




Data-Driven and Model-Driven Approaches in Predictive Modelling for Operational Efficiency: Mining Industry Use Case

Oussama Hasidi^{1,2} , El Hassan Abdelwahed¹, My Abdellah El Alaoui-Chrifi³, Aimad Qazdar⁴, François Bourzeix², Intissar Benzakour³, Ahmed Bendaouia^{1,2}, and Charifa Dahhassi⁵

¹ LISI Laboratory, Faculty of Sciences Semailia, Cadi Ayyad University, Marrakesh, Morocco

Oussama.hasidi@ced.uca.ma

² AI Department, Moroccan Foundation for Advanced Science, Innovation and Research (MAScIR), Rabat, Morocco

³ R&D and Engineering Center, Reminex, Managem Group, Marrakesh, Morocco

⁴ ESTIDMA Laboratory, ENSA, Ibn Zohr University, Agadir, Morocco

⁵ Math and Computer Science Department, ENSAM, Moulay Ismail University, Meknes, Morocco

Abstract. In this study, we explore the effectiveness of a hybrid modelling approach that seamlessly integrates data-driven techniques, specifically Machine Learning (ML), with physics-based equations in Simulation. In cases where real-world data for industrial processes is insufficient, a simulation tool is employed to generate an extensive dataset of process variables under varying operating conditions. Subsequently, this dataset is utilized for training the Machine Learning model. The paper showcases a practical use case of this hybrid modelling approach, revealing a model that consistently demonstrates strong predictive accuracy and reliability within the specific industrial context we investigate. By merging the insights derived from physics-based understanding with the adaptability of data-driven Machine Learning, the hybrid model offers a comprehensive solution for precise and accurate predictions.

Keywords: Predictive Modelling · Machine Learning · Simulation · Hybrid approach · Process Monitoring

1 Introduction

In today's world, modelling has become incredibly important for understanding complex systems, making predictions, and making smart choices. With everything being so interconnected and data-driven, modelling is like a guiding light that helps us make sense of complicated things. Researchers, industries, and

practitioners are investing a lot in modelling to better understand how systems work and to find solutions that can make a real impact.

In the realm of modelling, two primary approaches are employed. One involves using extensive data and computer learning, while the other relies on well-established scientific principles. These are known as data-driven modelling and physics-based modelling. Data-driven modelling employs computers to uncover patterns within large datasets. It's particularly valuable in fields such as healthcare and finance. Conversely, physics-based modelling applies scientific laws to construct simulated versions of real-world scenarios. This aids in comprehending phenomena like fluid motion or material behavior.

Machine learning and simulation both share a common objective: the anticipation of system behavior through data analysis and mathematical representation [1]. In the context of our study, these approaches hold intrinsic significance. Machine learning, as evident in domains like image classification, linguistic analysis, and socio-economic exploration, has exhibited exceptional achievements. These triumphs are especially pronounced when dealing with scenarios characterized by limited causal insights but extensive datasets. Conversely, simulation finds its historical roots in disciplines such as natural sciences and engineering. Notably, computational fluid dynamics relies on simulation to understand intricate causal inter-plays, while areas like structural mechanics employ it for evaluating structural performance encompassing reactions, stresses, and displacements.

This paper introduces a hybrid approach for developing predictive models and discusses a use case of model development for industrial froth flotation process using Machine Learning and Simulation. By leveraging a simulator to generate data and Artificial Neural Networks to construct the core virtual model, this hybrid approach harnesses the strengths of both model-driven and data-driven methods. This strategic combination overcomes the limitations posed by each approach - the scarcity of real data in model-driven methods and the complexities faced by data-driven methods in handling intricate systems.

The remainder of this paper is organized as follows: Sect. 2 further discusses the data-driven and physics-driven modelling approaches, the advantages of each one and their challenges. Section 3 elucidates the hybrid approach where the both approach could assist each other and mitigate the challenges faced by each one. Section 4 outlines a use case of model development of froth flotation process using hybrid approach for minerals processing advanced monitoring. Section 5 presents the results and analysis of the study, including the evaluation of the predictive model for a functioning flotation cell. Finally, Sect. 6 concludes the paper with a summary of key findings and recommendations for future research.

