Silica grade forecast in flotation processes: Evaluation and Statistical Diagnosis of Machine Learning Methods

Ray da Silva Basilio[1], Iuri da Silva Diniz[2], Alex Júnior Guimarães[3], Gilmar Vagner dos Santos[1], Alan Kardek Rêgo Segundo[3], Rodrigo César Pedrosa Silva[4]

[1]Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Ouro Preto – MG, Brazil

[2]PPG em Instrum., Controle e Automação Universidade Federal de Ouro Preto e Instituto Tecnológico Vale – MG, Brazil

[3]Departamento de Controle e Automação (DECAT), Escola de Minas, Universidade Federal e Ouro Preto (UFOP) – MG, Brazil

[4]Departamento de Computação, Universidade Federal de Ouro Preto – MG, Brazil

{ray.basilio,iuri.diniz}@aluno.ufop.edu.br

{alex.guimaraes,gilmar.santos}@aluno.ufop.edu.br

{alankardek,rodrigo.pedrosa}@ufop.edu.br

Abstract. *This study evaluates machine learning models for one-step-ahead ($T + 1$) silica grade forecasting in iron ore flotation, comparing autoregressive (AR) and autoregressive plus process variables (AR+Var) strategies on a private industrial dataset and a public Kaggle dataset. The evaluated models were Decision Tree, Random Forest, XGBoost, and LSTM, with a Naive persistence forecast used as baseline. The results show that the Naive method is a strong benchmark, especially in the industrial dataset, where several tree-based models did not outperform it. The best predictive result in the industrial scenario was obtained by the AR-LSTM ($R^2 = 0.75$, MAE = 0.37), but residual diagnostics indicated remaining autocorrelation and bias. In the public dataset, Random Forest achieved consistent performance ($R^2 = 0.66$ in both AR and AR+Var) and passed residual diagnostic tests, while LSTM achieved the highest R^2 in AR (0.67) but failed residual adequacy tests.*

1. Context

Iron ore flotation plays a central role in concentrate quality control, especially for impurity reduction such as silica (SiO_2) [Filippov et al. 2014, Suhasini et al. 2015, Drumond et al. 2018]. However, flotation performance is affected by process variability, reagent action, and physicochemical conditions, which makes real-time control difficult [Fan et al. 2020]. In practice, product quality variables are often obtained from delayed laboratory measurements, creating time lags that limit operational response and process optimization [Warne et al. 2004, Perera et al. 2023].

Machine learning has been increasingly applied to mineral processing for grade prediction and process support [McCoy and Auret 2019, Szmigiel et al. 2024,

Nakhaei et al. 2022, Pural 2023]. Although prior studies report promising predictive results, many evaluations emphasize error metrics while giving limited attention to residual diagnostics in time-series forecasting [Pu et al. 2020, Pu et al. 2021]. Since residual independence and near-zero mean are important adequacy conditions in forecasting analysis [Hyndman and Athanasopoulos 2018, de Oliveira e Lucas et al. 2020], this work compares AR (only target lags) and AR+Var (target lags plus process variables) strategies using both predictive metrics and statistical residual validation across two datasets. In addition, the problem is treated as a one-step-ahead forecasting task ($T+1$), in which the silica grade is estimated before the corresponding laboratory result becomes available, helping reduce the impact of analysis delays and supporting timely corrective actions in operation.

2. Approach adopted

This work followed an experimental proof-of-concept approach for short-term silica forecasting using two datasets: a private industrial dataset (*base_1*) and a public flotation dataset from Kaggle (*base_2*) [Oliveira 2017]. *base_1* comprises 18 months of operation (Jan/2022–Jul/2023), with 59 process and quality variables (e.g., levels, pressures, physicochemical variables, and reagent dosages), and the target variable is the silica grade in the concentrate (SiO_2 C flot). *base_2* spans approximately 6 months of production (Mar/2017–Sep/2017) and contains 23 operational and quality variables, including feed quality, reagent flow rates, pulp conditions, and flotation column parameters, with mixed sampling frequencies (hourly and every 20 s).

The adopted process comprised: (i) preprocessing and temporal alignment; (ii) model configuration and hyperparameter optimization; and (iii) training, dynamic validation, and statistical evaluation (Fig. 1). Two input strategies were tested: **AR** (only target lags) and **AR+Var** (target lags plus process variables).

