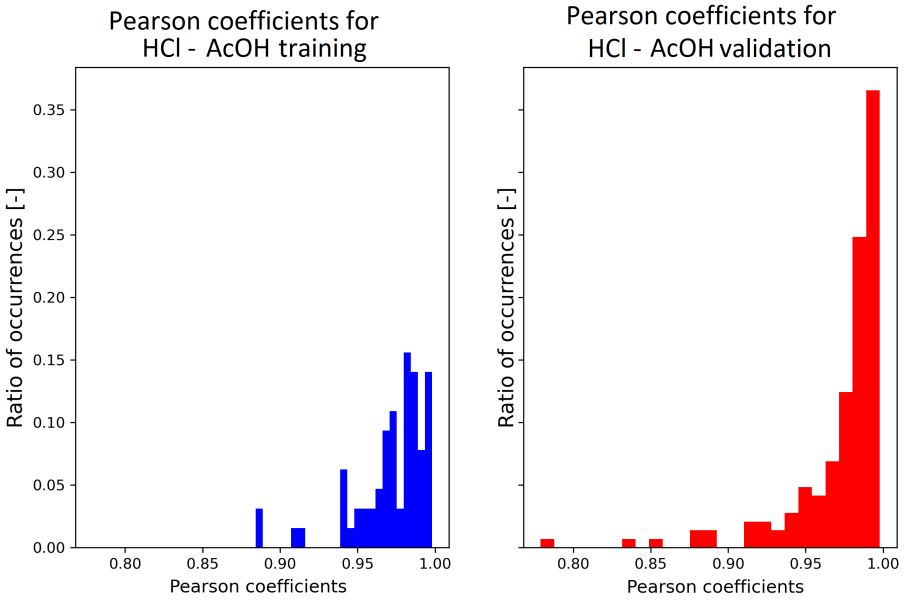


FIGURE 5.11: The histograms of Pearson correlation coefficients (carbonate rock samples)

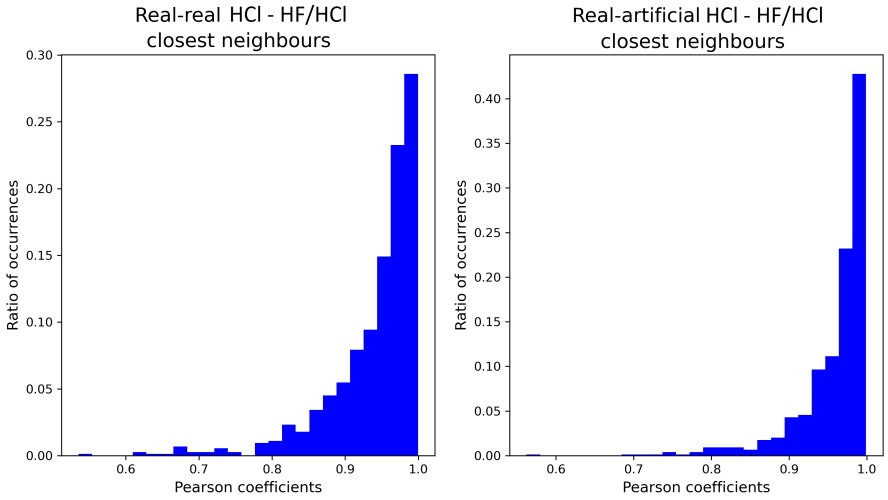


FIGURE 5.12: Histograms of the Pearson correlation coefficients for real-real and real-artificial closest neighbors in sandstone samples.

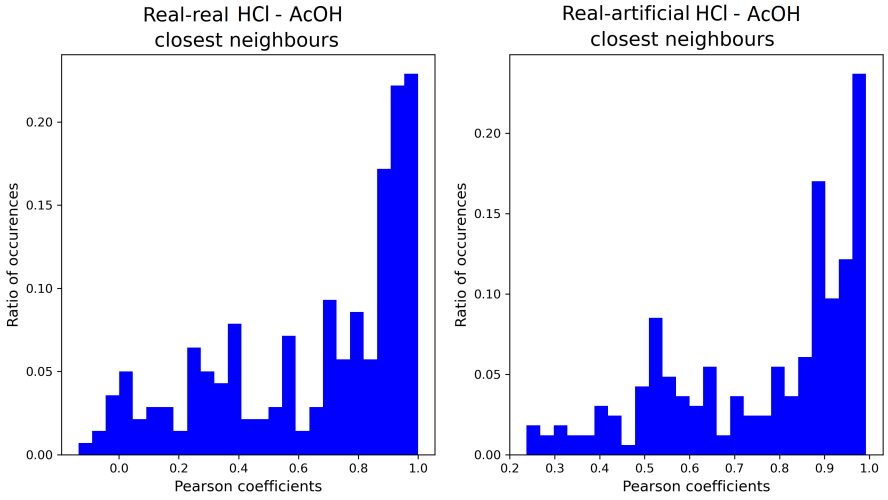


FIGURE 5.13: Histograms of the Pearson correlation coefficients for real-real and real-artificial closest neighbors in carbonate rock samples.

5.4 Discussion

The methodology can be used to generate artificial spectra. The method in Figure 5.2 differs from the previous techniques in that it can purposefully generate synthetic data for sections with missing values. Artificial spectra were generated by adding noise to the appropriate wavenumber ranges [84]; in other cases, artificial spectra were generated using a Monte Carlo simulation. Han *et al.* used the Monte Carlo simulation to generate Raman spectra. Ten peaks in the Raman spectrum were selected, and the spectra were generated by adjusting the concentration and noise scale. The mixture spectra were all randomly and independently generated in this article [104], like the FTIR technique.

The other option to improve the models is to use spectral interference subtraction (SIS). The SIS enables the removal of a significant portion of known additive interference, resulting in a more straightforward chemometric model to interpret. If these interference effects are estimated and extracted in the preprocessing stage, it reduces costs and improves interpretability. SIS preprocessing can be beneficial when it is challenging to generate calibration samples that capture the total variability of the expected future sample qualities [105]. Another empirical approach is that Tong *et al.* created simulated spectra for 381 diesels and 401 NIR spectral data points based on a mathematical formula [106]. Selzer *et al.* emphasise that infrared spectroscopy is extremely useful for identifying substances, for example, when comparing the experimental spectrum with the reference spectra in the database. This article explores organic compounds and utilises neural network techniques to facilitate rapid access to reference spectra. This method is based on experimental data that examine whether the database containing 13373 infrared spectra provides sufficient spectral information to perform prediction experiments. Although the prediction of organic compounds is highly complex, the technique nonetheless provides good prediction results. The correlation coefficient for six compounds is 0.9 compared to the simulated and experimental spectral sets from the 16 compounds. The correlation coefficient for the other six spectra ranged from 0.8 to 0.9, indicating high similarity. Two spectra showed a correlation coefficient of 0.7-0.8, which is acceptable. Only two spectra have a correlation coefficient of less than 0.2, indicating poor similarity. The validation methodology presented in ref. [107] is similar to our verification methodology. Correlation coefficient values are also calculated to compare real and artificial spectra. This technique shows that we can generate artificial spectra using neural networks that match

the spectra of real samples with high accuracy. These approaches differ from the methodology presented in this paper. In our case, we do not want to create new patterns with noise or Monte Carlo simulation, but we want to generate patterns for our entire variability space purposefully.

5.5 Chapter summary

Tested on rocks, it shows that artificially produced and real spectra are very similar. Furthermore, we have created artificial spectra that meet the following requirements:

1. are similar to the measured spectra,
2. their solubility characteristics are similar to those of the scale (also in all three solubilities (HCl, HF-HCl, AcOH)),
3. they were generated to fill in our missing parameter space with few real samples,
4. with the solution, we can improve our original distributions in a targeted way, so that we generate samples where the number of real samples is small.

The method can also be used in other systems. Among the tested case studies, there is another example in the Appendix A.6. In this case too, the artificial spectra can be produced in a suitable quality. The method's performance improved if the chosen number of principal components could explain as much variance as possible. The performance of the presented method depends to a large extent on the number of PCs used for spectrum generation. The method is suitable for generating infrared spectra that enrich the parameter space along the constraints previously laid down. The method's performance improved if the chosen number of principal components could explain as much variance as possible. The number of principal components increases with the number of features of the system, which contributes to the performance of the method. The presented method is capable of generating infrared spectra for synthetic samples that lie within the parameter space.

Figure 5.2 shows the schematic diagram of the methodology, based on which we created artificial spectra. The quality of the spectra produced in this way underwent statistical and visual validation and can be considered adequate. The essence of the methodology is that we produce artificial spectra where there are few or no data points in our starting data series. Various restrictions must be met before an artificial spectrum can be created. Outlook, future plans: the high number of generated samples makes the usage of a convolutional neural network feasible.

Chapter 6

Edge computing and machine learning-based framework for software sensor development

6.1 Introduction

Software sensors determine critical parameters of complex chemical processes that are difficult to measure. The development and application of software sensors in the chemical industry have been prevalent in the last decade [2]. However, no suitable solution has been developed for their economic operation and life cycle tracking, so the number of devices is low today. The development of a methodology for cost-, energy- and resource-efficient operation of models facilitates continuous real-time software sensors [3],[4]. Several sensors are used in chemical processes to monitor critical process variables such as product quality and process safety. Samples awaiting analysis are taken manually from the process and analyzed in laboratories. Sampling frequencies are often too low for process monitoring and control [5]. The accuracy of models built on databases with relatively small and inadequate standard deviations may give unsatisfactory results. Therefore, the beginning of modeling requires exploration and analysis of basic statistics [6].

Our goal is to present a solution that meets the above criteria and continuously supports the qualification processes. To this end, we have developed a quality assurance architecture that summarizes the building blocks required to develop such

a solution. In addition, we have developed a methodology that supports the application of ML, and we also present a case study detailing the applicability. The technology offers a solution for several different laboratories. In addition to the arguments listed above, the development aimed to reduce the environmental impact of laboratory activities and use software sensors in various industrial processes. The various ML algorithms have been developed to calculate critical parameters of the materials based on fast, environmentally friendly, and inexpensive spectroscopic measurements. The ML algorithms can learn essential parts of spectral information that can predict qualitative and quantitative parameters. For example, the chemometrics and ML methods are successful tools for testing the quality and quantity of beers [108]. Furthermore, the combination of Raman spectroscopy and ML is becoming a fast, non-destructive method for verifying the nature or origin of foods [109]. Moreover, another review focuses on biomedical FTIR applications published between 2009 and 2013, which are used for early detection of cancer by qualitative and quantitative analysis [110]. The excellent results were using these algorithms also obtained in distinguishing the origin of honey [111].

These review articles show how popular the development and application of ML algorithms based on data from laboratory devices are in various industries. First, however, we need to apply state-of-the-art methodologies to ML algorithms, such as etc. Auto ML CRoss Industry Standard Process for Machine Learning (CRISP-ML), which allow these algorithms to be updated. From the literature reviewed, it can be concluded that these models are used many times, but only for a short time, as they deteriorate over time and the development part needs to be restarted. Building and maintaining the right IT industry framework is essential for developing and day-to-day application of ML models. Our goal is to develop a framework that can be used in an industrial environment, proposing solutions to the problems outlined above and helping with quality assurance and process control. The developed framework will be developed and tested on oil industry data but can also use in medicine, the pharmaceutical industry, the food industry and waste management.

In order to ensure the quality of the products manufactured, samples taken from the production of the company processes must be subjected to quality assurance laboratory testing. Therefore, a vital issue is predicting the arrival of production samples in the laboratory, which will help allocate resources. The CRoss Industry

Standard Process for Data Mining (CRISP-DM) system is used to solve this problem. The system consists of three iteration processes, and an AutoML procedure has been used to allow the comparison and configuration of ML algorithms [112].

The process system engineering (PSE) is now more than fifty years old in the chemical engineering industry, mainly focusing on computer power and the further development of chemical processes using them to promote better plant design, operation, and better product quality for more prosperous, more environmentally friendly, and more efficient production [113]. The key areas such as IoT, cloud-, fog-, edge computing, and ML contribute to a more economical, environmentally friendly, and efficient operation of various processes. ML algorithms have now been adopted to track the quality of multiple industrial processes effectively [114]. In addition to the various ML solutions, increasing the efficiency, development and maintenance of standard data models and ML algorithms is still to be worked out [115]. Due to the complexity of chemical processes, it is challenging to incorporate ML models into continuous or batch production processes. Therefore, improving the integration capacity of corporate governance systems and ML processes is needed. The analysis of processes seems to be a prevalent and innovative solution from the pharmaceutical industry. This topic is called process analytical technology (PAT)[116]. The basis for achieving the primary objectives mentioned above is that the available IoT and edge computing tools continuously support operational activities with ML models. The models need to be updated based on historical data and practical information. In addition, ML models, like machines, need maintenance because the models can land or break over time. Therefore, continuous monitoring and maintenance are required for more accurate and robust model results. An industrial data science framework will help address these challenges. Furthermore, companies need to pay more attention to maintaining their ML competencies. In addition to maintenance and supervision, a well-developed architecture and a well-documented framework are key. The edge computing performed by IoT devices communicating with the remote cloud, plays an essential role in industrial digitization. The edge computing architecture can be an ideal solution to minimize delays for intelligent factories and smart cities [117]. The IoT and edge use a gateway to communicate.

A literature review shows that many of the articles uses Industrial 4.0 devices, but the prevalence of a large number of software sensors is not yet visible. The problem is that an installed software sensor specializes in basic parameters that are

difficult to measure. As a result, specialists are required to interpret laboratory measurements. In addition, the maintenance of the model and the tuning of its parameters require continuous monitoring. The purpose of this chapter is to explore how software sensors can be developed, deployed, and continuously monitored and maintained with edge and cloud computing.

The following main points show the roadmap that will contribute to the methodology we have developed.

- Section 6.2 describes the related work, overview of cloud - and edge computing studies used in chemical engineering. The literature review shows that there are quite a few initiatives in these areas, mainly in the healthcare and pharmaceutical industries.
- Section 6.3 presents the elements of a framework proposed to address the challenges of a general quality assurance laboratory. The framework helps to develop and maintain models.
- Section 6.4 presents a case study supporting the work of the quality assurance laboratory by comparing the performance of different ML models.
- The final Section 6.5 summarizes conclusions and research recommendations.

6.2 Overview of cloud computing and software sensor development in chemical engineering

This second section presents the importance of the topic, the related literature, patents. The preferred reporting items for systematic reviews and meta-analyses (PRISMA) methodology is used to review the many scientific sources systematically.

6.2.1 Literature review

The PRISMA statement includes a report outlining the area of study and assisting the researcher in selecting relevant literature in a systematic review [118]. The analysis makes it easy to review the literature on Scopus or even the Web of

Science [119]. Resources related to the topic should be described with a systematic overview and a high degree of methodological detail. The flowchart is an integral part of the methodological description of the PRISMA review. The use of data-driven predictive models is becoming increasingly popular in the engineering and manufacturing sectors. During the literature research, we searched for literature with several word combinations in the search for Scopus. First, the central area of the topic was the edge, computing, software and sensor; the number of articles was 388, of which 14 were chemistry papers. The keywords of other searching were cloud-, egde-, fog computing and ML or ensemble learning 168 review articles from which were 6 chemistry. Next, the chemistry laboratory and ML chose as keywords; there were 207 articles from which 22 were relevant and related to chemical engineering. Finally, there were 17 relevant pieces of literature on edge computing, ML, and chemical engineering. Each combination search shows a few scientific studies on chemical industry software sensors, edge computation, and ML.

The selection criteria were the relevant literature on edge computing and software sensors used in the industry. As a result, the studies found on Scopus were processed using the PRISMA methodology (Figure 6.1). The network diagram summarizes which keywords appear in the scientific journals "edge computing", "software", and "sensors" during the first search of 388 articles (Figure 6.2). Red shows the connection between the network and the application device. The colors green and blue illustrate the devices, methods, and data connecting IoT to applications. The yellow keywords summarize the computing and the sensors connected to extreme computing and IoT.

The five groups shown in the Figure 6.2 are the following. The red group contains the wireless sensor networks, the IoT industrial solutions for the wireless networks. The purple group includes 5G technologies and visualizations. Yellow focuses on the fog- and cloud computing parts, while the green group deals with ML, edge, and big data. Finally, the blue group is for in-depth learning of artificial intelligence, energy efficiency and visualization. It can also be seen from the network of keywords that the edge computer, IoT and ML algorithms have been intertwined technologies for years. However, little research has been presented on the maintenance and monitoring of the algorithms presented in the literature.