2 Modelling Approaches

In this section, we elucidate the two modelling methodologies through a conceptual framework designed to enhance their transparency and facilitate meaningful comparisons between their respective components.

2.1 Data-Driven Approach (i.e Machine Learning)

Data-driven modelling with machine learning has gained significant traction in recent years due to the explosion of available data and advancements in algorithms. This approach involves training models on large datasets to learn complex patterns and relationships, enabling them to make accurate predictions and inform decision-making processes.

The benefits of data-driven modelling are numerous and impactful. By harnessing the power of big data, organizations can make data-informed decisions, enhancing efficiency and accuracy. Moreover, these models excel at handling non-linear relationships and can uncover hidden insights within data. For example, in genomics, machine learning has facilitated the identification of novel disease-associated genes and pathways, revolutionizing our understanding of complex genetic disorders. Furthermore, data-driven modelling's adaptability to diverse data types, from text to images, has paved the way for innovations in natural language processing, image recognition, and autonomous vehicles.

Nevertheless, several challenges accompany this approach. One key challenge is the issue of data quality and bias. Biased or incomplete datasets can lead to skewed model outcomes and reinforce existing prejudices. Another challenge is model interpretability, where complex algorithms like deep neural networks can be difficult to explain, raising questions about transparency and accountability in decision-making. Addressing these challenges necessitates a comprehensive understanding of data collection practices, algorithmic transparency, and adherence to regulatory guidelines.

2.2 Physics-Driven Approach (i.e Simulation)

Physics-driven modelling through simulation is a potent tool for understanding and predicting complex real-world phenomena. However, this approach is not without challenges that need to be carefully considered. The goal of a simulation is to predict the behavior of a system or process based on the underlying physical principles. To achieve this, a mathematical model is formulated using differential equations that capture the causal relationships between different variables. These models are often developed through extensive research, starting with the derivation of equations from theoretical physics principles, followed by validation through experiments.

One major challenge is the dependence on accurate input parameters and assumptions in the mathematical models. Small inaccuracies in these inputs can lead to significant discrepancies in the simulation results. In complex systems with numerous interacting components, it can be difficult to precisely estimate all necessary parameters, introducing uncertainties in the predictions. Moreover, assumptions made during model formulation might not hold true in all scenarios, further affecting the reliability of simulations. Addressing these challenges requires rigorous sensitivity analysis to identify the impact of input variations and a thorough understanding of the assumptions' validity.

Computational intensity is another challenge in physics-driven simulations, particularly when dealing with complex behaviors such as quantum mechanics or highly nonlinear interactions. Simulating such phenomena demands substantial computational resources and advanced numerical techniques. As the complexity of the system increases, the computational cost can become prohibitive, requiring researchers to strike a balance between accuracy and computational feasibility.

Furthermore, while physics-driven models are adept at capturing well-defined causal relationships, they can struggle to represent emergent behaviors and complex interactions that arise in systems with many interdependent components. For instance, in agent-based simulations, where individual agents interact based on predefined rules, it can be challenging to predict emergent patterns that arise from the collective behavior of the agents. This limitation can be addressed by integrating data-driven approaches that leverage machine learning techniques to capture these complex interactions.

Table 1. Summary of advantages and limitations of data driven and model driven approaches.

	Data-driven approach	Model-driven approach
Advantages	Simplicity	Better understanding of complex systems
	Faster decision-making	Generalized to similar problems
	Increased accuracy with more data	Accuracy: limited errors
Limitations	Quality and quantity of data	Lack of flexibility
	High cost of data collection, storage, and analysis	Technical expertise required
	Technical expertise required	Time-consuming

In summary, as represented in Table 1 the data-driven approach may be advantageous in situations where the physical system is complex and difficult to model or where there is a large amount of data available. However, this approach can be limited by the quality and availability of data, and may not always be suitable for systems with complex nonlinear behaviors [2]. On the other hand, the model-driven approach may be advantageous in situations where the physical system is well-understood and there is a good understanding of the underlying physics. However, this approach can be limited by the accuracy of the mathematical model and the assumptions made in developing the model.

3 Combination of Data-Driven and Model-Driven Approach

Considering the advantages and limitations of both approaches, we propose a hybrid modelling approach for the development of accurate and reliable predic-

tive models for leveraging the advantages of the both methodologies. The combination of data-based and knowledge-based modelling is motivated by applications that are partly based on causal relationships, while other effects result from hidden dependencies that are represented in huge amounts of data [1] (Fig. 1).