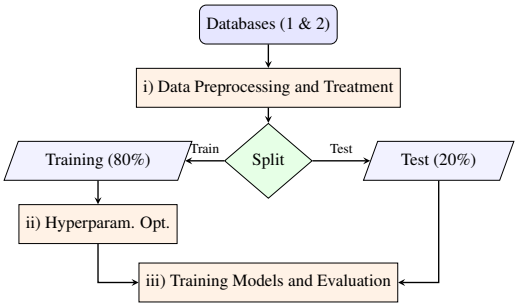

Figure 1. Flowchart of the proposed methodology.

The evaluated models were Naive (persistence baseline), Decision Tree, Random Forest, XGBoost, and LSTM. Hyperparameters were optimized using RandomizedSearchCV, and forecasts were generated in a dynamic one-step-ahead scheme. Performance was assessed by MSE ($\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$), MAE ($\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$), RMSE ($\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$), R^2 ($R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2}$), and Theil's U_2 ($U_2 = \sqrt{\frac{\sum_{t=2}^{n}(\hat{y}_t-y_t)^2}{\sum_{t=2}^{n}(y_t-y_{t-1})^2}}$), with residual diagnostics (Ljung-Box and t-test) to evaluate autocorrelation and bias.

2.1. Data Preprocessing and Treatment

To ensure analytical consistency, the following steps were applied:

1. **Expert-based filtering** (*Base 1*): records inconsistent with process physics were removed, constraining feed flow rate to ≤ 350 m^3/h and density to ≤ 1.3 t/m^3.
2. **Outlier and missing-value handling**: negative values (sensor errors) were removed; additional outliers were treated using the 3σ rule ($\mu(x) \pm 3\sigma(x)$); missing values were imputed using the mean.
3. **Synchronization and resampling**: *Base 2* was resampled to hourly frequency (first value of each hour), and both datasets were aligned to a common 6-month horizon.
4. **Variable selection and feature engineering**: iron content in the concentrate was excluded due to its high correlation with the target (silica in the concentrate). To represent temporal dynamics, the target variable was included as input via lags ($t-1$ to $t-6$) and moving averages (3, 6, and 12 periods), avoiding data leakage.
5. **Normalization and partitioning**: Min-Max scaling [Guo et al. 2021] was applied ($X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$), followed by a strict temporal split (80% training, 20% testing).

3. Solution: Results and benefits for the organization and software practice

Table 1. Comprehensive Performance (Metrics and Statistical Diagnostic).

Model	BASE 1 (Private - Mining Company)												BASE 2 (Public - Kaggle)											
	AR						AR + Variables						AR						AR + Variables					
	MSE	MAE	RMSE	R^2	U^2	Diag.	MSE	MAE	RMSE	R^2	U^2	Diag.	MSE	MAE	RMSE	R^2	U^2	Diag.	MSE	MAE	RMSE	R^2	U^2	Diag.
Naive (Last)	0.96	0.40	0.98	0.64	1.0	$\times_L$	0.96	0.40	0.98	0.64	1.0	$\times_L$	0.52	0.45	0.72	0.63	1.00	$\times_L$	0.52	0.45	0.72	0.63	1.00	$\times_L$
Decision Tree	1.26	0.67	1.12	0.55	1.14	$\times_{LT}$	1.14	0.56	1.07	0.57	1.09	$\times_{LT}$	0.50	0.50	0.71	0.64	0.99	✓	0.51	0.49	0.72	0.64	0.99	✓
Random Forest	1.22	0.64	1.10	0.57	1.12	$\times_{LT}$	**0.93**	**0.52**	**0.97**	**0.65**	**0.98**	$\times_{LT}$	0.48	0.48	0.69	0.66	0.96	✓	**0.48**	**0.48**	**0.69**	**0.66**	**0.96**	✓
XGBoost	1.17	0.61	1.08	0.58	1.10	$\times_{LT}$	0.99	0.58	0.99	0.63	1.01	$\times_L$	0.53	0.50	0.73	0.62	1.01	✓	0.51	0.51	0.72	0.64	0.99	$\times_L$
LSTM	**0.54**	**0.37**	**0.74**	**0.75**	**0.76**	$\times_{LT}$	3.15	1.17	1.78	-0.19	1.81	$\times_{LT}$	**0.47**	**0.48**	**0.68**	**0.67**	**0.94**	$\times_{LT}$	0.55	0.52	0.75	0.61	1.03	$\times_{LT}$

Legend: Diag.: Statistical residual diagnostic. ✓: Passed (Independent and Unbiased). $\times_L$: Failed Ljung-Box Test (Correlated residuals). $\times_T$: Failed t-Test (Biased residuals/mean $\neq 0$). $\times_{LT}$: Failed both tests. **Bold**: Best results per metric and category.