Edge computing tools play a significant role in the maintenance, onsite access, and developed ML models. The aim of edge computing is to bring cloud resources and

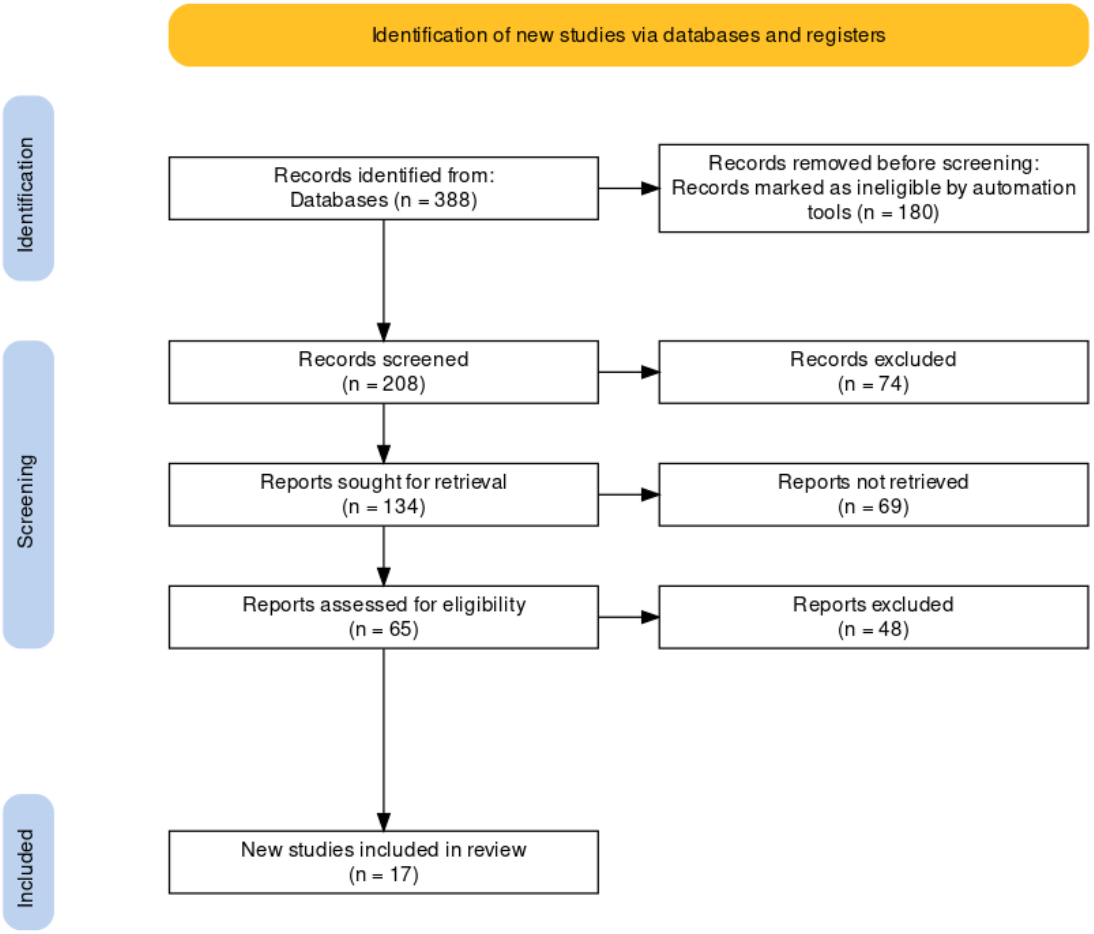


FIGURE 6.1: Grouping of studies according to PRISMA methodology. PRISMA chart representing the methodology of the literature review based on Scopus database. As can be seen 388 articles started the analysis, but 17 were included in the study.

services closer to the things which are generating data [120]. Cloud computing provides convenient, on-demand network access to a shared set of configurable computing resources that can be quickly deployed and released with minimal supervision [121]. A group of IoT infrastructures that connect different objects and allow them to be managed, accessed and mined by the data they generate and communicate with other devices [122]. In a broader sense, it extends network connectivity and computing power to objects, devices, sensors, or objects that are not computers [123]. Furthermore, IoT devices play a prominent role in the wireless detection and transmission of signals. Different gateways and devices on the edge of the internet play a vital role in the operation of modern companies [120].

The digitization of production lines plays a key role in the efficiency of several production units, such as predictive maintenance and quality assurance.

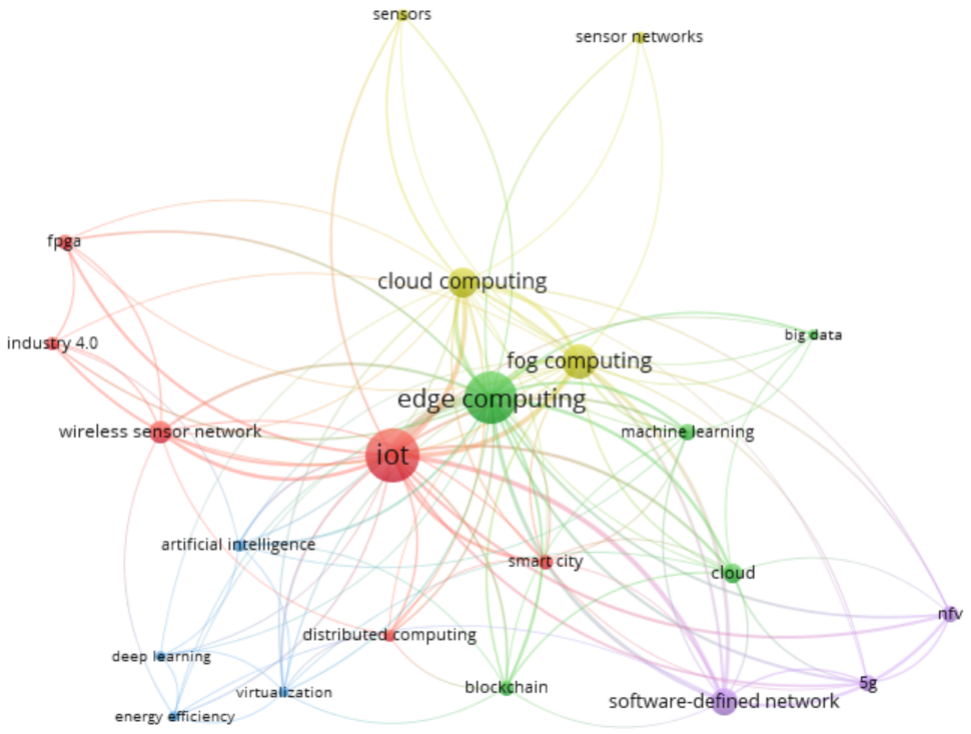


FIGURE 6.2: The co-occurrence network of the keywords of "edge computing" and "software sensor" -related articles in the Scopus database. As can be seen the papers are clustered into four categories. Red show the IoT, the green the edge, the yellow fog-, cloud computing and sensors, and blue colored module shows the artificial intelligent and deep learning modules.

Monitoring the condition and process of data-driven machines in a fog-based framework is of great importance in cyber manufacturing. The communication protocol presented in this article is MTConnect, an open set of standards on which it is based standard internet technologies, and Amazon Machine Image (AMI) defines the primary operating system. Manufacturers can use MTConnect to monitor real-time machining and process data, speed, temperature, emergency shutdown, and performance status. Furthermore, because this protocol is implemented as a web service, it is easily accessible to any device that connects to the machine's network [124]. In addition to fog and cloud calculations, edge calculations are also used in many cases. The point is to carry out on-site operations, make forecasts and thus speed up processes. Recently, prevalent topics like cloud, edge and fog computing and the IoT are essential for developing smart factories. Osmotic Computing has elements that enable more coordinated computing, networking, storage, data transfer, and management between cloud and IoT devices in computing layers of edge [125].

6.2.2 Related patents, trends and benchmarks

The patent research can be seen from the results that this article's topic is becoming more and more popular year after year, not only the number of articles in the literature but also the number of patents shows a significant increase (Figure 6.3).

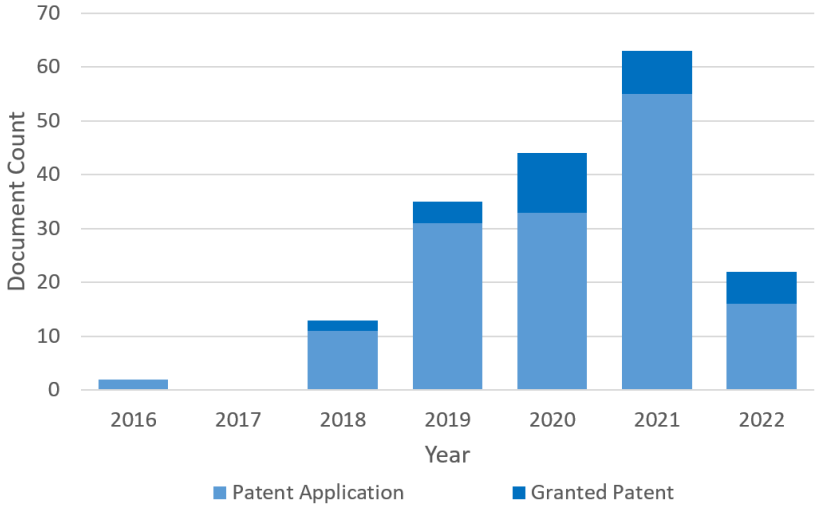


FIGURE 6.3: Number of patents in recent years. The patents researched on Lens database, the keywords were "edge computing" and "software sensor". In addition, there is a significant increase in their annual breakdown, which shows the relevance of the topic today. (Access the data at *lens.org* as of 31 March 2022)

The patents review shows that Fraunhofer Ges Forschung is at the forefront of edge computing and software sensor technology. The Fraunhofer is leading applied research organization of the world. Prioritizing future-relevant technologies and commercializing its findings in business and industry plays a significant role in the innovation process, such as data innovation development in the different industries, the architecture of the IoT, data mining and ML algorithms development. This company had 49 patents at the end of 2021, but even Hewlett Packard, Version Patent, Sony, Abb, and Intel Corporation hold quite a few patents this research was based on *lens.org*. The patents demonstrate the security capabilities of intelligent computing and Industrial IoT devices. For example, one presents a network device that analyzes size and influences packet delivery by a threshold [126]. Furthermore, there are patents in which neural networks transmit the results of each model to the final edge computing. The neural network transformation system can be carried forward by using the disguised input data as input to the neural network model. Applying it to the teaching data generated at the first level is the input to the neural level at the next level. The process can be further adapted to

pass output data to clients [127]. Another patent discloses disabling live devices that include a processing resource that communicates with a memory resource [128]. There is also a patent that demonstrates the distributed computational mechanism of ML models. The essence of the patent is that it optimizes to run multiple calculations in a hierarchical system, so solving a cost function can give better results [129]. The assignment of ML models to devices is addressed in several patents, one of which presents a method that provides estimates and the score of estimates [130]. Another patent offers a solution for optimizing laboratory procedures. The invention facilitates alternative processes and supports laboratory processes through cost optimization. The essence of the patent is to store data of laboratory processes in an aggregated and structured form that can be easily interpreted and reproduced in laboratories [131].

Based on research in the literature and patents in the field, it can be concluded that ML tools are becoming more widespread in industrial environments. However, there is a tendency in research topics to focus on data collection and model development in the cloud solution, usually using good ML models to ensure quality in minor proof of concept (PoC) projects. It can be explained by the fact that maintaining the accuracy of the models requires constant maintenance, as the performance of the models may deteriorate over time. Maintenance is time-consuming and resource-intensive, but this challenge can solve with the correct methodology, edge- and cloud computing methods, and appropriate architecture.

6.3 The proposed framework

The following section describes the elements of CRISP-ML following principles similar to CRISP-DM and presents the main steps in the sequence of model development (Section 6.3.1). The concept of cloud-based development of software sensors and its essential tools such as IoT and edge computing are described in Section 6.3.2. Follow the predictive model markup language (PMML) in section 6.3.3 to help you apply, develop, and monitor your models, as well as the lean six sigma principles that are essential for development (Section 6.3.4).

6.3.1 CRISP-ML for the sustainability of the models

The following data science technology concept is to make data and models available to laboratories and plants at any time of the day. Of course, the goal is to use the latest models as accurately as possible to support chemical processes. The enterprise cloud service needs to be supplemented in a short period with the results of fast, environmentally friendly, and inexpensive measurements of the samples so that predictions can be made from the results obtained quickly for the broad qualification of the products. In addition to uploading data from devices that perform fast measurements, it is also essential to access enterprise resource planning (ERP) data. In addition to data transport, pretreatment, model development, continuous development and maintenance of models are paramount. The application of the CRISP-ML methodology helps in this. The difference between CRISP-DM and CRISP-ML is that the CRISP-DM focuses on data mining and does not cover the application of different ML models inferring in real-time over a long period. Furthermore, the CRISP-DM does not give guidance on the Quality Assurance methodology. This shortcoming is evident in the standards of information technology and the process models for data mining [115]. The life cycle of the development of data science models is shown in the Figure 6.4.

In order to monitor quality assurance in an enterprise environment, it is essential to establish standard process modeling for the development of ML models. In contrast, there are still many developments where this is not happening. Due to the growing demand and recent quality assurance for the models, the CRISP-ML methodology based on the CRISP-DM data mining model has been developed. CRISP-ML quality assurance requirements include data quality, model robustness, and expected model performance. The essence of the approach is to articulate risks that could negatively affect application efficiency and success of ML models. For example, the patterns that make up the models can overwhelm the teaching pattern army, or outliers samples can degrade the accuracy of the models, or incorrectly selected and adjusted models can lead to over-fitting problems. During the prediction of properties that significantly affect the quality of products, the continuous validation of the models is essential, and the application of the CRISP-ML methodology helps in this (Figure 6.4).

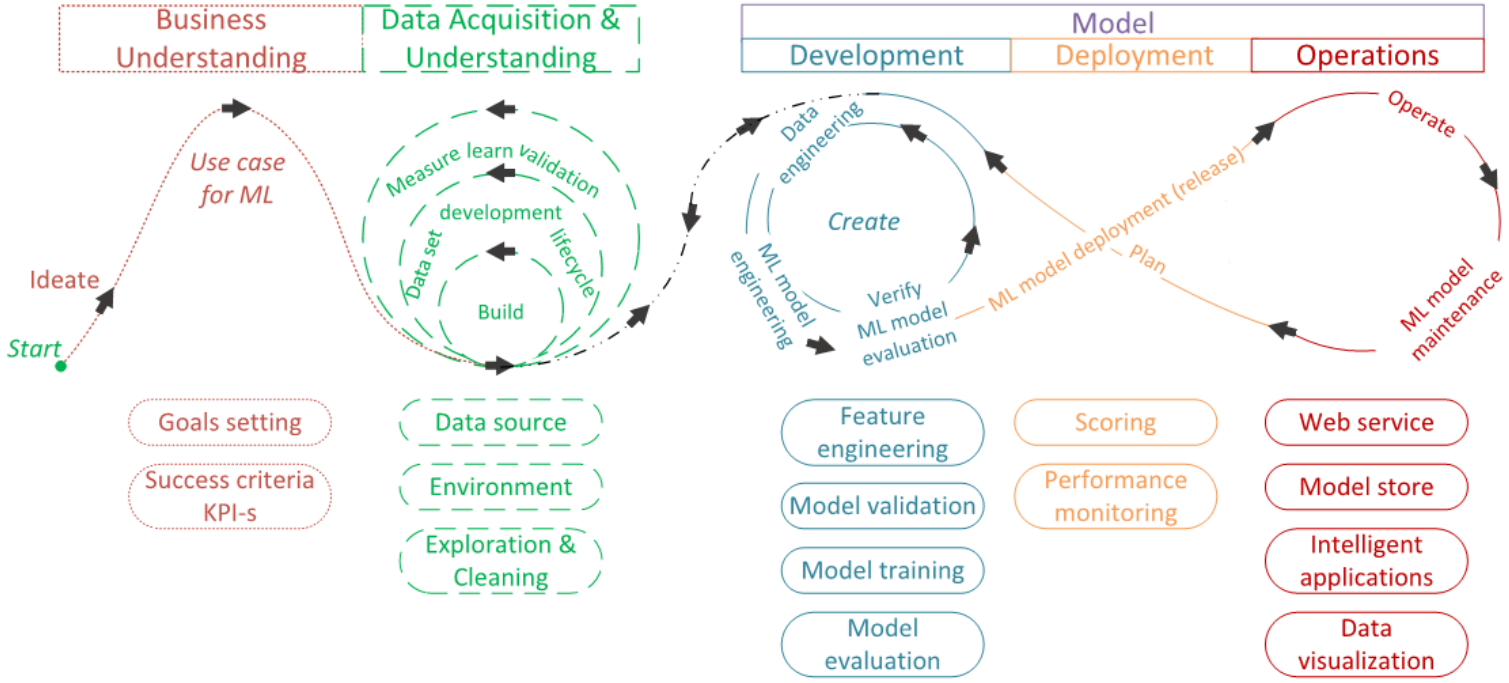


FIGURE 6.4: Data science life-cycle management and CRISP-ML The first part is understanding the business problem (brown) in which goals and success criteria need to be defined. The second stage is understanding the data (green), which involves exploring, cleaning up the data sources and building an environment. The next part is modeling, which has three sub-parts. Development of the model (blue), in which it is essential to compete and select different models. The next step is to deploy the models (orange), including monitoring the installed models. The final stage of the modeling is the operation (purple), which includes data visualization and the development of intelligent applications. The arrows in the figure also illustrate the cyclically of the development.

The different colors in Figure 6.4 show the different parts of the data scientist concept. It is important to note that this figure applies to the development of ML models in general.

- **Business Understanding**
  Projects for the development of ML applications are done by controlling data quality and identifying success criteria. The criteria should be clearly defined and measurable to decide whether the models developed are good or not. In our case, these parameters are the accuracy, reliability, and repeatability of conventional laboratory measurements. In addition to continuous tracking of numbers, it is essential to liaise with the parties designated by the company (e.g., chemical engineers, laboratory development engineers, technicians). For industrial applications, the ML Canvas framework recommends helping define the limitations and application requirements (robustness, scalability). A critical issue in the design of ML models is the quality of the data and the statistical evaluation of the data collected.

- **Data Acquisition and Understanding**
  The development of ML models begins with understanding business processes and issues to be solved. The next phase is followed by a detailed exploration of the data sets and examining the data quality. At the end of the section, it can be determined whether the data research project is feasible or not. If you want a good understanding of the business problem, use an Ishikawa chart that lists the factors that influence the goal and their other influencing factors. At this stage, even the success criteria of the models defined and measurable key performance indicators (KPIs) defined. Each research topic is determined by process control or laboratory quality assurance engineers at each step. ML Canvas supports the forecasting and learning parts of the ML application. In addition, each business site imposes restrictions on model compliance and application boundary conditions. ML Canvas offers the opportunity to outline the solution imagined by ML on a transparent map. The map outlined helps us see what is needed to implement it. In addition, team members provide information to see what else is needed for a successful ML project [132]. Part of the second phase of the CRISP-ML process assumes data sources, data cleaning, and building an environment. In this phase, its main task is to prepare the data for the ML

models. The second section also covers service design and data standardization, and appropriate data quality requirements [115]. In the next phase, its main task is to prepare the data for the ML models.