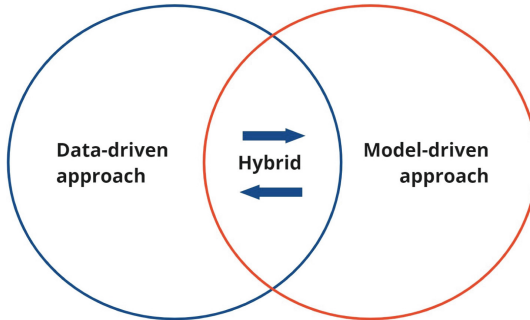


Fig. 1. Data-driven and Model-driven modelling approaches combination.

The data-driven approach could assist the model-driven approach in various ways. Tiny Machine Learning-based algorithms could be advantageous for the inference of complex measurements that are hard to measure or that are not bound to be directly measured. Statistical methods could also play a major role in studying and investigating hidden parameters dependencies. Moreover, Data assimilation and Reinforcement Learning play an important role in guarantying the Predictive Layer accurate performance after deployment.

Inversely, the model-driven approach could as well assist the data-driven approach. In the modelling phase for instance, additional training data could be generated based on physical and phenomenological equations to be fed to Machine Learning models. Furthermore, testing Machine Learning models is less risky if firstly validated in a simulation-based environment that is built with physics and phenomenological equations.

Overall, the hybrid approach has the potential to improve the accuracy and robustness of predictive models. Diverse are the applications that include the combination of the two methods in different modelling areas [3–5]. In the next section, we will showcase one way of combining data-driven and model-driven approaches in predictive modelling for operational efficient in the mining industry using Simulation and Machine Learning.

4 Use Case: Froth Flotation Predictive Model Using Hybrid Approach

4.1 Froth Flotation Overview

Froth flotation is the most common way to separate minerals in the mining industry. Since it was first used in factories in 1905, many researchers have

contributed to improving how we understand the process [6]. This method is the most widely used way to separate valuable minerals from useless rock. It works by taking advantage of the different ways minerals react with water. Some minerals don't like water (hydrophobic), while others do (hydrophilic). When we stir the mixture and add air, minerals that don't like water stick to tiny bubbles and rise to the top as foam. Minerals that like water stay in the liquid. This helps us get the valuable minerals out for further use (Fig. 2).

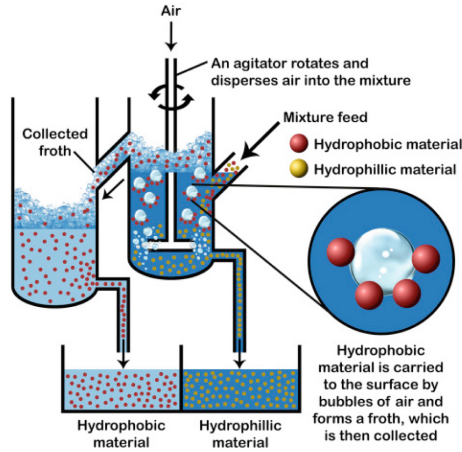


Fig. 2. A schematic representation of flotation process.

Advances in control and optimisation of the froth flotation process are of great relevance since even very small increases in recovery lead to large economic benefits [6]. However, the implementation of advanced control and optimisation strategies has not been completely successful in flotation since ever. This is because flotation performance is affected by a great number of variables that interact with each other, while unmeasurable disturbances in the process further complicate the implementation of efficient strategies [7]. Besides the multiple chemicals reagents (Collectors, Modifiers and Frothers), froth flotation is influenced by several operating factors. Many parameters involved in the process and have a crucial impact on the quality of the froth floated and therefore the concentration of the minerals resulted: the pH within the liquid, the rate of oxidation of the ore, the grain size of the feed, The viscosity of the pulp, air flows, agitators speed,...etc.