- **Model Development**
  The third phase is the ML model development of CRISP-ML, this is the very iterative process. Occasionally, we may need to review business objectives, define other KPIs, and modify the results of the ML model using available engineering from the available data. In the final phase, the ML workflow is packaged into a process to create repeatable modeling. The modeling phase follows the model evaluation phase, in which the performance of the trained model evaluates on a test data set. In addition, the robustness of the models should be tested on noisy or poor input data. After testing, a requirement level should be formulate against which ML methods can apply. In the final phase, before installing the models, the algorithms must meet a success criterion in which ML experts must evaluate the performance [133]. All settings and results for the modeling and evaluation phases should create a detailed document. The introduction of ML models means integrating models into a software system. For example, deploying ML models means that the predictive function is packaged as an interactive dashboard, as a predictive forecast, as a component of the ML model snap-in, into a kernel software architecture, or as a web service endpoint in a distributed system. The implementation of the ML model includes the following tasks: determination of hardware inference evaluation of the model in a live environment. In addition, provide online testing, such as A/B tests, and statistics test, user acceptance and usability testing, and, in extreme cases, plan for model downtime to gradually introduce a new model. Once the ML model is in production, continuous monitoring and maintenance of its performance are essential. A good solution for this is to display the indicators of ML models on a dashboard [133], [134]. For example, a depleted model where the main risk realized is the effect of "model obsolescence" when the performance of the ML model decreases when it begins to operate on samples of unseen production parameters or data from measurements of exceptionally rocks.

- **Model Deployment**
  The next phase is the commissioning of ML models in production. The complexity, size, and complexity of ML models depend on the business problem

to be solved [135]. The fourth phase is strongly related to those in front of it, which provides continuous feedback. At this stage, it is essential to select and enter the ML model. One of the main challenges for ML projects is reproducibility and robustness. Therefore, it is crucial to store all metadata related to the data (instrument, measurement setting parameters, environmental conditions, date) and the exact settings of the models (e.g. pre-processing, training, validation data set division, hyper-parameters, model, structure). All information about the deployed models should be stored using the predictive model markup language (PMML) as well as the machine learning model operationalization management (MLOps) methodology [136].

- **Model Operations**

  The final modeling phase is the maintenance of installed and continuously running models. In this phase, the available models must be continuously accessed through intelligent applications, and the data must be displayed continuously, for example, visualization on a dashboard. The use of MLOps is constructive in the third and fourth phases. MLOps is based on hands-on experience designed to monitor the efficient and reliable operation and maintenance in a live environment of the ML models. Cloud infrastructure services provide significant amounts of computing power at a relatively low cost. A significant advantage is that multiple users can share codes and capacities simultaneously. According to the methodology, the models are tested and developed in an isolated experimental system when the model is ready for deployment before being simulated sharply by data scientists and ML engineers to migrate the system. The daily application of ML models is a significant challenge for their application in industrial environments [137]. MLOps and compound of development and operations (DevOps) are very similar in their efforts to automate and improve production models while meeting standards and requirements. MLOps cover the entire modeling lifecycle, including diagnostics, fine-tuning deployments, and monitoring business metrics [136]. The use of MLOps assists in the installation and automation of ML models, the reproducibility of forecasting, the diagnostics and scalability of models, and the monitoring and, if necessary, management of their interaction. Saved and documented information increases the efficiency, transparency, and explainability of the reproducibility of ML models. One way to do this is to use the "Model Cards Toolkit". In addition, ML

models are increasingly used to perform highly complex tasks. The performance of the models aided by the version number of the packages used, and detailed documentation helps to understand the task. One way to do this is to create different model cards to help with the structured documentation of the models [138].

The best practice to prevent model performance degradation is to perform the observation task when performance evaluation of the models continuously to determine if retraining is required. Moving models from a monitoring task can lead to updating the ML model. In addition to tracking and retraining, tracking business processes and reflecting on ML models can help determine the mineral composition of oil fields more accurately [139] and make production plants more cost-effective and stable to produce a better product [140]. The Appendix A.7 briefly summarize the phases of the CRISP-ML methodology in a table.

6.3.2 Concept of cloud and edge based software sensor development

The CRISP-ML methodology presented in the previous section requires the development of an appropriate architecture that, in addition to the above, ensures the continuous availability of the models on-site and the secure and continuous data collection. The external elements of the architecture presented in this chapter are edge- and cloud computing solutions. Cloud infrastructure services provide significant amounts of computing power at relatively low cost. In addition, virtual services are available at a pre-determined hourly rate in these services so that we can pay as much for the service as before. A significant advantage is that multiple users can share codes and capacities at the same time. The cloud computing and MLOps greatly facilitate the development, monitoring, and subsequent operation of ML models. Our concept is essential for storing laboratory data in the cloud and for the joint handling of data related to the manufacturing process, such as temperature, pressure, analytical measurements. Data is transferred from laboratories using various edge computing devices and from production using IoT. The data analysis thus collected can provide rapid support in product quality using the results of ML algorithms and the condition of the machines involved in production. Furthermore, data transmission and models should work seamlessly in terms

of data availability. The architecture related to the concept is illustrated in Figure 6.5. The figure shows that the relevant architecture consists of two main parts (factory, cloud) and three sub-parts (laboratory, reporting, development). The main parts of the environment are defined by the factory process tracking and intervention, by the laboratory data collection and model running on-prem environment, while building the data pipeline, algorithms development, ML services, model monitoring and reporting do online.

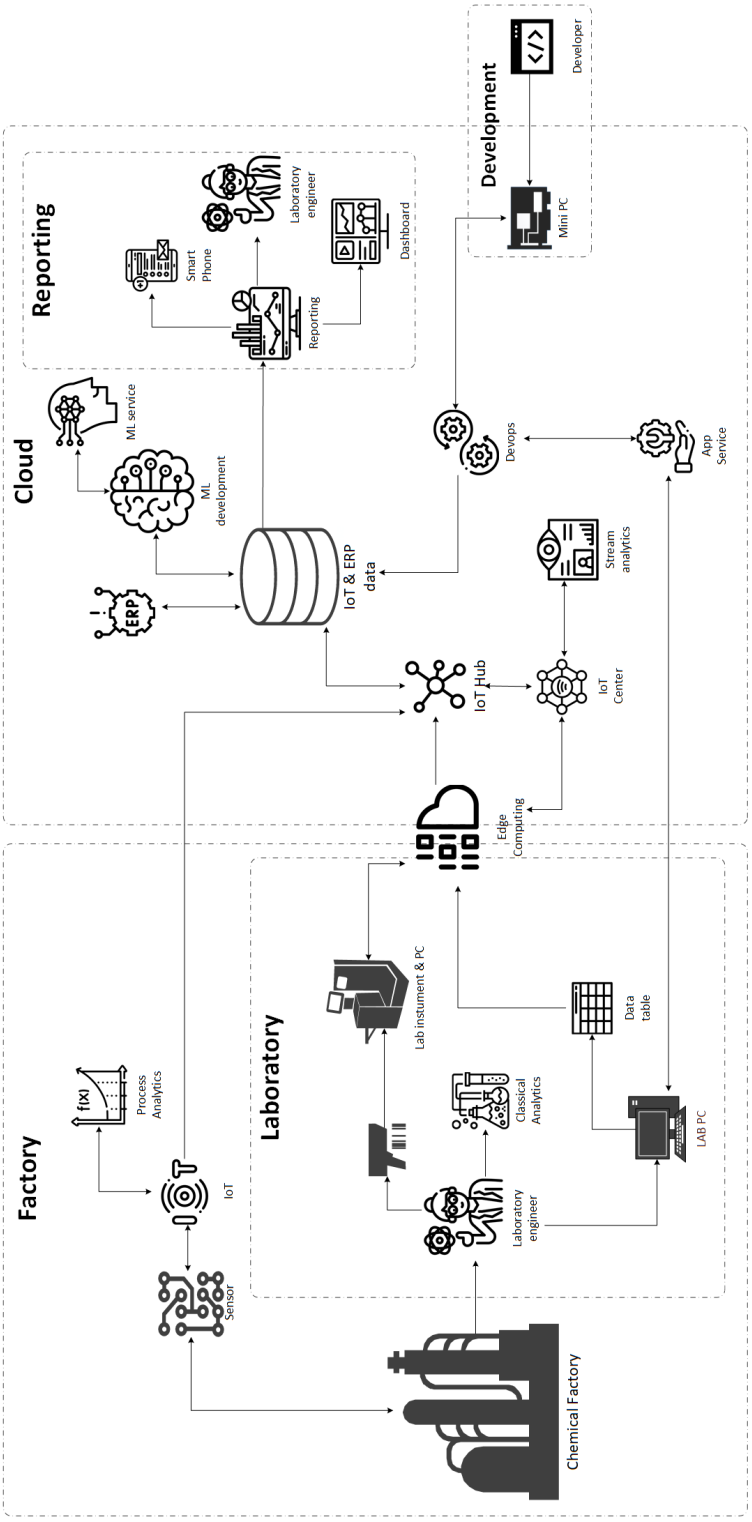


FIGURE 6.5: Architecture supporting measurements of the quality control in chemical processes. The dashed lines indicate the boundaries of the two main parts (factory, cloud) and three sub-parts (laboratory, reporting, development), the edge computing device connects the cloud and onsite area.

- **Process tracking and intervention**
  Process control colleagues constantly monitor industrial sensors with various software that connects to IoT devices via a LAN cable. Process engineers monitor various parameters such as temperature, pressure, material flow rate. From these parameters, the best conclusions can be drawn about the products' goodness. They can also get accurate quality results by predicting ML models of laboratory equipment. The samples of the process are transported to the laboratory, where colleagues prepare the samples and perform measurements using classical or rapid innovative measurement techniques.

- **Data collection and model running on the edge**
  The results of the classical measurements are manually uploaded to the enterprise system. Data entry for rapid measurements is done with a QR code reader for easier, faster and simpler use. The computing devices in the field are connected to the edge device with a LAN cable, which transmits the data to the cloud. On lab computers, colleagues are able to run ML models developed in the cloud and tested on a mini computer. As the figure shows, the critical part of the architecture is the edge computer. This device establishes a connection between the factory and the cloud service to be real-time and continuous data transfer.

- **Machine learning model building and development in cloud**
  Another critical part of the architecture is the IoT and ERP data market, where data engineers carefully compile data from different sources, which data researchers will then process. ML models are being developed in a cloud environment, moving into cutting-edge computing through data flow analysis and the IoT center. Maintenance of models and continuous monitoring of their performance is critical. It is essential for the production unit in the field always to have the best models available. Maintenance of models and constant monitoring of their performance is vital. It is necessary for laboratories always to have the best models available. By validating laboratory measurements and ML models, robust and efficient models can be developed that must be monitored continuously and intervened when warranted. Testing new, better models before the live operation for continuous model development is essential. It is imperative to separate these tests from the existing system completely.

- **Machine learning model testing**
  The new models are tested through a virtual unit, simulated as if sharp samples were running. In all cases, experts in data science and the business process should perform this activity with due care. Then, when the models have proven to be suitable, they can deploy the new ones on the edge device with an update. The great strength of the architecture is the continuous development and application of ML models, we can teach and update models every minute.

- **Reporting and quality control**
  Applications in Industry 4.0 solutions allow continuous evaluation and real-time monitoring of results. Reporting professionals can easily track the results of a plethora of lab samples on a dashboard, even on a smartphone. In addition, the dashboards are easy to customize and provide users with live data at any time.

Continuous data collection aims to make the most efficient use of data from industrial units to monitor processes. For example, the intermediate component of the oil fields or the different element content of the product is essential. In the Figure 6.6, the layers show the different levels of data processing. The first level is the secure collection and transmission of data. After collecting the laboratory data, the second level is to clean the data and prepare the fundamental analyses and reports. The fourth level is aggregation, which begins with communication between machines and then includes data integration and aggregation forecasting. Finally, the level of analysis begins with predictive analysis, then with ML, and finally with AI. The data from the IoT or edge device units send on a pyramid, and the point is that the measured raw data is under AI control.

6.3.3 Secure data collection and running on the edge device

An essential aspect of the development project is to make the developed models available for production and certification even if something goes wrong between the cloud and the terrain. If we have some issue with edge computing, troubleshooting is also easier. Edge computational analysis and knowledge generation occurs at or near the source of data and computational performance, away from centralized points toward the edges of the network. Edge computing should emphasize that

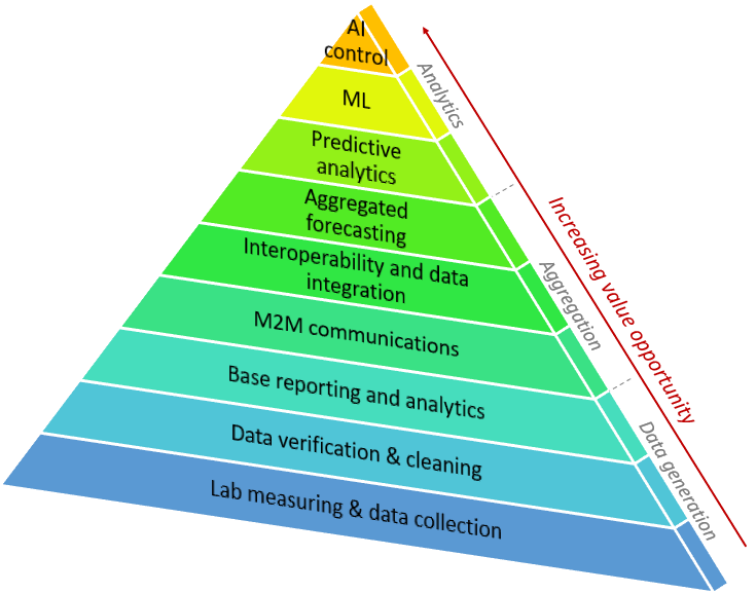


FIGURE 6.6: Level of the data processing. The layers show the integration of data into the corporate control system. The higher the level of the pyramid, the more complex the data-based processes.

this model does not rely on data centers but gets ready-packed models developed in the cloud. Edge computing is a distributed computing platform that brings computing and data storage closer to shortening response times and minimizing potential distance challenges and problems. As a result, it increases the speed and efficiency of responding to information. This computing platform is similar to a cloud-based platform, only closer to applications. Edge computing analyzes some data from IoT devices on the edge of the local network and transfers them to the cloud. In the technique we have developed, laboratory information management system (LIMS) and ERP data must be available on the edge device in addition to the measurement results. Therefore, selecting the optimal edge device in the market is crucial. Many manufacturers produce a variety of sharps, the parameters of which can vary significantly. The edge device of our choice is a mini personal computer. An essential aspect of the research was that the device could be used in extreme field conditions (the temperature varies between -40 °C and 85 °C), not just in the laboratory. The carefully selected edge tool securely transmits the collected data to the cloud and stores and runs the models packaged after the appropriate command.

A possible solution to eliminate possible attacks is to use block-chain technology. The technology offers a suitable capability for secure data transfer and ML model

deployment to IoT and edge devices [141]. However, there are other secure solutions besides or with block-chains.

6.3.4 Implementation of software sensor and machine learning model monitoring

Once models are developed, their maintenance is critical because they can become obsolete over time and their performance decreases compared to their development. Therefore, to always have a suitable model available in the field, we monitor the accuracy of the models and measured performance (Figure 6.4, *Deployment, Operations*).

The PMML is an XML-based specification for the representation of statistical and data mining models [142]. This can be used in the CRISP-ML approach that makes appropriate ML models available for quality assurance, helping the development, deployment and operation of ML models (Figure 6.4, *Development, Deployment, Operations*) [143]. ML model version numbers, settings, data dictionary and conversion, developer information, licenses, package release numbers are all built-in. PMML is an accessible markup language created for ML models. PMML is similar to HTML, but it is the hypertext markup language for web pages. PMML is an XML derivative developed specifically by the developers of the Data Mining Group (DMG) consortium to provide statistical and data mining for sharing between software and programs [142]. The great advantage of PMML is that it is vendor-neutral and conforms to any standard that is widely accepted and easy to use as a markup language for enterprise databases [144]. This reduces the potential for conflict and an open-ended platform that allows ML models to be developed and deployed. PMML is an open access de-facto standard for storing and exchanging predictive models [145], such as cluster models, regression models, trees, or supporting vector machines. In addition, development and deployment are separate, allowing data scientists and software professionals to develop models separately and quickly (Figure 6.7). With the power of a markup language, you can decide in minutes whether or not a model can be put into service for years. With PMML, models can be easily logged and consist of main components: header, data dictionary, data transformations, model. Of course, the pre-processing and post-model post-processing steps can also be stored before the models, and the model explanation allows performance to be evaluated. The PMML represents

not only a wide range of statistical techniques, but also the data transformations needed to turn input data and raw data into meaningful functions [146].

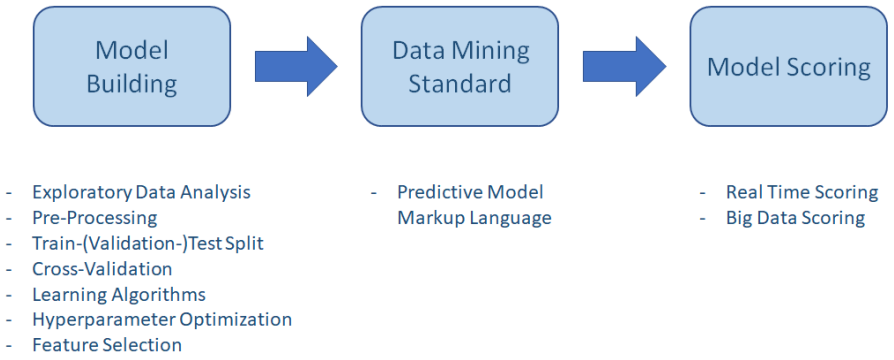


FIGURE 6.7: Predictive model markup language based data mining activity. The three main sections show the main stages in the development of the models.

The performance of the models can be measured by various tools such as lean six sigma (LSS) and statistical process control (SPC) [147]. Improving the efficiency of processes is essential for environmental and economic reasons. The increase in efficiency is due to the combined effect of the LSS principle, and the ML algorithms [148]. Six sigma can be used to measure products quality and ML model performance. Since the accuracy of a ML algorithm can be quantified, the goal is continuous improvement. The goal of the models is to reach the accuracy of six sigma, so we can reduce mistake product volume, which will increase revenue. It is essential to mention that all these findings also play an essential role in developing the models. The continuous data collection, models' re-learning, and algorithms' experiences contribute to achieving the best predictive results.

These three metrics are key indicators of each laboratory measurement that the standards are provided. These numbers also affect the goodness of the models, as the reference data are with these numbers [147]. An essential tool in enterprise quality management is SPC [149]. It can effectively and verifiably distinguish abnormal fluctuations in product quality. Therefore, intelligent and efficient SPC is of great importance to factories, especially Industry 4.0 [150]. The property of SPC is that it focuses on histogram pattern recognition and can mathematically support the detection of manufacturing differences [150]. Different pipelines can be used to easily track the performance of the SPC models [151]. The continuous integration/continuous delivery (CI/CD) process introduces monitoring and automation to improve the application development process, especially during the integration and testing phase, further during shipping and installation. The CI/CD

is a methodology in software development that combines continuous integration with continuous delivery. The added value of CI/CD pipelines is achieved through automation, but it is even possible to perform each CI/CD process step manually [152]. The CI/CD automation keeps the deployed ML models up to date without causing disruptions to production (Figure 6.4, *Deployment*) [153].

The main elements of the proposed framework are: following the CRISP-ML methodology, applying it to the developed and validated ML models using MLOps, PMML for model tracking and archiving, CI/CD pipeline for easier use of the models. Select the appropriate cloud service and edge device for the required devices, considering computing needs and connectivity options. Choose the right reporting tool if it has the option of even a smartphone-compatible dashboard service.

6.4 Case study

This section presents a study that provides an opportunity for complex companies to predict difficult-to-measure and critical parameters. During the development, the possible deterioration of the quality of the models should be monitored, in which the CRISP-ML approach can help. This section describes the reason for the development (Section 6.4.1), the technology & the tasks encountered (Section 6.4.2), method implementation (Section 6.4.3), the ML models used (Section 6.4.4) and lessons learned (Section 6.4.5) by this case study.

6.4.1 Background

In addition to the production of motor fuels, the production plants of the integrated oil companies also produce lubricating greases. Therefore, the product range of the bread material production unit is very diverse. Sourcing requirements and standards determine the exact product mix. In the case of ML algorithms, it is essential to emphasize that the number of models is determined by the number of products and their parameters. Therefore, the development and maintenance of ML models are essential for companies. The best version of the models should always be available on-site. The wide range of products poses a severe challenge to the continuous presence of the best models. Without CRISP-ML, MLOps and

PMML there would be plenty of untraceable models that could not be operated in the long run. The company has a data team responsible for moving data, developing models, maintaining and reporting. Measuring the penetration and metal content parameters of lubricants and greases under operating conditions has so far proved impossible. However, ML models built on laboratory measurements have proven that this can be done with software sensors installed in the right place in plants. On-site deployment of live computing tools and cloud computing is essential for developing quality assurance models.

6.4.2 Technology & task

The development goal is to create a unique application that can automate the work in the laboratory and help day-to-day activity in the laboratory colleagues. Furthermore, another goal is to verify and collect laboratory data and production data of the process. Continuous monitoring of difficult-to-measure parameters with software sensor lines provides our plants with accurate material flow quality information or our well analysis of drilling samples. Furthermore on the well samples can we use for this methodology prediction the mineral composition.

Reducing the response time of laboratories and measurements using less hazardous substances are of paramount importance in laboratory developments. Our goal is to get the most information out of a lab sample and do it all in the fastest way possible. Fast and non-destructive measurements include various spectroscopic measurements such as Infrared (IR) -, Raman Spectroscopy, X-ray, Gas-Chromatography. The essence of these measurements is that the device makes a curve from a small amount of material, which has much helpful information about the samples. Furthermore, the measurements do not require the use of hazardous substances. The measurement process can be automated. If the appropriate sample is prepared, the devices can be left alone until all the completed measurements have been completed. The measurements listed above provide different information about chemicals, so storing these measurements in a standard "data lake" is an essential part of laboratory development. The Industry 4.0 devices help to store measurement results in one place. For example, the edge computing or IoT sensors described above are essential for moving data. Laboratory measurements can easily connect to the corporate data, even with minute updates.

6.4.3 Framework implementation

An essential aspect in the construction of models is the quality of parameter on which the model can be built upon. In addition, an important consideration is where and how a given parameter can predict. So the models for laboratory measurements help the installation of software sensors for operational and even drilling intelligent sensors. The first phase of the CRISP-ML methodology Figure 6.4, *Business Understanding* business task is to understand that the estimation of nitrogen from the operating parameters and the quartz content from the drilling rock samples gives great potential for estimating ML models. The success criterion of nitrogen model estimating was determined by the reproducibility value of the classical measurement in the quartz model, although the degree of error of the model and the speed associated with the estimation. The developed model meets the first phase of the CRISP-ML criteria in both cases. In the second phase, in understanding the data, an important test was whether, in both cases, the traditional measurement could replace by a fast, non-destructive model, and the models built in this way would be a good starting point for the installation of later software sensors. The data understanding phase (Figure 6.4, *Data Understanding*), what measurement data and what errors do we have in our measurements (reproducibility, repeatability). Data sources in both cases were the edge tool and ERP and LIMS, respectively. During the modeling, we used particular train test splitting for both target variables, which can monitor the data distribution from the two data sets. The distribution of the train and test data sets with the application was similar. We used ten-fold cross-validation (10-cv) to develop the models and PMML to deploy the models. Colleagues can track the results and accuracy of deployed models using a visualisation tool, a personal computer application, a web browser, or even a smartphone.

Newer and newer measurements from the edge device must review through validation (Figure 6.5, *Edge computing*). ML maintenance shows whether the sample is worth incorporating into the model or not. In addition, newer and newer samples help track the performance of models currently in service (Figure 6.5, *Reporting*). The built models must be able to handle such changes, so the models are maintained, and the data is displayed through an application (Figure 6.5, *Reporting*). The model development steps for a parameter of material flow are shown in the figure below (Figure 6.8).

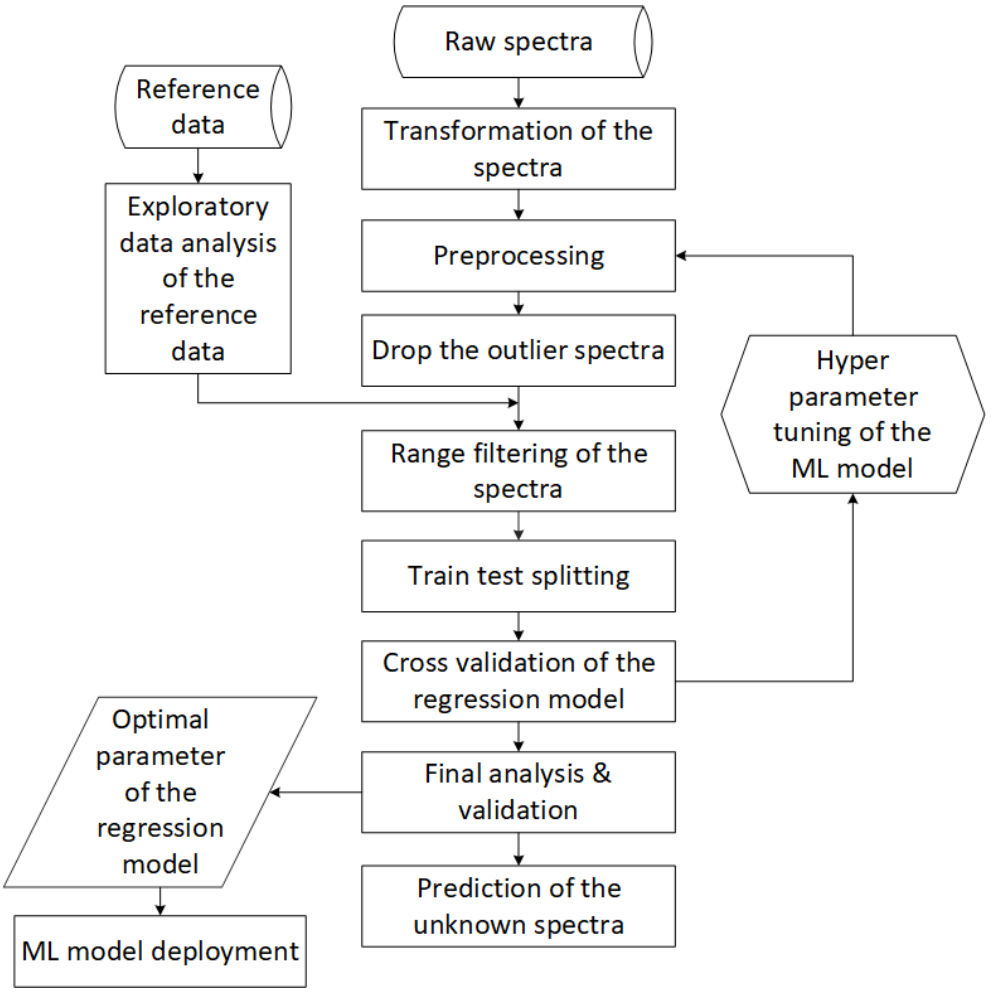


FIGURE 6.8: Major stages in the development of ML models. The parts are considered general laboratory, quality assurance and industry independent.

The main parts of the development of ML models outlining the simplified steps of data processing and modeling are exploratory data analysis (EDA), pre-processing, outlier detection, train & test splitting, with a special technique that considers the distribution of the target variable. Then, the iteration process shows the fine-tuning of the model parameters, and finally, the low-error models with the appropriate settings are deployed. This process must be set separately for each parameter (nitrogen, quartz content etc.) in each family of laboratory samples. Laboratory results from measurements can often not be used directly for interpretation or modeling. It must be tied to some calibration to understand business, or in many cases, some mathematical technique must be used (Figure 6.4, *Business Understanding*), in all cases involving the business colleagues. To determine what influences specific parameters the most, we use the Ishikawa diagram mentioned in the previous section 6.3.1) (Figure 6.4, *Data Acquisition*), which shows the target variable and the factors and sub-factors that most influence it. Following the

CRISP-ML methodology, this figure is constantly expanding. Therefore, the role of each factor in the design of the models should be examined. If the accuracy of the model can be easily affected by these factors, the model must be prepared to solve these challenges with robustness (Figure 6.9).

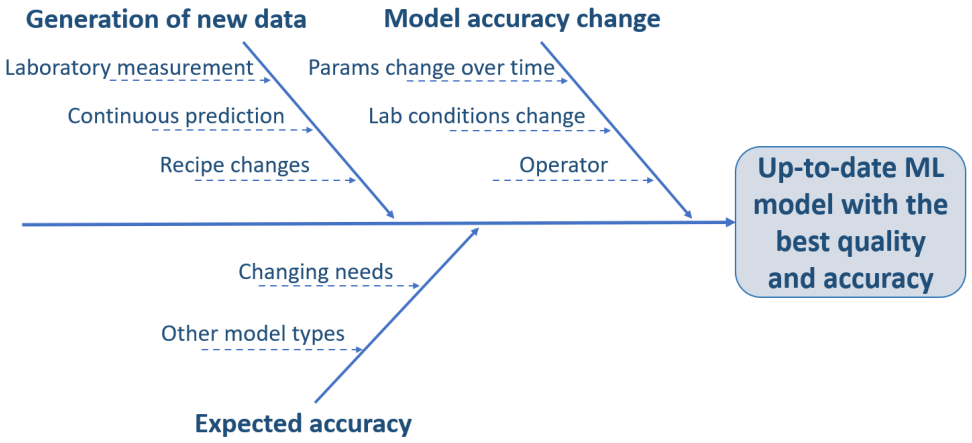


FIGURE 6.9: Isikawa diagram related to the development of ML models. The accuracy and applicability of the ML model can influenced by the three main factors, which are affected by two or three things.

The distribution of the modeling data sets of the ML models constructed in the two laboratories presented in the case study is illustrated in Figure 6.10. The x-axis of the figure shows the given property to be measured as a percentage, and the y-axis shows the density. The quartz content in the upstream laboratory and the nitrogen content in the lubricant laboratories are measured. The distribution of quartz data is much more favourable for modeling than the nitrogen content. It can be explained by the fact that the variability of the nitrogen content during stable operations is much smaller than the quartz content of the rock sample from several oil fields Figure 6.4, *Business and Data Understanding*. Tuning the models and testing their robustness for variables with a high skewness (>3) value is paramount. In addition to calibration samples, other samples should be included in the model, such as products manufactured under extreme manufacturing conditions or products of poor quality produced under laboratory conditions.

The quartz content is based on the X-ray diffraction measurement, and the nitrogen content is the target variable from the Kjeldahl measurement method. In both cases, the FTIR spectra give the predictor data set of the model. The ML models are validated with ten-fold cross-validation.

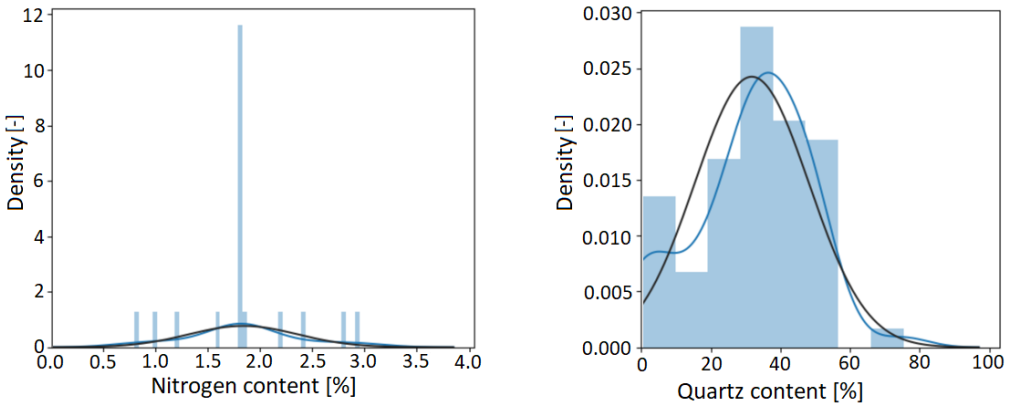


FIGURE 6.10: Histogram of the target variables. The left diagram shows the nitrogen content related to lubricants, and the right diagram represents the quartz content of the Exploration and Production laboratory. The black curves show the normal distribution of the given modeling data set, and the blue is the actual distribution.

6.4.4 Evaluation and type of ML models

During the real-time operation of ML models, it is essential to continuously check the accuracy of the models to determine when a particular model is considered excellent, good, or unsuitable. When evaluating the models, the three "R" index of the classical measurements must be considered [154]. In general, a model is considered adequate if the prediction accuracy of the new samples is within the reproducibility limit. Models that exceed the reproducibility value of conventional measurements are considered unsuitable (Figure 6.4, *Business Understanding*). Monitoring models allows them to be ranked based on percentage overshoot. The monitoring system displays the models with the most significant errors at the beginning of the ranking, in which case the intervention is urgent (Figure 6.4, *Deployment of the Model*). Model KPIs are similar to different metrics in traditional laboratory measurement techniques. For decades, we have used circular measurements of various standards to validate devices periodically. Therefore, the calculations are very similar when using the indicators of the ML models. Correlation coefficient (R^2), root-mean-square deviation ($RMSE$), and relative percent differences (RPD) are important indicators for tracking ML models. In this study, we compete with three different algorithms to estimate the given parameter with the best algorithm. A special linear regression is complemented by a particular calculation that can also handle non-linearity problems. The other two tree-based algorithms are prevalent random forest and extreme gradient boosting. A vital consideration in the selection was to choose an algorithm that would qualify the

samples. During the measurements, the ML algorithms must be robust, not sensitive to outlier samples (Figure 6.8, *EDA*), and the methodology of the competing algorithms is different. The three algorithms must be optimized and tested for each target variable, and then the best of the three is implemented on the edge tool. For installed models, the model type may have changed during development.

A brief theoretical overview of the three model types selected is provided below. The partial least squares regression (PLSR) model is possible to let the score matrix represent the data matrix. A simplified model would consist of a regression between the scores for the $\mathbf{X}$ and $\mathbf{Y}$ block [77].

$$\mathbf{X} = \mathbf{S}_X \mathbf{L}'_X + \mathbf{E}_X \tag{6.1}$$

One can build the outer relation for the $\mathbf{Y}$ block in the same way:

$$\mathbf{Y} = \mathbf{S}_Y \mathbf{L}'_Y + \mathbf{E}_Y \,, \tag{6.2}$$

where $\mathbf{S}$ is the score, $\mathbf{L}$ is loading matrix and $\mathbf{E}$ represents errors.

The partial least squares has been a gaining popularity as a multivariate data analysis tool due to it is ability to cater for noisy, co-linear and incomplete datasets. The PLSR was supplemented by a nonlinear iterative partial least squares (NIPALS) algorithm supplemented by a non-linear iterative calculation, based on a recursive computation of co-variance matrices and gradient-based techniques to compute eigenvectors of the relevant matrices [155].

Random Forest is a tree based algorithm that combines the outputs of multiple decision trees to create the final output. The term "random" is because this algorithm is a forest of randomly generated decision trees. The simpler decision tree algorithm was not chosen because it has a significant drawback that causes over-matching, which can be limited in implementing random forest regression (RFR). Another significant advantage is that the Random Forest algorithm can be very fast and robust compared to other algorithms.

The following formula shows how to calculate the RFR:

$$F(x_t) = \frac{1}{B} \sum_{i=0}^{B} F_i(x_t) \tag{6.3}$$

Where:

- x_t = test samples
- B = Time for random sampling with replacement from the original data. This sample functions as the training set for growing the tree.
- F_i = It is a function of each decision tree, each tree is grown as much as possible without pruning.
- F = Outputs function, in the case of a regression problem, we take the average of the predictions for each tree.