Acquiring a specific knowledge of the operational behavior of the process and trying to ensure a certain stability of the flotation yield (in terms of concentration/grade) is considered a complicated and tough challenge. Hence the necessity of the development of an accurate and reliable predictive models that comprehends the underlying phenomenon of the process and that incorporates these affecting parameters as an advanced operations monitoring system [8]. In

the further section we will discuss the development of Machine Learning based predictive model of the froth flotation cell using phenomenological, kinetic and physics-based simulation data.



Fig. 3. Froth flotation virtual model development: Hybrid approach methodology.

4.2 Methodology

Given the scarcity of sufficient real-world data records for essential process parameters in industrial plants, a simulation tool was employed to generate an extensive dataset of process variables across various operating conditions (Fig. 3). This simulator, known as HSC Chemistry, operates on a model-driven approach, integrating well-established physical and chemical principles that govern the froth flotation process. The simulator incorporates essential equations, such as the calculation of the overall flotation rate k (1/min) of a cell, which is derived from the overall recovery (mass pull) R (%) and cell residence time W (min) using the formula:

$$k = \frac{R}{W(100 - R)} \quad (1)$$

The Eq. 1 play a crucial role in accurately simulating the flotation process and provide valuable insights into the system's behavior under different operating conditions.

By running simulations with HSC Chemistry, we obtained a wealth of data that enabled the training of the data-driven component of our hybrid approach.

4.3 Operational Data Generation and Collection

In this study, we conducted data collection using HSC Chemistry [9], a minerals processing simulator, as illustrated in Fig. 4. Given the limited availability of operational data from the industrial flotation plant, we used a small amount of industrial data to pre-configure the simulations in the software. These data comprised crucial variables such as feed rate, water flow rate, pulp density, flotation rate and mineral grades. Through multiple simulations, we generated a diverse dataset encompassing a wide range of operating conditions. HSC Chemistry was

chosen as the simulator due to its remarkable accuracy and flexibility in simulating various mineral processing scenarios. The simulator allowed us to obtain an extensive amount of data that would have been challenging and costly to acquire through traditional experimental methods in real industrial settings. Following data generation, we carefully curated the dataset to ensure it captured the most pertinent and critical process parameters. This comprehensive dataset served as the foundation for developing the predictive model. This approach not only optimized data collection efforts but also ensured a high level of accuracy and efficiency in the development of predictive models for the froth flotation process.

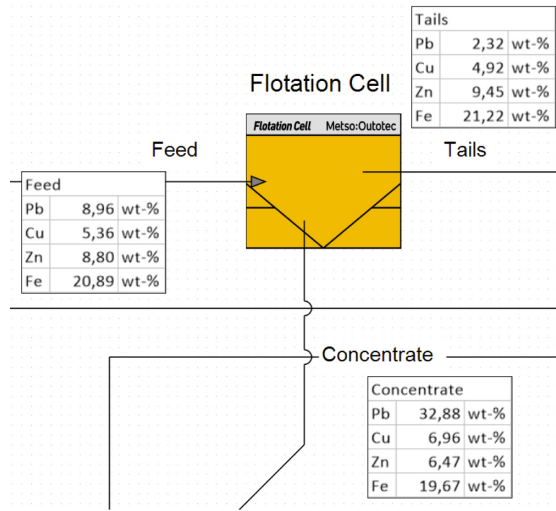


Fig. 4. Single flotation cell simulation using HSC Chemistry Sim.

4.4 Artificial Neural Networks Training

Following the data collection, the gathered dataset was employed to train a Neural Network model, enabling predictions of the froth flotation process's behavior under new operating conditions. This Neural Network model was built on a data-driven approach, learning the intricate nonlinear relationships between the input and output variables. Through training the ANN model with the dataset generated from the simulator, we established a predictive model for the froth flotation process.

The predictive model facilitated the estimation of mineral grades before and after entering a specific flotation cell, based on feed characteristics and control variables. For this study, a feed-forward backpropagation network structure was adopted, with the appropriate configuration determined by adjusting the number of neurons, hidden layers, transfer functions, and optimization methods.

Satisfactory outcomes were achieved through careful selection of the number of layers and neurons in each layer. The final architecture network is presented in Table 2.

Table 2. Architecture of the trained Feed Forward Neural Network.