The extreme gradient boosting (XGBoost) is a popular algorithm for gradient-increased trees. The method of the algorithm tries to accurately predict the desired target variable by combining estimates from simpler, weaker models. XGBoost minimizes the regularized (L1 and L2) objective function, which combines a convex loss function (the difference between predicted and target outputs) and a penalty term for the complexity of the model. The training is done iteratively by adding new trees, which predicts the remnants or defects of the previous trees, which are then combined with the previous trees to make the final forecast. In addition to using a unique method to build and prune trees, it also has custom optimization. It is an excellent advantage as it makes computing faster on substantial data sets.

$$S = \frac{\sum_{i=1}^{n} R_i^2}{\sum_{i=1}^{n} [PP_i(1 - PP_i)] + \lambda} \tag{6.4}$$

Where:

- S = Similarity Score
- R_i = Residual is a different of actual value and between predicted value (observed value - predicted value)
- PP = Previous probability is the probability of an event calculated at a previous step. The initial probability is assumed to be 0.5 for every observation, which is used to build the first tree. For any subsequent trees, the previous probability is recalculated based on initial prediction and predictions from all prior trees.

- λ = Lambda is a regularization parameter. Increasing it reduces the effect on the leaves with little observation, while many observations have little effect on the leaves.

An essential element in the development of robust models is the examination of the sensitivity of the models. Sensitivity analyses evaluate changes in system inputs and the individual effects of each variable on the output and provide information about the different impacts of each variable tested. In addition, it is essential to produce a sufficient number of samples and rare samples to install good models. Extreme samples can be prepared by the design of the experiment (DoE) for the latter process, these samples help to achieve the robustness of the models. During development, we calculated the accuracy of the models for each laboratory property for validation and test data set. The models were optimized so that KPIs did not differ significantly in training-, validation - and test data set, thus protecting the models from over-fitting.

The following two tables summarize the accuracy of the ML models built on the two tested properties. It is important to note that the pretreatment of the spectra before the three model types was the same for both properties (Table 6.1 and 6.2). The '10-cv' ten-fold cross-validation results are represented by the 'perf.' metric that represents the performance of the model on samples not used in the teaching of the models. From the results presented in these two tables, it can be concluded that XGBoost is overfitted and performs the worst despite hyper-parameter tuning. The PLSR shows a balanced average performance, yet the RFR is the best-tuned ML model out of the three models. These model results show that we can discuss the two important parameters included in the study with ML models. By applying the models, we can determine specific key parameters much faster, with which we are already able to reduce the load and response time of the laboratory significantly. Furthermore, after testing the developed models, the installation of factory software sensors can be solved with the involvement of factory technologists. In the case of lubricants, the development provides support for where to install sensors, while in the case of upstream wells, software sensors can be allowed in the wells. The parameters required by the plant are to reduce overall equipment effectiveness (OEE) during lubricant production and to find the proper reservoir for upstream drilling. With the help of the models, scrap products are reduced during the production of lubricants, and in the case of quartz models, we get a more accurate picture of the geological formations.

Table 6.1: Results of the ten-fold cross-validation (10-cv) and results of the performance dataset (perf.) nitrogen content of ML models.

Nitrogen content	*RMSE*		R^2		*RPD*	
	10-cv	perf.	10-cv	perf.	10-cv	perf.
PLSR	0.010	0.035	0.999	0.975	57.73	6.36
RFR	0.089	0.084	0.972	0.929	5.98	3.77
XGBoost	0.005	0.112	0.999	0.747	31.62	1.98

Table 6.2: Results of the ten-fold cross-validation (10-cv) and results of the performance dataset (perf.) the quartz content of ML models.

Quartz content	*RMSE*		R^2		*RPD*	
	10-cv	perf.	10-cv	perf.	10-cv	perf.
PLSR	2.032	2.407	0.900	0.731	3.165	1.930
RFR	1.434	4.671	0.621	0.937	1.625	4.010
XGBoost	1.966	4.660	0.913	0.870	3.406	2.779

The models were currently available to laboratories monitored through reporting and web application. For the samples examined, there are different ranges at which the system indicates the difference between the prediction and the classical measurement. After the ten indications, the web application automatically indicates the validation required for the ML model. Then, the data scientist colleagues review the poorly predicted samples and develop the model if they deem it.

6.4.5 Lessons learned

The advanced analytical models on the production and research laboratories can quickly measure many more samples. The architecture presented in the previous and the models developed can reduce laboratory workload and facilitate measurements with lower health and safety executive (HSE) risk. Instead of classical measurements containing difficult-to-measure, hazardous materials, the accuracy of ML algorithms deployed on edge computing devices for different qualification properties can change significantly over time. This solution may cause changes in the production program, such as different raw materials or new geological rock samples not yet known by the model. The accuracy of the models may also be affected by the operating time of the devices, the degradation of the light sources, the relocation of the devices within the laboratory, or the extreme measurement

conditions of the measurement of the samples (e.g. human factor, temperature, humidity). Fortunately, the infrared measurement technique presented in the present study is less sensitive to measurement conditions and instrument ageing. However, changes in sample quality can easily affect the accuracy of models. Checking the accuracy of models should become a daily practice for manufacturing and research laboratory engineers. They can report to data scientists or model developers who can solve the problem quickly. After installing the system, monitoring and maintaining the models of the edge device is also essential. In addition, the tool is responsible for real-time data transfer and accessing the latest models onsite. The edge device selected in the study is the MOXA-8200, the configuration and operation of which posed a severe challenge during development. MOXA is an excellent tool for collecting data and managing a few models, but increasing the number of models results in severe limitations when using the device. The market for edge computing devices is changing very dynamically, so it is worth reviewing the devices you use from time to time. The tool tested in the case study was hired from your local support company, so it is easy to ensure that the best tool is always onsite.

The case study presented in this chapter can estimate difficult-to-measure, problematic parameters using different ML algorithms. The strength of the models developed is that the right ones are constantly available. Tracking and keeping models up-to-date is a challenge for research and manufacturing laboratories, with cloud and edge computing techniques providing a solution. They offer turnkey solutions for data transfer, design, model development and deployment. However, the two techniques present a severe opportunity and difficulty for the safe and continuous supply of industrial processes. Therefore, it is essential to ensure the real-time accuracy and availability of the models (Figure 6.5).

Applying the CRISP-ML methodology presented in this chapter significantly reduces the time required to collect, create, and develop data and deploy ML models. Experience has shown that the steps of the first models took a total of 150 working hours by three colleagues, a laboratory technician, a data scientist, and a technologist. Furthermore, introducing the first ML model took about 60 working hours from a data scientist and data engineer. Building a new average ML model from the beginning with CRISP-ML involves data mining, cleaning, outlier filtering, and creating a basic model of about two and a half hours. Testing and commissioning take one and a half hours. Finally, it takes another half hour to

evaluate and interpret the results of colleagues. The model is built and installed fully automatically using CRISP-ML. The development and implementation time of the new ML model is about 2% compared to the data understanding, the development, and implementation of the ML model, and the working time reduced to one-fiftieth alone guarantees a return.

6.5 Chapter summary

With the development of Industry 4.0 and the opportunities offered by digitalization, it is crucial to bring science and research closer and closer to production, and sensors play an essential role in this. Nowadays, software sensors are gaining more and more space, which can predict critical parameters that are difficult to measure in production processes. However, software sensors require the development of special ML algorithms that must be continuously monitored, operated, and maintained. The methodology outlined in the scientific study and the case study discussed in detail present a possible solution for the possibility of using software sensors. The introduction of ML models into production involves several nested components and processes. CRISP-ML is a systematic process model for ML software development that raises awareness of potential risks and emphasizes quality assurance to reduce these risks to ensure the success of the ML project. The CRISP-ML methodology consists of five parts of a sizeable cyclical process that helps build traditional research and development digitization PoC projects into a thriving, sustainable and long-term system. The main elements of the application of the CRISP-ML methodology are model development, continuous data cleaning, feature engineering, model validation, performance monitoring, data visualization. The other essential elements of this methodology are edge and cloud computing, which are needed for the continuous development of models, serial data transfer, and onsite access to the models. The ML models used in the two laboratory measurements presented in the case study are suitable for the use of software sensors. Furthermore, the architecture presented is related to the methodology using elements of edge- and cloud computing. The ML models presented in this chapter meet industry requirements and are suitable for estimating parameters. Our next goal is to build similar models to predict as many parameters as possible, which can help ensure quality assurance, better production.

Our future goal is to install software sensors for various process units using the framework to improve manufacturing processes further. The CRISP-ML methodology helps develop models consistently and systematically, and it is essential not to have to develop a separate model for each sensor. In the case of application and monitoring of the developed models, sensor replacements and maintenance can cause problems in the accuracy of the models, and the developed methodology must provide a solution for these (e.g. method and model transfer).

Chapter 7

Conclusions

This dissertation provides a thorough insight into the development of soft sensors, followed by an in-depth exploration of four key areas crucial for their application within Industry 4.0 advancements. Regarding soft sensors, methodologies generally fall into three categories: model-based (white-box) models, empirical (black-box) models, and hybrid (grey-box) models. I presented a comprehensive literature review on soft sensors, explaining their crucial role in Industry 4.0 advancements. Chapter 2 details their ideal applications, outlines their advantages and disadvantages, and discusses the challenges inherent in their deployment. Additionally, it examines the diverse applications of soft sensors across various industries.

A significant portion of this work focuses on black-box models. However, the balance error reconciliation technique employed in the hierarchical time series analysis represents a hybrid approach. The detailed methodology presented a novel approach integrating data reconciliation with machine learning to address the challenge of satisfying hierarchical constraints in complex system models. Here, it may fit better to write that I investigated three distinct machine learning model development approaches: one without reconciliation, another using reconciled measurement data, and a third directly fine-tuning model predictions based on modeling errors. Through three diverse case studies, my research demonstrates that directly reconciling machine learning predictions significantly enhances accuracy and reliability while ensuring all hierarchical constraints are met. This method facilitates data integration from various measurement techniques, taking into account their inherent errors within the prediction process. Ultimately, this approach leads to

more accurate, flexible, scalable, and robust models. However, its effectiveness hinges on precise knowledge of the system's hierarchical structure, data quality, and defined constraints (Chapter 3).

Chapter 4 is about data reconciliation in hierarchical time series, my dissertation here I discuss another prominent area within soft sensors: data fusion. This research rigorously tested various data fusion (DF) techniques, including our novel Complex-Level Ensemble Fusion (CLF), built on analytical chemical spectra. CLF, which combines GA feature selection, PLS projection, and XGBoost stacking of MIR and Raman data, consistently outperformed single models and traditional DF methods, significantly reducing prediction errors for both industrial lubricant additives and RRUFF minerals. Beyond superior accuracy, CLF offers practical benefits like tool-chain compatibility, modular expandability, and full reproducibility. However, its effectiveness hinges on rigorous preprocessing and feature selection, and mid-level fusion provided no benefit. In essence, DF, especially CLF, significantly boosts ML model performance in industrial digitalization by leveraging complementary spectral data, although proper preprocessing remains critical. Future work will extend CLF to classification, real-time online soft sensors, and validation on larger, multi-site industrial datasets, establishing it as a valuable, transferable tool for enhanced quality control and geochemical screening.

Chapter 5 presents a robust method for artificial data generation, which is a crucial part of industrial 4.0 models. The generation of artificial infrared spectra of rocks demonstrates strong similarity to real measured spectra. The generated spectra successfully replicate solubility characteristics across multiple acids (HCl, HF-HCl, AcOH) and are systematically produced to fill gaps in sparsely sampled parameter spaces, allowing for targeted improvement of data distributions. This method effectively enriches the parameter space with synthetic samples that adhere to predefined constraints. The quality of the generated spectra was validated both statistically and visually. A key future direction involves leveraging the increased number of generated samples to enable the use of convolutional neural networks for further analysis.

Chapter 6, the development of the model and its life cycle framework for one branch of the application of soft sensors is presented. With the rise of Industry 4.0, integrating science and production through advanced sensors is vital. Software sensors, predicting difficult-to-measure parameters, are increasingly crucial but

demand continuous ML algorithm management. My study presents a methodology and case study for their effective implementation.

I propose CRISP-ML, a systematic framework that guides ML software development from concept to a sustainable system, mitigating risks and ensuring quality. This methodology leverages edge and cloud computing for continuous model development, data transfer, and on-site access. The presented ML models meet industrial requirements for parameter estimation, and future work aims to expand their application for enhanced quality assurance. The consistent development provided by CRISP-ML will also address challenges like sensor replacements, ensuring model accuracy through method and model transfer (Chapter 6).

In summary, this dissertation addresses four critical areas within the domain of soft sensors for Industry 4.0. Each research topic is comprehensively supported by detailed case studies, which empirically demonstrate the utility of the developed methodologies using both industrial and benchmark datasets.

Regarding future work, the primary objective is to integrate the four key steps outlined in this dissertation into a unified, process-centric system. This system would be deployed, for instance, on an industrial front-end or SCADA platform. The core functionality of this integrated framework would be its ability to operate autonomously and dynamically. A crucial component of this automation would involve the continuous generation of new data points using the artificial spectra methodology presented. This would specifically target areas of the parameter space where real-world data is sparse or entirely absent, ensuring the training data remains comprehensive and representative. Simultaneously, the framework would be capable of monitoring the performance of deployed machine learning models in real-time. If a model's accuracy were to fall below a predefined threshold, the system would automatically initiate a retraining process to update and optimize the model. Furthermore, the rules governing data reconciliation errors — for instance, the physical and chemical constraints — would be customizable through a user-friendly interface. This would allow operators to fine-tune the system's behavior in response to changing process conditions without requiring expert programming knowledge. Ultimately, the dissertation's overarching goal is to go beyond theoretical development by creating and demonstrating practical, industry-ready digital solutions that fully leverage the capabilities of Industry 4.0. By developing this comprehensive framework, we aim to transform disjointed research methods

into a cohesive, automated system that can be seamlessly implemented in industrial environments, thereby improving efficiency, reliability, and decision-making in real-world applications. This will establish a robust and transferable blueprint for the next generation of smart manufacturing and process control systems.

Chapter 8

Thesis findings

The following list contains my new scientific results in four thesis findings.

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 characterized by high predictive power, practicality, and robustness to measurement errors.**

   1.1 I developed a data reconciliation technique for hierarchical time series that improves the usability of machine learning models.

   1.2 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.

   1.3 Through the three case studies, I demonstrated the utility of the technique and its importance for application in complex systems.

   Related publications: [R1], [R2].

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.**

   2.1 I developed a data fusion technique that is capable of creating higher-performing machine learning models by fusing data from different measurement spectra.

2.2 Based on the two datasets I presented, the developed complex-level ensemble model demonstrated superior performance compared to other data fusion techniques.

2.3 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: [R3] [R4].

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.**

3.1 I have developed a technique for generating artificial spectra that can efficiently generate data of appropriate quality, especially for sparse or incomplete data sets.

3.2 I created the artificial infrared spectra using the developed methodology, PCA and ANN techniques.

3.3 I tested the authenticity of the artificial spectra in detail and presented the methodology for their validation.

Related publications: [R5], [R6].

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.**

4.1 I developed an industrial-grade framework that uses cloud and edge-based computing to predict laboratory data in near-real time.

4.2 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.

4.3 I tested the framework on an industrial network, and its accuracy was demonstrated through two case studies involving machine learning models.

Related publications: [R7], [R8], [R9].

Other publications that are tangentially related to the research topic

Related publications: [F1], [F2], [F3] [F4].

Appendix A

Appendices

A.1 Graphical summary for the spectrum generation

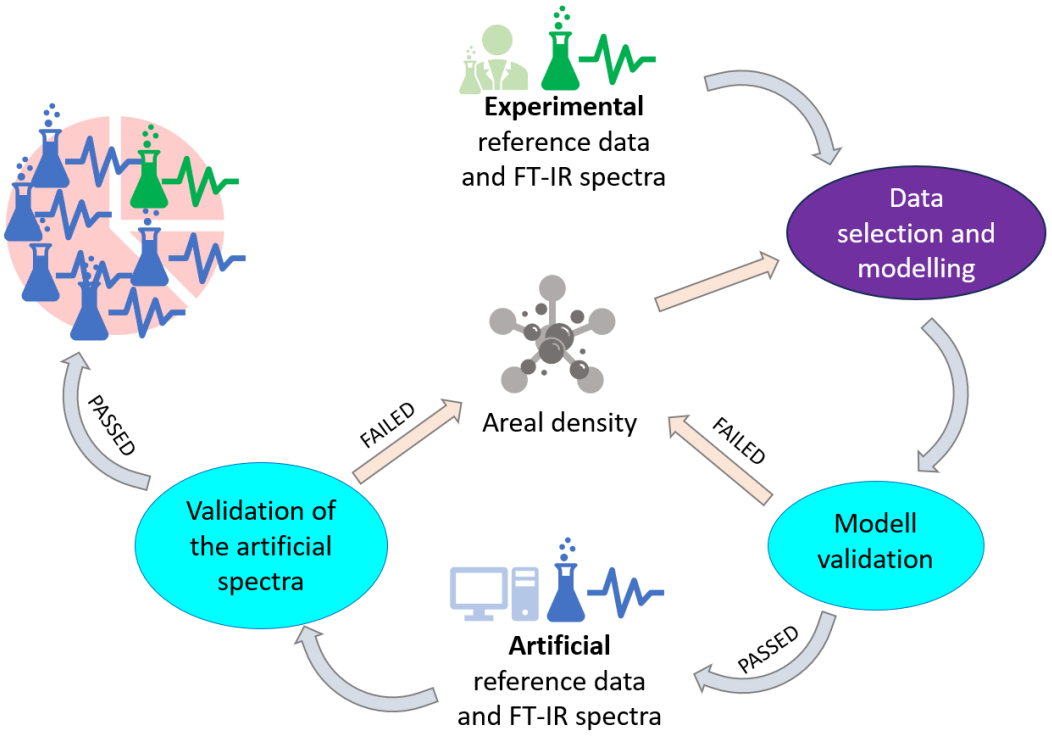


FIGURE A.1: Graphical summary for the generation of realistic infrared spectra using artificial neural networks

A.2 Comperision of Raman, MIR, NIR measuring

TABLE A.1: Comparison of Raman, MIR, NIR measurements [56]

	Raman	**MIR**	**NIR**
Wavenumber	50 - 4000 cm^{-1}	200 - 4000 cm^{-1}	4000 - 12500 cm^{-1}
Bonds	homonuclear bonds such as C-C, C=C, S-S	polar bonds such as C=O, C-O, C-F	H-containing bonds such as C-H, O-H, N-H, S-H
Absorption bands due to	scattered radiation	absorbed radiation (basic vibration)	absorbed radiation (overtones and combination)
Absorption	strong	weak	weak
Absorption bands	well-resolved, assignable to specific chemical groups	well-resolved, assignable to specific chemical groups	series of overlapping bands
Signal intensity	poor	good	good
Quantification	intensity (I) $\sim$ concentration	log $I_0/I \sim$ concentration (Lambert-Beer law)	log $I_0/I \sim$ concentration (Lambert-Beer law)
Excitation conditions	change of polarizability, α	change of diploe moment, μ	change of diploe moment, μ
Selectivity	high	high	low, requires calibration and chemometrics
Interference	broad fluorescence baseline	water	water, physical attributes (e.q., sample size, shape, and hardness)
Particle size	independent	dependent	dependent
Applicability for atline, online, inline	good	poor	good
Radiation source	monochromatic (laser VIS/NIR region)	polychromatic by globar tungsten	polychromatic by globar tungsten
Sample preparation	none	reduced (except ATR)	none

A.3 Visualization of principal components in the C path

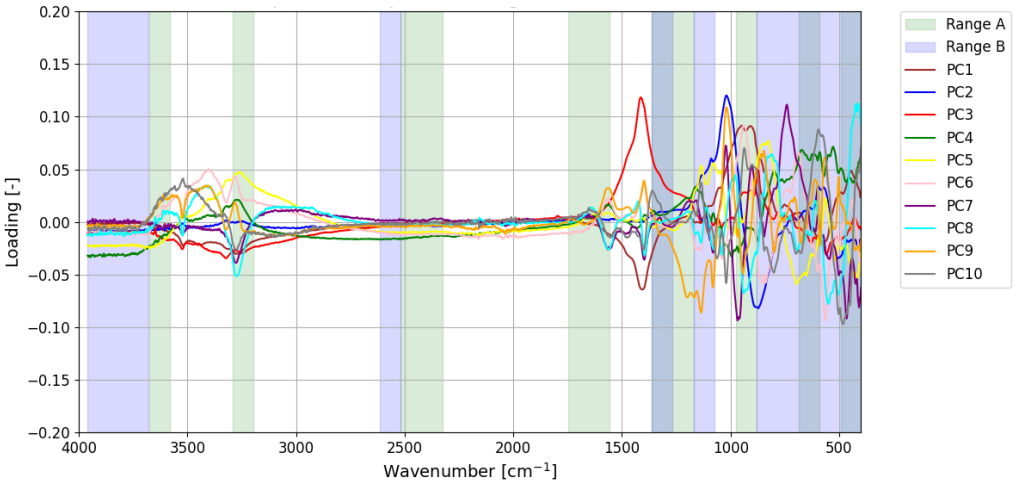


FIGURE A.2: Feature selection MIR A, B, C method

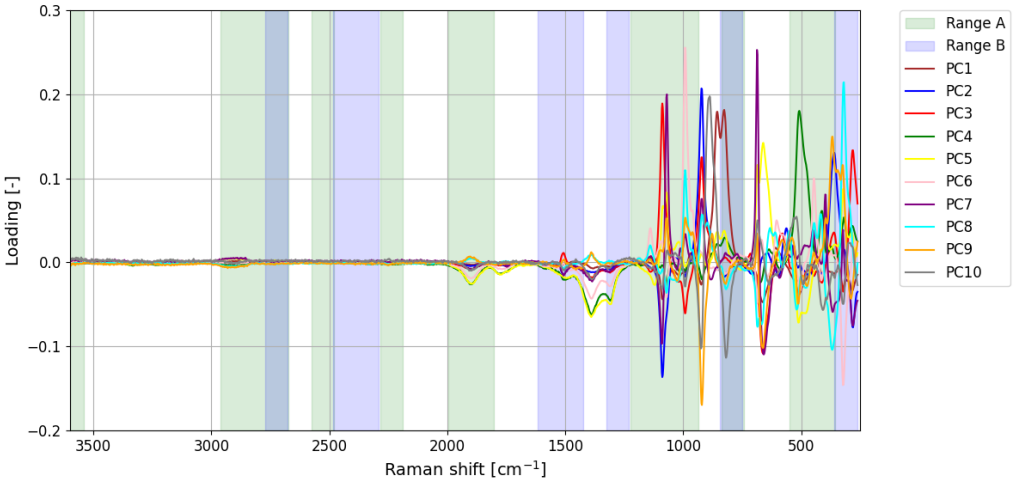


FIGURE A.3: Feature selection Raman A, B, C method

A.4 Comparison of the PLSR, XGBR and ANN modell performance parameters

Performance Metrics		**PLSR**			**XGBR**			**ANN**	
		HCl	*HF/HCl*	*AcOH*	*HCl*	*HF/HCl*	*AcOH*	*HCl*	*HF/HCl or AcOH*
RMSE [%]	train	4.963	5.375	5.338	1.992	2.516	1.437	2.51	3.394
RMSE [%]	cv	5.02	5.718	5.623	3.001	4.495	3.077	2.978	4.363
RMSE [%]	pred	4.199	5.119	2.222	4.368	4.534	1.403	3.794	4.539
R^2	train	0.954	0.823	0.954	0.997	0.981	0.999	0.994	0.979
R^2	cv	0.953	0.8	0.949	0.989	0.911	0.988	0.992	0.964
R^2	pred	0.937	0.969	0.552	0.943	0.829	0.855	0.987	0.968
RPD	train	4.674	2.374	4.643	12.249	5.105	18.628	9.25	4.829
RPD	cv	4.622	2.232	4.408	6.616	2.379	6.52	7.799	3.757
RPD	pred	2.235	2.116	1.423	2.149	2.389	2.254	2.591	3.526

The parameters were calculated using the formulas below [1]:

$$RMSE = \sum_{i=1}^{N} \sqrt{(\hat{y}_i - y_i)^2} \tag{A.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{A.2}$$

$$RPD = \sqrt{\frac{1}{1 - R^2}} \tag{A.3}$$

A.5 Highlighting the prediction results of the ANN model

FIGURE A.4: Histogram of ANN model predictions for solubility of cross-validation samples in (A) HCl, (B) HF/HCl, and (C) AcOH. The bars are the number of occurrences, the red line in the given range is the standard error of prediction [1].

We concluded that we could not use the CNN properly if the sample size is of less than a few thousand [156].

A.6 Case study no.1

Besides the above task, still based on mid-infrared spectra, we have checked the applicability of this method in a better defined system that contained fewer ambiguities than samples deriving from a more complex matrix. This system was a solution of three pure compounds [ethanol (EtOH), methyl ethyl ketone (MEK),

and ethyl acetate (EtOAc)] and the method was applied with 20 different mixtures of these three compounds. The method was then checked for two test mixtures of these three compounds and the predicted and the real infrared spectra were compared - see Figure: A.5. The real and artificially created spectra were much more similar in this simple system than they were in the more versatile geological samples. It might derive from two sources: the variance explained by the three principal components (99.3%) was higher than in the case of the rock solubility samples, and their solution was simpler compared to the rock samples. In this model, we used the first three principal components that explained 99.3% of the total variance to establish a relationship between the compositional data and the PCA values of the infrared spectra of the mixtures. For training, we used 15 of the 20 samples, 5 left for validation purposes. Figure: A.6 shows real and predicted points completely overlapping, in contrast to the rocky case study. Figure: A.7 shows the compositional data and the PCA values for both the training and the validations sets. They show an excellent match between the real and the predicted data pairs in the training and the validation samples. The comparison of the real and predicted infrared spectra delivered RMSE values of 0.0295, 0.0211, and 0.0442 for the training, validation, and test samples, respectively.

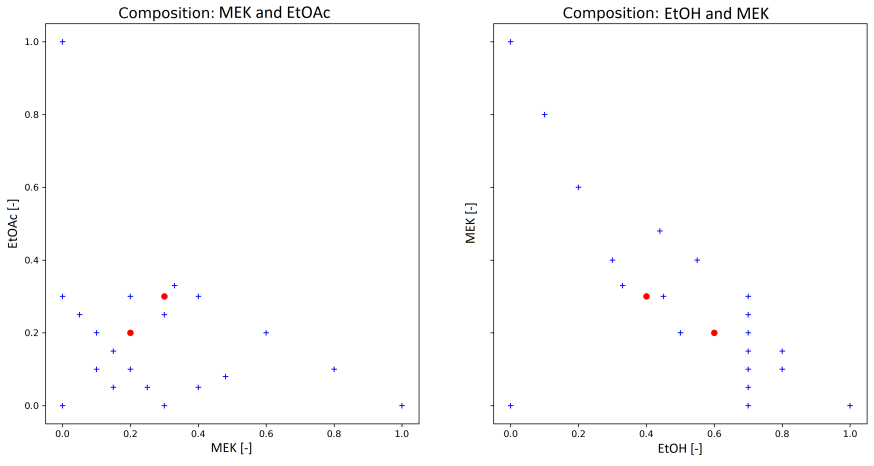


FIGURE A.5: The mixtures' compositions: blue crosses represent the known compositions, red dots represent the test mixtures

Table A.2 contains the Pearson correlation coefficients of these groups of samples. With 15 training samples at a 95% level of confidence, the threshold value is 0.514,

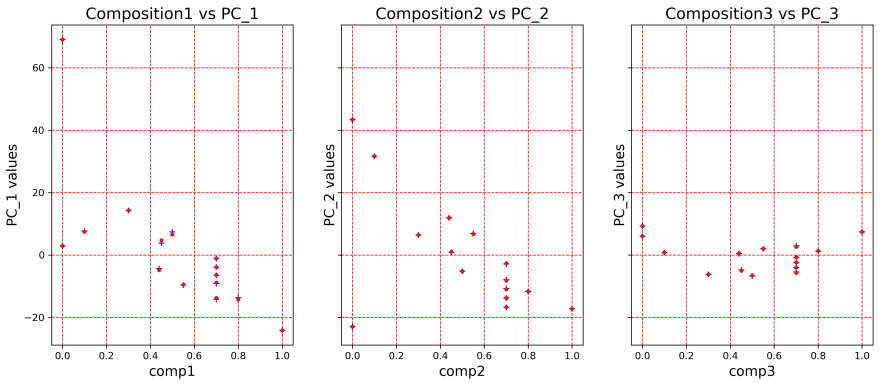


FIGURE A.6: Real [blue crosses] and predicted [red dots] PC values for the training set

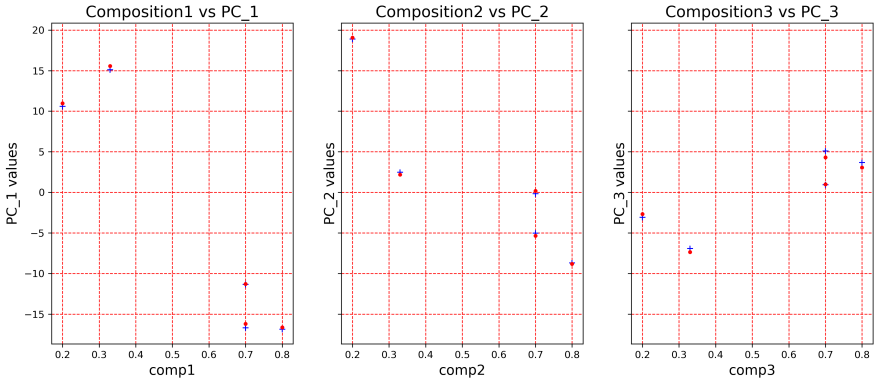


FIGURE A.7: Real [blue crosses] and predicted [red dots] PC values for the validation set

with five validation samples at a 95% level of confidence, the threshold value is 0.878 [103]. In the study of liquid mixtures, we have found that the artificial spectra generated by the linear combinations of pure compound spectra cannot be used for model building due to the interactions between the compounds of the mixture (e.g hydrogen-bonds). These interactions can cause changes in the peaks' shape and their locations (band shifts). Since these interactions are not incorporated into the model of the linear combination of the pure compounds, the spectra generated by the linear combination greatly differ from the real samples' infrared spectra. When choosing the number of training and validation samples, we took into account the recommendations of Rácz *et al.* [157]. Furthermore, in this case study, the generation of the artificial spectrum tested on the rock data presented in detail in the article was intentionally tested on the spectra of

a few other materials. Our goal with this test was to prove that the developed method can also perform well on small sample sets. When applying validation, we took into account the two common mistakes that can be made. The data set was properly divided into training and validation data sets, and we made sure that no validation loops were formed [158].

TABLE A.2: Comparison of the indicators of the three data sets.

Comparison	**Training set**	**Validation set**
number of samples	15	5
average Pearson coeff.	0.9986	0.9988
min Pearson coeff.	0.9918	0.9983
max Pearson coeff.	0.9999	0.9992

Figure A.8 shows the separately handled unknown spectra and their predicted versions.

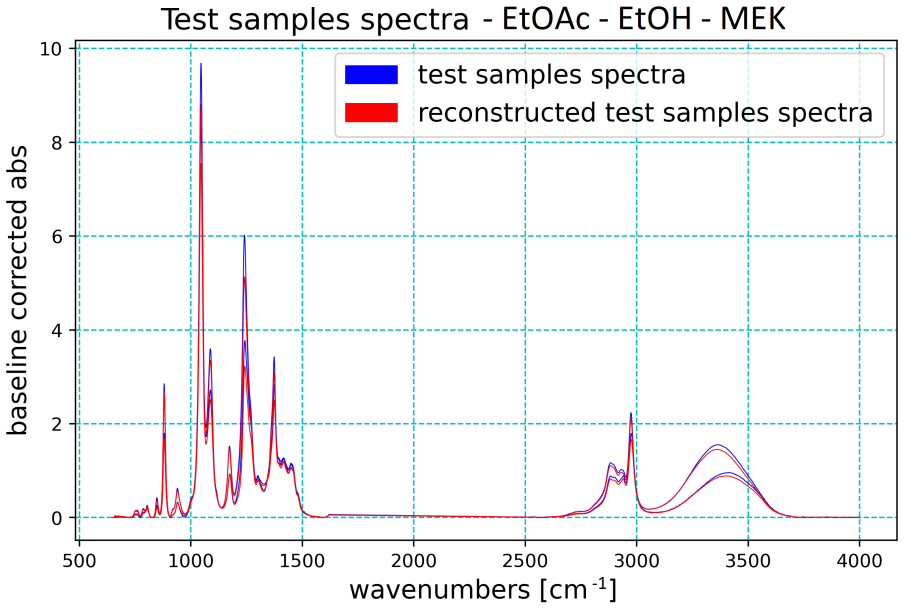


FIGURE A.8: Test mixtures' infrared spectra and their artificial versions

A.7 Phase of the CRISP-ML methodology

TABLE A.3: Phase of the CRISP-ML methodology

	Business Understanding	**Data Acquisition and Understanding**	**Model Development**	**Model Deployment**	**Model Operation**
Process	Agile business understanding for ML initialization.	Factor identification via Ishikawa diagrams; meticulous data preparation as a prerequisite for robust ML models.	Iterative ML model development and hyperparameter optimization.	Deploying validated and pre-configured ML models into production environments.	Continuous manual or automated deployment of production-ready models.
Objective	Quantifying business objectives through precise KPIs.	Profiling data quality, defining metrics with stakeholders, and executing multi-stage data cleaning and formatting.	Robust model development and numerical validation on standard data splits.	Production deployment of the top-performing robust model.	Integrating models into industrial systems for real-time production use.
Success	Cross-functional alignment of business goals and ML problem statements.	Detecting outliers and patterns for stakeholder validation and structured dataset preparation.	Metric-based evaluation and validation of deployment-ready models.	Parameter-driven continuous deployment.	Comprehensive MLOps integration for live model lifecycle management.
Considerations	Pre-project alignment of cross-functional stakeholders on shared objectives.	Statistical data profiling and multi-step preprocessing (encoding, scaling, and outlier handling).	Optimizing model performance via hyperparameter tuning and stakeholder-validated usability.	Ongoing verification of model consistency between development and production.	Real-time performance tracking to address model degradation, drift, and infrastructure scaling.
Tasks	Goal setting, KPI definition, and stakeholder alignment on success criteria.	Data source mapping, exploratory analysis, and multi-stage preprocessing.	Feature engineering, iterative model training, and comparative performance evaluation under MLOps standards.	Model scoring, versioning, and continuous performance tracking via MLOps.	Model serving and storage via web services, dashboards, and intelligent application integration.