Structure's parameters	ANN
Number of Layers	3
Number of neurons in the input layer	64
Number of neurons in the hidden layer	64
Number of neurons in the output layer	4
Activation function of the input hidden layer	ReLU
Activation function of the hidden layer	ReLU
Activation function of the output layer	Linear
Optimizer function	Adegrad
Learning rate	0.01
Loss function	MSE
Number of epochs	500
Batch size	32

To optimize the model's accuracy, the training and testing sets were allocated 80% and 20% of the samples, respectively. While maintaining a fixed structure for the Artificial Neural Network (ANN), the learning hyperparameters were fine-tuned using the training set. This iterative process helped enhance the performance of the model and improve its predictive capabilities.

5 Results and Discussion

The integration of operational records from the industrial plant into the HSC Chemistry software allowed us to establish simulations for the froth flotation process. Utilizing these simulated datasets, we trained a Neural Network to estimate mineral grades in the concentrates based on feed and control variables. The Neural Network exhibited exceptional accuracy, achieving a predictive capability of 95% with an MSE of less than 2. This confirms the model's capability to accurately estimate mineral grades in the concentrates using input variables.

Figures 5, 6, 7, and 8 visually demonstrate the comparison between actual and predicted values of minerals-of-concern grades, showcasing the robust predictive capability of the Neural Network model. Combining the simulator with the Neural Network model resulted in a powerful hybrid model. The simulator provided a physics-based foundation that encompassed the process's underlying principles, while the Neural Network model incorporated data-driven insights to capture complex nonlinear relationships between input and output variables.

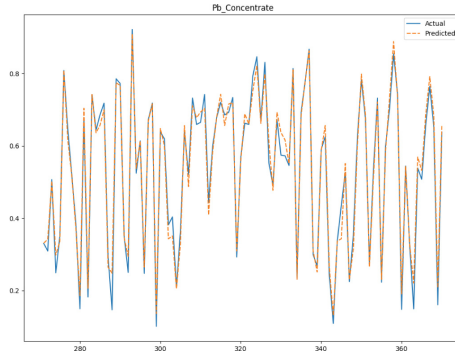


Fig. 5. Grades of Lead Pb% in the flotation cell's concentrates: Actual values are represented in blue and the Neural Networks predictions values are represented in orange. (Color figure online)

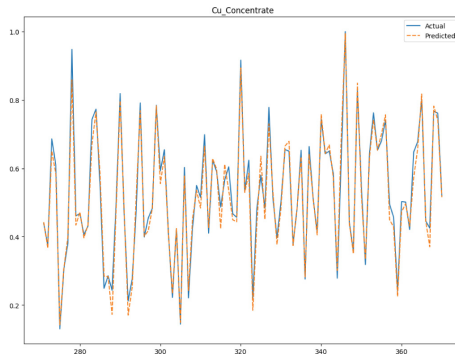


Fig. 6. Grades of Copper Cu% in the flotation cell's concentrates: Actual values are represented in blue and the Neural Networks predictions values are represented in orange. (Color figure online)

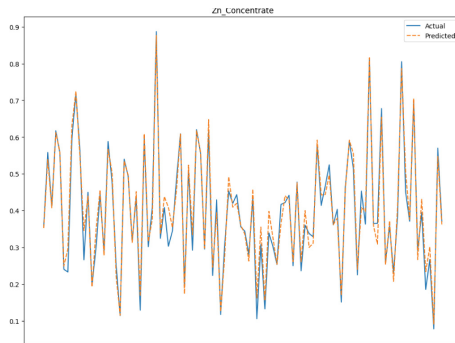


Fig. 7. Grades of Zinc Zn% in the flotation cell's concentrates: Actual values are represented in blue and the Neural Networks predictions values are represented in orange. (Color figure online)

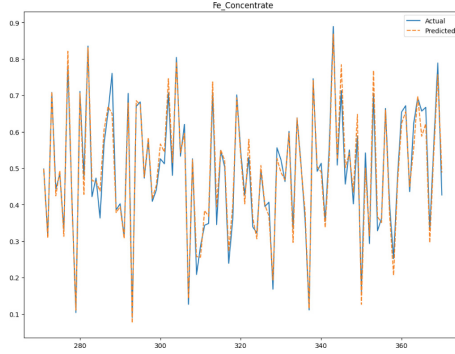


Fig. 8. Grades of Iron Fe% in the flotation cell’s concentrates: Actual values are represented in blue and the Neural Networks predictions values are represented in orange. (Color figure online)

This hybrid model effectively predicted the froth flotation process’s behavior under novel operating conditions, offering valuable insights for process optimization. Our approach exemplifies the potential of hybrid modelling techniques in developing predictive models for intricate industrial processes, such as froth flotation. By synergizing physics-based and data-driven models, we successfully captured the process’s intricate and nonlinear behavior, empowering us to make precise predictions under diverse operating conditions.