Acronyms

a	element of the incidence matrix
A	incidence matrix
AcOH	acetic acid
AI	artificial intelligence
AMI	amazon machine image
ANFIS	adaptive neural fuzzy inferential system
ANN	artificial neural network
ARIMA	autoregressive integrated moving average
ASTM	American Society for Testing and Materials
ATR-FTIR	attenuated total reflectance Fourier transform infrared
$\tilde{b}$	stands for the constant values
BOD	biological oxygen demand
CA	California state
CI/CD	continuous integration/continuous delivery
CLF	complex-level-ensemble fusion
CNN	convolutional neural network
CRISP-DM	cross industry standard process for data mining
CRISP-ML	cross industry standard process for machine Learning
DevOps	compound of development and operations
DF	data fusion
DLS	dynamic light scattering
DLSS	deep learning based soft sensor
DMG	data mining group
DoE	design of the experiment

DR	data reconciliation
EAND	evolutionary algorithm with numerical differentiation
EDA	exploratory data analysis
ϵ	error of the predicted value
ERP	enterprise resource planning
FIR-CNN	finite impulse response convolutional neural networks
FTIR	Fourier-transform infrared spectroscopy
GA	genetic algorithm
HCl	hydrogen chloride
HF	hydrogen fluoride
HPLC	high-performance liquid chromatography
HSE	health and safety executive
HTS	hierarchical time series
HWES	holt–winters exponential smoothing
I	identity matrix
IBK	instance-based k-nearest neighbors
IoT	internet of things
IR	infrared
k	index of the hierarchical structure level
K	lowest level of the hierarchical structure
KPIs	key performance indicators
LCBS	laser-induced breakdown spectroscopy
LC-MS	liquid chromatography-mass spectrometry
LDA	linear discriminant analysis
LIMS	laboratory information management system
LSS	lean six sigma
LWPLS	locally weighted partial Least squares
LWDA-SAE	layer-wise data augmentation stacked autoencoder
MALS	multiangle light scattering
MIR	mid-infrared
ML	machine learning
MLOps	machine learning model operationalisation management
MSC	multiplicative scatter correction

MSE	mean squared error
n	number of the samples
MIR	mid-infrared
mUVE	uninformative variable elimination
NIPALS	nonlinear iterative partial least squares
NIR	near-infrared
NNR	near-neighbor regression
OEE	overall equipment effectiveness
OPA	outer product analysis
$\hat{P}$	projection matrix
PAT	process analytical technology
PC	principal component
PCA	principal component analysis
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
PLSR	partial least squares regression
PMML	predictive model markup language
PoC	proof of concept
PRISMA	preferred reporting items for systematic reviews and meta-analyses
PSE	process system engineering
R^2	correlation coefficient
ReLU	rectified linear unit
RFR	random forest regression
RI	refractive index
RMSE	root-mean-square deviation
RPD	relative percent differences
RPLS	recursive Partial Least Squares
S	summation matrix
SCGP	Shell coal gasication process
SIMCA	soft independent modeling of class analogy
SIS	spectral interference subtraction
SLS	static light scattering
SNV	standard normal variate

SPC	statistical process control
STAR	structured additive regression
t	timestamp
θ	parameter of the model
UV-Vis	ultraviolet–visible spectroscopy
Vis-NIR	visible-near-infrared
X	independent variable
XGBoost	extreme gradient boosting
XRF	X-ray fluorescence
y	dependent variable
10-cv	ten-fold cross-validation

Related publications to theses

[R1] 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. Scimago Journal Ranking: Q1, Impact Factor: 8.980.

[R2] 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.

[R3] 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.

[R4] 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. Scimago Journal Ranking: Q1, Impact Factor: 4.775.

[R5] László Győry, Szilveszter Gergely, and Pál Péter Hanzelik. Generating realistic infrared spectra using artificial neural networks. *Journal of Chemometrics*, 38(9):e3573, 2024. Scimago Journal Ranking: Q3, Impact Factor: 2.189.

[R6] 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. Scimago Journal Ranking: Q3, Impact Factor: 1.569.

[R7] 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. Scimago Journal Ranking: Q1-Q2, Impact Factor: 4.532.

[R8] 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, Availabil-ity:https://secure.confis.cz/chisa2022/UserPages/ContribListProgramPre.aspx.

[R9] Pál Péter Hanzelik, Alex Kummer, Ádám Ipkovich, and János Abonyi. Per-formance monitoring of machine learning models. Conference presentation, September 2022. at the Process Control Systems (PCS) conference, Availab-ility: https://pcsmeeting.hu/PCS2022/.

Further publications

[F1] 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. Scimago Journal Ranking: Q3, Impact Factor: 0.885.

[F2] 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. Scimago Journal Ranking: Q1, Impact Factor: 5.168.

[F3] 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.

[F4] S Puskas, A Vago, M Toro, T Ordog, Gy Kalman, P Hanzelik, Zs Bihari, J Blaho, R Tabajdi, I Dekany, et al. Surfactant-polymer eor from laboratory to the pilot. In *SPE EOR Conference at Oil and Gas West Asia*, page D011S001R003. SPE, 2018.

Bibliography

[1] 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.

[2] A Jiménez, G Beltrán, MP Aguilera, and M Uceda. A sensor-software based on artificial neural network for the optimization of olive oil elaboration process. *Sensors and Actuators B: Chemical*, 129(2):985–990, 2008.

[3] Francisco AA Souza, Rui Araújo, and Jérôme Mendes. Review of soft sensor methods for regression applications. *Chemometrics and Intelligent Laboratory Systems*, 152:69–79, 2016.

[4] Bhawani Shankar Pattnaik, Arunima Sambhuta Pattanayak, Siba Kumar Udgata, and Ajit Kumar Panda. Machine learning based soft sensor model for bod estimation using intelligence at edge. *Complex & Intelligent Systems*, 7(2):961–976, 2021.

[5] Zhenyu Wang and Leo Chiang. Monitoring chemical processes using judicious fusion of multi-rate sensor data. *Sensors*, 19(10):2240, 2019.

[6] Pascal Dufour, Sharad Bhartiya, Prasad S Dhurjati, and Francis J Doyle Iii. Neural network-based software sensor: training set design and application to a continuous pulp digester. *Control Engineering Practice*, 13(2):135–143, 2005.

[7] David Wang, Jun Liu, and Rajagopalan Srinivasan. Data-driven soft sensor approach for quality prediction in a refining process. *IEEE Transactions on Industrial Informatics*, 6(1):11–17, 2009.

[8] Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4):795–814, 2009.

[9] Ali Al-Jlibawi, Mohammed Lutfi Bin Othman, Muayed S Al-Huseiny, Ishak Bin Aris, and Samsul Bahari. The efficiency of soft sensors modelling in advanced

control systems in oil refinery through the application of hybrid intelligent data mining techniques. In *Journal of Physics: Conference Series*, volume 1529, page 052049. IOP Publishing, 2020.

[10] A Thiruneelakandan, Gaganpreet Kaur, Geetha Vadnala, N Bharathiraja, K Pradeepa, and Mervin Retnadhas. Measurement of oxygen content in water with purity through soft sensor model. *Measurement: Sensors*, 24:100589, 2022.

[11] Esin Iplik, Ioanna Aslanidou, and Konstantinos Kyprianidis. Hydrocracking: A perspective towards digitalization. *Sustainability*, 12(17):7058, 2020.

[12] Martin Mojto, Karol L'ubuškỳ, Miroslav Fikar, and Radoslav Paulen. Data-based design of inferential sensors for petrochemical industry. *Computers & Chemical Engineering*, 153:107437, 2021.

[13] Daniela CM de Souza, Luís Cabrita, Cláudia F Galinha, Tiago J Rato, and Marco S Reis. A spectral automl approach for industrial soft sensor development: Validation in an oil refinery plant. *Computers & Chemical Engineering*, 150:107324, 2021.

[14] Xiaofeng Yuan, Chen Ou, Yalin Wang, Chunhua Yang, and Weihua Gui. A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process. *IEEE transactions on neural networks and learning systems*, 32(8):3296–3305, 2019.

[15] Gustavo AP de Morais, Bruno HG Barbosa, Danton D Ferreira, and Leonardo S Paiva. Soft sensors design in a petrochemical process using an evolutionary algorithm. *Measurement*, 148:106920, 2019.

[16] A Jiménez, G Beltrán, MP Aguilera, and M Uceda. A sensor-software based on artificial neural network for the optimization of olive oil elaboration process. *Sensors and Actuators B: Chemical*, 129(2):985–990, 2008.

[17] Bhawani Shankar Pattnaik, Arunima Sambhuta Pattanayak, Siba Kumar Udgata, and Ajit Kumar Panda. Machine learning based soft sensor model

for bod estimation using intelligence at edge. *Complex & Intelligent Systems*, 7(2):961–976, 2021.

[18] Kangcheng Wang, Chao Shang, Lei Liu, Yongheng Jiang, Dexian Huang, and Fan Yang. Dynamic soft sensor development based on convolutional neural networks. *Industrial & Engineering Chemistry Research*, 58(26):11521–11531, 2019.

[19] Ling Yi, Jun Lu, Jinliang Ding, Changxin Liu, and Tianyou Chai. Soft sensor modeling for fraction yield of crude oil based on ensemble deep learning. *Chemometrics and Intelligent Laboratory Systems*, 204:104087, 2020.

[20] Murali K Maruthamuthu, Scott R Rudge, Arezoo M Ardekani, Michael R Ladisch, and Mohit S Verma. Process analytical technologies and data analytics for the manufacture of monoclonal antibodies. *Trends in biotechnology*, 38(10):1169–1186, 2020.

[21] Evangelos Spiliotis, Mahdi Abolghasemi, Rob J Hyndman, Fotios Petropoulos, and Vassilios Assimakopoulos. Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing*, 112:107756, 2021.

[22] George Athanasopoulos, Puwasala Gamakumara, Anastasios Panagiotelis, Rob J Hyndman, and Mohamed Affan. Hierarchical forecasting. *Macroeconomic forecasting in the era of big data: Theory and practice*, pages 689–719, 2020.

[23] Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, 55(9):2579–2589, 2011.

[24] Lukas Neubauer and Peter Filzmoser. Rediscovering bottom-up: Effective forecasting in temporal hierarchies. *arXiv preprint arXiv:2407.02367*, 2024.

[25] Jooyoung Jeon, Anastasios Panagiotelis, and Fotios Petropoulos. Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, 279(2):364–379, 2019.

[26] George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Anastasios Panagiotelis. Forecast reconciliation: A review. *International Journal of Forecasting*, 40(2):430–456, 2024.

[27] Rob J Hyndman and George Athanasopoulos. Optimally reconciling forecasts in a hierarchy. *Foresight: The International Journal of Applied Forecasting*, (35):42, 2014.

[28] Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob J Hyndman. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, 306(2):693–706, 2023.

[29] Tim Van Erven and Jairo Cugliari. Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and stochastic learning for forecasting in high dimensions*, pages 297–317. Springer, 2015.

[30] Peter Nystrup, Erik Lindström, Pierre Pinson, and Henrik Madsen. Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, 280(3):876–888, 2020.

[31] Rob J Hyndman, Alan J Lee, and Earo Wang. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational statistics & data analysis*, 97:16–32, 2016.

[32] Julien Leprince, Henrik Madsen, Jan Kloppenborg Møller, and Wim Zeiler. Hierarchical learning, forecasting coherent spatio-temporal individual and aggregated building loads. *arXiv preprint arXiv:2301.12967*, 2023.

[33] Sajjad Taghiyeh, David C Lengacher, Amir Hossein Sadeghi, Amirreza Sahebi-Fakhrabad, and Robert B Handfield. A novel multi-phase hierarchical forecasting approach with machine learning in supply chain management. *Supply Chain Analytics*, 3:100032, 2023.

[34] Mahsa Ashouri, Rob J Hyndman, and Galit Shmueli. Fast forecast reconciliation using linear models. *Journal of Computational and Graphical Statistics*, 31(1):263–282, 2022.

[35] 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 *Computer Aided Chemical Engineering*, volume 52, pages 1379–1384. Elsevier, 2023.

[36] Shankar Narasimhan and Cornelius Jordache. *Data reconciliation and gross error detection: An intelligent use of process data*. Elsevier, 1999.

[37] José Antonio Vélez Godiño and Francisco José Jiménez-Espadafor Aguilar. Joint data reconciliation and artificial neural network based modelling: Application to a cogeneration power plant. *Applied Thermal Engineering*, 236:121720, 2024.

[38] Michal Dabros, Michael Amrhein, Dominique Bonvin, Ian W Marison, and Urs von Stockar. Data reconciliation of concentration estimates from mid-infrared and dielectric spectral measurements for improved on-line monitoring of bioprocesses. *Biotechnology progress*, 25(2):578–588, 2009.

[39] O Bennouna, N Heraud, M Rodriguez, and H Camblong. Data reconciliation and gross error detection applied to wind power. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 221(3):497–506, 2007.

[40] Shankar Narasimhan and Nirav Bhatt. Deconstructing principal component analysis using a data reconciliation perspective. *Computers & Chemical Engineering*, 77:74–84, 2015.

[41] Pál Péter Hanzelik, Alex Kummer, and János Abonyi. Edge-computing and machine-learning-based framework for software sensor development. *Sensors*, 22(11):4268, 2022.

[42] Arun Senthil Sundaramoorthy. Probabilistic graphical models for data reconciliation and causal inference in process data analytics. M.sc. thesis, University of Alberta, 2021.

[43] V Balaram and SS Sawant. Indicator minerals, pathfinder elements, and portable analytical instruments in mineral exploration studies. *Minerals*, 12(4):394, 2022.

[44] Zhiyong Xu, Bahne Carl Cornilsen, Domenic C Popko, Wayne D Pennington, James R Wood, and Jiann-Yang Hwang. Quantitative mineral analysis by FTIR spectroscopy. *Internet Journal of Vibrational Spectroscopy*, 5(4), 2001.

[45] Mark D. Raven and Peter Self. Outcomes of 12 years of the reynolds cup quantitative mineral analysis round robin. *Clays and Clay Minerals*, 65:122–134, 2017.

[46] Oladipo Motoso, Douglas McCarty, Stephen Hillier, and Reinhard Kleeberg. Some successful approaches to quantitative mineral analysis as revealed by the reynolds cup contest. *Clays and Clay Minerals*, 54, 12 2006.

[47] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2022.

[48] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

[49] Lenka Zavíralová, R Šomplák, Martin Pavlas, J Kropác, Pavel Popela, Ondřej Putna, and Jiří Gregor. Computational system for simulation and forecasting in waste management incomplete data problems. *Chemical Engineering Transactions*, 45:763–768, 2015.

[50] Sergio De-la Mata-Moratilla, Jose-Maria Gutierrez-Martinez, Ana Castillo-Martinez, and Sergio Caro-Alvaro. Prediction of the behaviour from discharge

points for solid waste management. *Machine Learning and Knowledge Extraction*, 6(3):1389–1412, 2024.

[51] Ivan Eryganov, Martin Roseckỳ, Radovan Šomplák, and Veronika Smejkalová. Forecasting the waste production hierarchical time series with correlation structure. *Optimization and Engineering*, 26(2):781–803, 2025.

[52] Martin Pavlas, Radovan Somplak, Veronika Smejkalova, Vlastimir Nevrly, Lenka Zaviralova, Jakub Kudela, and Pavel Popela. Spatially distributed production data for supply chain models-forecasting with hazardous waste. *Journal of Cleaner Production*, 161:1317–1328, 2017.

[53] Prajakta S Kalekar et al. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi school of information Technology*, 4329008(13):1–13, 2004.

[54] Yizhuo Fan, Jiaqiang Wang, Gao Shu, Shu Xiao, Yit Jing Eb, and Kamaruzzaman Sopian. A real-time correction model for carbon emission measurement data and carbon emission factors in coal-fired power plants based on data fusion. In *Journal of Physics: Conference Series*, volume 3001, page 012033. IOP Publishing, 2025.

[55] Robert Schimanek, Pinar Bilge, and Franz Dietrich. Data fusion for improved circularity through higher quality of prediction and increased reliability of inspection. *International Journal of Sustainable Manufacturing*, 5(2-4):164–199, 2022.

[56] Metrohm. A guide to near-infrared spectroscopic analysis of industrial manufacturing processes, 2014.

[57] Davide Ballabio, Elisa Robotti, Francesca Grisoni, Fabio Quasso, Marco Bobba, Serena Vercelli, Fabio Gosetti, Giorgio Calabrese, Emanuele Sangiorgi, Marco Orlandi, et al. Chemical profiling and multivariate data fusion methods for the identification of the botanical origin of honey. *Food Chemistry*, 266:79–89, 2018.

[58] Shima Ghanavati Nasab, Mehdi Javaheran Yazd, Federico Marini, Riccardo Nescatelli, and Alessandra Biancolillo. Classification of honey applying high performance liquid chromatography, near-infrared spectroscopy and chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 202:104037, 2020.

[59] Matthew Dirks, David Turner, and David Poole. Spectral sensor fusion for prediction of li and zr in rocks: Neural network and pls methods. *Chemometrics and Intelligent Laboratory Systems*, 240:104915, 2023.

[60] Clara Pérez-Ràfols, Núria Serrano, and José Manuel Díaz-Cruz. Authentication of soothing herbs by uv–vis spectroscopic and chromatographic data fusion strategy. *Chemometrics and Intelligent Laboratory Systems*, 235:104783, 2023.

[61] Maogang Li, Jia Xue, Yao Du, Tianlong Zhang, and Hua Li. Data fusion of raman and near-infrared spectroscopies for the rapid quantitative analysis of methanol content in methanol–gasoline. *Energy & Fuels*, 33(12):12286–12294, 2019.

[62] S Hamed Javadi and Abdul M Mouazen. Data fusion of xrf and vis-nir using outer product analysis, granger–ramanathan, and least squares for prediction of key soil attributes. *Remote Sensing*, 13(11):2023, 2021.

[63] Jingyi Zhu, Xia Fan, Lu Han, Chong Zhang, Jiahong Wang, Leiqing Pan, Kang Tu, Jing Peng, and Mingzhi Zhang. Quantitative analysis of caprolactam in sauce-based food using infrared spectroscopy combined with data fusion strategies. *Journal of Food Composition and Analysis*, 104:104130, 2021.

[64] Brigitta Nagy, Dulichár Petra, Dorián László Galata, Balázs Démuth, Enikő Borbás, György Marosi, Zsombor Kristóf Nagy, and Attila Farkas. Application of artificial neural networks for process analytical technology-based dissolution testing. *International journal of pharmaceutics*, 567:118464, 2019.

[65] Jia Chen, Fayin Ye, and Guohua Zhao. Rapid determination of farinograph parameters of wheat flour using data fusion and a forward interval variable selection algorithm. *Analytical methods*, 9(45):6341–6348, 2017.

[66] BP Geurts, J Engel, B Rafii, L Blanchet, A Suppers, E Szymańska, JJ Jansen, and LMC Buydens. Improving high-dimensional data fusion by exploiting the multivariate advantage. *Chemometrics and Intelligent Laboratory Systems*, 156:231–240, 2016.

[67] Shouxin Ren and Ling Gao. Combining artificial neural networks with data fusion to analyze overlapping spectra of nitroaniline isomers. *Chemometrics and Intelligent Laboratory Systems*, 107(2):276–282, 2011.

[68] Sihai Li, Yangyang Wang, Hang Song, and Mingqi Liu. Multi-spectral fusion and self-attention mechanisms for gentiana origin identification via near-infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 246:105068, 2024.

[69] Fabricio S Terra, Raphael A Viscarra Rossel, and Jose AM Dematte. Spectral fusion by outer product analysis (opa) to improve predictions of soil organic c. *Geoderma*, 335:35–46, 2019.

[70] Erin Gibbons, Richard Léveillé, and Kim Berlo. Data fusion of laser-induced breakdown and raman spectroscopies: Enhancing clay mineral identification. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 170:105905, 2020.

[71] Omneya Attallah. Multitask deep learning-based pipeline for gas leakage detection via e-nose and thermal imaging multimodal fusion. *Chemosensors*, 11(7):364, 2023.

[72] Barbara Lafuente, Robert T Downs, Hexiong Yang, Nate Stone, Thomas Armbruster, Rosa Micaela Danisi, et al. The power of databases: the rruff project. *Highlights in mineralogical crystallography*, 1:25, 2015.

[73] Eric Deconinck, Celine Duchateau, Margot Balcaen, Lies Gremeaux, and Patricia Courselle. Chemometrics and infrared spectroscopy–a winning team for the analysis of illicit drug products. *Reviews in Analytical Chemistry*, 41(1):228–255, 2022.

[74] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, 2009.

[75] M.S. Dhanoa, S.J. Lister, R. Sanderson, and R.J. Barnes. The link between multiplicative scatter correction (msc) and standard normal variate (snv) transformations of nir spectra. *Journal of Near Infrared Spectroscopy*, 2(1):43–47, 1994.

[76] Eigenvector Research Documentation Wiki. Advanced preprocessing: Noise, offset, and baseline filtering.

[77] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[78] Riccardo Leardi. Genetic algorithms in feature selection. In *Genetic algorithms in molecular modeling*, pages 67–86. Elsevier, 1996.

[79] John McCall. Genetic algorithms for modelling and optimisation. *Journal of computational and Applied Mathematics*, 184(1):205–222, 2005.

[80] A guide to near-infrared spectroscopic analysis of industrial manufacturing processes.

[81] Alok Kumar and J Mayank. Ensemble learning for ai developers. *BApress: Berkeley, CA, USA*, 2020.

[82] Robert T Downs and M Bonner Denton. Report on the progress of the rruff project: An integrated database of raman spectra, x-ray diffraction, and chemical data for minerals. *Gems & Gemology*, 42(3), 2006.

[83] Rafaella de F Sales, Luan Cássio Barbosa-Patrício, Neirivaldo C da Silva, Lívia Rodrigues e Brito, Maria Eduarda Fernandes da Silva, and Maria Fernanda Pimentel. Gasoline discrimination using infrared spectroscopy and virtual samples based on measurement uncertainty. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 303:123248, 2023.

[84] Éva Szabó, Szilveszter Gergely, and András Salgó. Linear discriminant analysis, partial least squares discriminant analysis, and soft independent modeling of class analogy of experimental and simulated near-infrared spectra of a cultivation medium for mammalian cells. *Journal of Chemometrics*, 32(4):e3005, 2018.

[85] Tom O'Haver. A pragmatic introduction to signal processing. *University of Maryland at College Park*, 1997.

[86] Jinhyung Kwon, Jiseok Kim, Hanjin Kim, SongHyun Kim, Seungsoo Jang, Janghee Lee, and Young-su Kim. Development of gamma-spectrum data generation method by monte carlo simulation. *Journal of the Korean Physical Society*, 82(7):658–670, 2023.

[87] PA Mazzali and LB Lucy. The application of monte carlo methods to the synthesis of early-time supernovae spectra. *Astronomy and Astrophysics*, 279:447–456, 1993.

[88] Sheng-Yang Tsui, Chiao-Yi Wang, Tsan-Hsueh Huang, and Kung-Bin Sung. Modelling spatially-resolved diffuse reflectance spectra of a multi-layered skin model by artificial neural networks trained with monte carlo simulations. *Biomedical optics express*, 9(4):1531–1544, 2018.

[89] Krzysztof B Beć and Christian W Huck. Breakthrough potential in near-infrared spectroscopy: Spectra simulation. a review of recent developments. *Frontiers in chemistry*, 7:48, 2019.

[90] Christian P Minor, Joseph C Gezo, and Kevin J Johnson. Information measures for multisensor systems. Technical report, 2013.

[91] Ruihao Luo, Juergen Popp, and Thomas Bocklitz. Deep learning for raman spectroscopy: A review. *Analytica*, 3(3):287–301, 2022.

[92] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and

Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[93] Wartini Ng, Budiman Minasny, Wanderson de Sousa Mendes, and José Alexandre Melo Demattê. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *Soil*, 6(2):565–578, 2020.

[94] Subrato Bharati, Prajoy Podder, and M Mondal. Artificial neural network based breast cancer screening: a comprehensive review. *arXiv preprint arXiv:2006.01767*, 2020.

[95] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[96] Michael J Economides, Kenneth G Nolte, et al. In *Reservoir stimulation*, volume 2, pages 17–1 – 18–28. Prentice Hall Englewood Cliffs, NJ, 1989.

[97] Acid solubility.recommended practices for core analysis recommended practice 40. Standard, American Petroleum Institute, USA, March 1998.

[98] Ronald W Kennard and Larry A Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.

[99] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[100] T Davies and Tom Fearn. Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, 16(6):20, 2004.

[101] Shuxia Guo, Thomas Bocklitz, Ute Neugebauer, and Jürgen Popp. Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417, 2017.

[102] Károly Héberger. Frequent errors in modeling by machine learning: A prototype case of predicting the timely evolution of covid-19 pandemic. *Algorithms*, 17(1):43, 2024.

[103] Philip R Bevington and D Keith Robinson. Data reduction and error analysis. *McGraw Hil, New York*, 2003.

[104] Ningren Han and Rajeev J Ram. Bayesian modeling and computation for analyte quantification in complex mixtures using raman spectroscopy. *Computational Statistics & Data Analysis*, 143:106846, 2020.

[105] Harald Martens and Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of pharmaceutical and biomedical analysis*, 9(8):625–635, 1991.

[106] Peijin Tong, Yiping Du, Kaiyi Zheng, Ting Wu, and Jiajun Wang. Improvement of nir model by fractional order savitzky-golay derivation (fosgd) coupled with wavelength selection. *Chemometrics and Intelligent Laboratory Systems*, 143:40–48, 2015.

[107] Paul Selzer, Johann Gasteiger, Henrik Thomas, and Reiner Salzer. Rapid access to infrared reference spectra of arbitrary organic compounds: scope and limitations of an approach to the simulation of infrared spectra by neural networks. *Chemistry–A European Journal*, 6(5):920–927, 2000.

[108] Nattane Luíza da Costa, Maxwell Severo da Costa, and Rommel Barbosa. A review on the application of chemometrics and machine learning algorithms to evaluate beer authentication. *Food Analytical Methods*, 14(1):136–155, Jan 2021.

[109] Yi Xu, Peng Zhong, Aimin Jiang, Xing Shen, Xiangmei Li, Zhenlin Xu, Yudong Shen, Yuanming Sun, and Hongtao Lei. Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC Trends in Analytical Chemistry*, 131:116017, 2020.

[110] Andrei A. Bunaciu, Vu Dang Hoang, and Hassan Y. Aboul-Enein. Applications of ft-ir spectrophotometry in cancer diagnostics. *Critical Reviews in Analytical Chemistry*, 45(2):156–165, 2015.

[111] Camila Maione, Fernando Barbosa, and Rommel Melgaço Barbosa. Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review. *Computers and Electronics in Agriculture*, 157:436–446, 2019.

[112] António João Silva, Paulo Cortez, and André Pilastri. Chemical laboratories 4.0: A two-stage machine learning system for predicting the arrival of samples. *IFIP International Conference on Artificial Intelligence Applications & Innovations*, pages 232–243, 2020.

[113] Marco S. Reis and Pedro M. Saraiva. Data-centric process systems engineering: A push towards pse 4.0. *Computers & Chemical Engineering*, 155:107529, 2021.

[114] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[115] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. Towards crisp-ml (q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413, 2021.

[116] Hae-Woo Lee, Hemlata Bhatia, Seo-Young Park, Mark-Henry Kamga, Thomas Reimonn, Sha Sha, Zhuangrong Huang, Shaun Galbraith, Huolong Liu, and Seongkyu Yoon. Process analytical technology and quality by design for animal cell culture. *Cell Culture Engineering: Recombinant Protein Production*, pages 365–390, 2019.

[117] Inés Sittón-Candanedo, Ricardo S Alonso, Sara Rodríguez-González, José Alberto García Coria, and Fernando De La Prieta. Edge computing architectures in industry 4.0: A general survey and comparison. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 121–131. Springer, 2019.

[118] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88:105906, 2021.

[119] Marta Abelha, Sandra Fernandes, Diana Mesquita, Filipa Seabra, and Ana Teresa Ferreira-Oliveira. Graduate employability and competence development in higher education—a systematic literature review using prisma. *Sustainability*, 12(15):5900, 2020.

[120] PJ Escamilla-Ambrosio, A Rodríguez-Mota, E Aguirre-Anaya, R Acosta-Bermejo, and M Salinas-Rosales. Distributing computing in the internet of things: cloud, fog and edge computing overview. In *Results of the Numerical and Evolutionary Optimization Workshop NEO Workshop September 20-24, 2016 in Tlalnepantla, Mexico*, pages 87–115. Springer, 2017.

[121] Peter Mell and Tim Grance. The NIST definition of cloud computing. 2011.

[122] Hugh Boyes, Bil Hallaq, Joe Cunningham, and Tim Watson. The industrial internet of things (IIoT): An analysis framework. *Computers in industry*, 101:1–12, 2018.

[123] Kenneth A Rose, Shaye Sable, Donald L DeAngelis, Simeon Yurek, Joel C Trexler, William Graf, and Denise J Reed. Proposed best modeling practices for assessing the effects of ecosystem restoration on fish. *Ecological Modelling*, 300:12–29, 2015.

[124] Dazhong Wu, Shaopeng Liu, Li Zhang, Janis Terpenny, Robert X. Gao, Thomas Kurfess, and Judith A. Guzzo. A fog computing-based framework for process monitoring and prognosis in cyber-manufacturing. *Journal of Manufacturing Systems*, 43:25–34, 2017.

[125] Massimo Villari, Antonio Celesti, and Maria Fazio. Towards osmotic computing: Looking at basic principles and technologies. In *Complex, Intelligent,*

and Software Intensive Systems, pages 906–915, Cham, 2018. Springer International Publishing.

[126] Jin Yang, Ye Huang, and Matthew W Nelson. System and method for ultra-low latency short data service, May 20 2021. US Patent App. 16/689,506.

[127] Kshitij Arun Doshi, Francesc Cesc Guim Bernat, and Suraj Prabhakaran. Ai model and data transforming techniques for cloud edge, August 17 2021. US Patent 11,095,618.

[128] Michael Stearns, Mark Barlow Hammer, Chanh V Hua, Sunil Gopalkrishna, and Yang Wang. Edge device disablement, December 15 2020. US Patent 10,867,076.

[129] Henrik Sundström, Basuki Priyanto, Andrej Petef, Lars Nord, and Anders Isberg. Mechanism for machine learning in distributed computing, December 24 2020. US Patent App. 16/970,479.

[130] Basuki Priyanto, Andrej Petev, Henrik Sundström, Anders Isberg, Anders Mellqvist, and Lars Nord. Method and device for computing estimation output data, October 31 2019. US Patent App. 16/295,048.

[131] Jean Peccoud. Methods, services, systems, and architectures to optimize laboratory processes, September 16 2021. US Patent App. 17/203,690.

[132] Zhibin Zhou, Lingyun Sun, Yuyang Zhang, Xuanhui Liu, and Qing Gong. Ml lifecycle canvas: Designing machine learning-empowered ux with material lifecycle thinking. *Human–Computer Interaction*, 35(5-6):362–386, 2020.

[133] Mahmoud Elsisi, Karar Mahmoud, Matti Lehtonen, and Mohamed MF Darwish. Reliable industry 4.0 based on machine learning and iot for analyzing, monitoring, and securing smart meters. *Sensors*, 21(2):487, 2021.

[134] Minh-Quang Tran, Mahmoud Elsisi, Karar Mahmoud, Meng-Kun Liu, Matti Lehtonen, and Mohamed MF Darwish. Experimental setup for online fault diagnosis of induction machines via promising iot and machine learning: Towards industry 4.0 empowerment. *IEEE Access*, 9:115429–115441, 2021.

[135] MG Sarwar Murshed, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain. Machine learning at the network edge: A survey. *ACM Computing Surveys (CSUR)*, 54(8):1–37, 2021.

[136] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. Who needs mlops: What data scientists seek to accomplish and how can mlops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE, 2021.

[137] Lucas Baier, Fabian Jöhren, and Stefan Seebacher. Challenges in the deployment and operation of machine learning in practice. In *ECIS*, 2019.

[138] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[139] MB Adamu. Fourier transform infrared spectroscopic determination of shale minerals in reservoir rocks. *Nigerian Journal of Basic and Applied Sciences*, 18(1):35–43, 2010.

[140] Carolina T Pinheiro, Ricardo Rendall, Margarida J Quina, Marco S Reis, and Lici´nio M Gando-Ferreira. Assessment and prediction of lubricant oil properties using infrared spectroscopy and advanced predictive analytics. *Energy and Fuels*, 31(1):179–187, 2017.

[141] Abebe Diro, Naveen Chilamkurti, Van-Doan Nguyen, and Will Heyne. A comprehensive study of anomaly detection schemes in iot networks using machine learning algorithms. *Sensors*, 21(24):8320, 2021.

[142] Julia Zeckl, Matthias Wastian, Dominik Brunmeir, Andrea Rappelsberger, Sergei B Arseniev, and Klaus-Peter Adlassnig. From machine learning to knowledge-based decision support—a predictive-model-markup-language-to-arden-syntax transformer for decision trees. In *Soft Computing for Biomedical Applications and Related Topics*, pages 89–99. Springer, 2021.

[143] Piero Molino and Christopher Ré. Declarative machine learning systems: The future of machine learning will depend on it being in the hands of the rest of us. *Queue*, 19(3):46–76, 2021.

[144] Xiaodong Zhu and Jianzheng Yang. An extended predictive model markup language for data mining. In *International Conference on Web-Age Information Management*, pages 218–231. Springer, 2010.

[145] Alex Guazzelli, Michael Zeller, Wen-Ching Lin, Graham Williams, et al. Pmml: An open standard for sharing models. *R J.*, 1(1):60, 2009.

[146] Max Ferguson, Kincho H. Law, Raunak Bhinge, David Dornfeld, Jinkyoo Park, and Yung-Tsun Tina Lee. Evaluation of a pmml-based gpr scoring engine on a cloud platform and microcomputer board for smart manufacturing. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2014–2023, 2016.

[147] Tibor Kulcsar, Peter Koncz, Miklos Balaton, Laszlo Nagy, and Janos Abonyi. Statistical process control based energy monitoring of chemical processes. In *Computer Aided Chemical Engineering*, volume 33, pages 397–402. Elsevier, 2014.

[148] Achinthya D Perera, Nihal P Jayamaha, Nigel P Grigg, Mark Tunnicliffe, and Amardeep Singh. The application of machine learning to consolidate critical success factors of lean six sigma. *IEEE Access*, 9:112411–112424, 2021.

[149] Gautam Dutta, Ravinder Kumar, Rahul Sindhwani, and Rajesh Kr Singh. Digitalization priorities of quality control processes for smes: A conceptual study in perspective of industry 4.0 adoption. *Journal of Intelligent Manufacturing*, 32(6):1679–1698, 2021.

[150] Tao Zan, Zhihao Liu, Zifeng Su, Min Wang, Xiangsheng Gao, and Deyin Chen. Statistical process control with intelligence based on the deep learning model. *Applied Sciences*, 10(1):308, 2019.

[151] Roberto González Velázquez, Iñaki Bravo-Imaz, Kerman López de Calle-Etxabe, and Aitor Arnaiz. A flexible data management system for the analysis of an electro-mechanical actuator on a test bench. In *PHM Society European Conference*, volume 6, pages 8–8, 2021.

[152] I Sessione di Laurea. *MLOps-Standardizing the Machine Learning Workflow*. PhD thesis, University of Bologna, 2021.

[153] Haining Zheng, Antonio R Paiva, and Chris S Gurciullo. Advancing from predictive maintenance to intelligent maintenance with ai and iiot. *arXiv preprint arXiv:2009.00351*, 2020.

[154] Hans E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:76, 2018.

[155] Alexander E Stott, Sithan Kanna, Danilo P Mandic, and William T Pike. An online nipals algorithm for partial least squares. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4177–4181. IEEE, 2017.

[156] L Norgaard, Martin Lagerholm, and Mark Westerhaus. Artificial neural networks and near infrared spectroscopy - A case study on protein content in whole wheat grain. *Foss White Paper*, 2013.

[157] Anita Rácz, Dávid Bajusz, and Károly Héberger. Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4):1111, 2021.

[158] S Guo, T Bocklitz, U Neugebauer, and J Popp. Common mistakes in cross-validating classification models. anal methods. 2017; 9 (30): 4410–7.