6 Conclusion

In conclusion, this study successfully showcased the creation of a predictive model for the froth flotation process through a hybrid approach, combining simulation and Machine Learning. The simulator generated crucial operational data, which was then utilized to train a Neural Network, resulting in the creation of a robust virtual model. The outcomes of the model exhibited impressive accuracy, with a predictive capability of 95% and an MSE of less than 2. This promising result underscores the potential application of this hybrid modelling approach using simulators and Machine Learning for an advanced operations monitoring.

Moreover, this research emphasizes the significance of Digital Twins in industrial productions and their transformative impact on the mining sector [8, 10]. By incorporating Industry 4.0 technologies such as the Internet of Things (IoT), Cyber-Physical Systems (CPS), Big Data, and Cloud Computing, mining operations can significantly enhance productivity and efficiency, fostering positive economic and environmental sustainability [11].

Acknowledgements. This work is supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco through Al-Khawarizmi program. This publication is part of the

work undertaken by the consortium of partners which is composed of MAScIR (Moroccan Foundation for Advanced Science, Innovation and Research), Reminex; the R&D and Engineering subsidiary of Managem group, UCA, ENSIAS and ENSMR. We would like to thank the Managem Group and its subsidiary CMG for allowing the conduction of this research on its operational site as an industrial partner of this project.

References

1. Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., Garcke, J.: Combining machine learning and simulation to a hybrid modelling approach: current and future directions. In: *Advances In Intelligent Data Analysis XVIII*, pp. 548–560 (2020)
2. Liao, L., Köttig, F.: A hybrid framework combining data-driven and model-based methods for system remaining useful life prediction. *Appl. Soft Comput.* **44**, 191–199 (2016)
3. Erge, O., Oort, E.: Combining physics-based and data-driven modelling in well construction: hybrid fluid dynamics modelling. *J. Nat. Gas Sci. Eng.* **97**, 104348 (2022). <https://www.sciencedirect.com/science/article/pii/S1875510021005436>
4. Song, H., Liu, X., Song, M.: Comparative study of data-driven and model-driven approaches in prediction of nuclear power plants operating parameters. *Appl. Energy* **341**, 121077 (2023). <https://www.sciencedirect.com/science/article/pii/S0306261923004415>
5. Zhang, S., et al.: Combing data-driven and model-driven methods for high proportion renewable energy distribution network reliability evaluation. *Int. J. Electr. Power Energy Syst.* **149**, 108941 (2023). <https://www.sciencedirect.com/science/article/pii/S0142061522009371>
6. Michaud, L.: Froth Flotation: A Century of Innovation (2017). <https://www.911metallurgist.com/blog/froth-flotation-century-innovation>
7. Bendaouia, A., et al.: Digital transformation of the flotation monitoring towards an online analyzer. In: Hamlich, M., Bellatreche, L., Siadat, A., Ventura, S. (eds.) *SADASC 2022. Communications in Computer and Information Science*, vol. 1677, pp. 325–338. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20490-6_26
8. Hasidi, O., et al.: Digital Twins-Based Smart Monitoring and Optimisation of Mineral Processing Industry. In: Hamlich, M., Bellatreche, L., Siadat, A., Ventura, S. (eds.) *SADASC 2022. Communications in Computer and Information Science*, vol. 1677, pp. 411–424. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20490-6_33
9. Roine, A.: HSC Chemistry® [Software], Metso Outotec, Pori (2021). Software available at www.mogroup.com/hsc
10. Sircar, A., Nair, A., Bist, N., Yadav, K.: Digital Twin in hydrocarbon industry. *Petrol. Res.* (2022)
11. Qassimi, S., Abdelwahed, E.H.: Disruptive innovation in mining industry 4.0. *Distrib. Sens. Intell. Syst.* 313–325 (2022)