Table 1 shows that the Naive method is a strong baseline, particularly for the industrial dataset (*Base 1*). In *Base 1* (AR), Naive reached $R^2 = 0.64$, outperforming Decision Tree ($R^2 = 0.55$), Random Forest ($R^2 = 0.57$), and XGBoost ($R^2 = 0.58$). This result reinforces the importance of strong baseline comparisons in forecasting studies, especially for noisy or persistent time series [Hewamalage et al. 2023].

For *Base 1*, the best predictive metrics were obtained by the AR-LSTM (MSE = 0.54, MAE = 0.37, RMSE = 0.74, $R^2 = 0.75$, $U_2 = 0.76$), indicating that historical target dynamics were highly informative in the private industrial series. However, residual diagnostics failed (Ljung-Box and t-test), indicating remaining autocorrelation and bias. Therefore, although the model improved predictive accuracy, the statistical evidence suggests that relevant temporal structure remained unexplained [Hyndman and Athanasopoulos 2018]. In the AR+Var configuration of *Base 1*, Random Forest was the best tree-based model ($R^2 = 0.65$, $U_2 = 0.98$), but residual diagnostics also failed, while LSTM performance degraded substantially ($R^2 = -0.19$).

In *Base 2* (public Kaggle dataset), Random Forest was the most consistent model across configurations, achieving $R^2 = 0.66$ and $U_2 = 0.96$ in both AR and AR+Var, while passing residual diagnostics in both cases. Decision Tree also passed diagnostics in both setups, with $R^2 = 0.64$. LSTM achieved the highest R^2 in *Base 2* under AR ($R^2 = 0.67$, $U_2 = 0.94$), but failed residual diagnostics, which limits its statistical adequacy despite strong predictive performance. XGBoost presented mixed behavior, passing diagnostics in AR ($R^2 = 0.62$) but failing Ljung-Box in AR+Var ($R^2 = 0.64$).

Overall, the results indicate that there is no universally superior model across datasets and feature configurations. The industrial dataset favored AR-LSTM in raw predictive accuracy, but not in residual adequacy, while the public dataset favored Random Forest as the most balanced option between predictive performance and statistical validation.

4. Conclusion

This study compared techniques from the literature and showed that there is no universally superior model for short-term silica grade forecasting across different datasets and input configurations. In the private industrial dataset, the autoregressive (AR) LSTM achieved the best predictive metrics, but it did not satisfy the residual diagnostic tests, indicating remaining autocorrelation and bias. In the public dataset, Random Forest showed the most balanced behavior, combining competitive predictive performance with adequate residual diagnostics. These results suggest that, although the models are promising, there is still room for refinement to capture temporal structures that remain unexplained.

As a practical contribution, this work proposes a safer model selection process for virtual sensing in flotation by combining error metrics with residual diagnostics and objectively comparing AR and AR+Var strategies. In addition, by enabling early prediction of silica grade, the approach can support operators in taking corrective actions in a timely manner, helping to keep the process within the grade ranges specified for operation.

However, this study is limited to the evaluation of only two datasets, which restricts the generalization of the results. As future work, we recommend expanding the number of analyzed datasets, broadening the set of statistical diagnostic tests, and incorporating conformal prediction to provide uncertainty measures that support operational decision-making and increase the practical usefulness of the solution in industry. Even so, based on the results obtained, this work provides a reproducible evaluation protocol for forecasting in an industrial context—integrating a baseline (Naive), predictive metrics, and residual validation—which strengthens validation criteria prior to deployment and can be reused in other applications involving delayed laboratory measurements and time-series prediction.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq), the Instituto Tecnológico Vale (ITV), and the Universidade Federal de Ouro Preto (UFOP).

References

de Oliveira e Lucas, P., Alves, M. A., de Lima e Silva, P. C., and Guimarães, F. G. (2020). Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks. *Computers and Electronics in Agriculture*, 177:105700.

Drumond, P., Kappes, D., De Moraes, C., et al. (2018). Soft sensor: Machine learning tradicional ou deep learning? In *Seminário de Automação e TI*, pages 231–242.

Fan, G., Wang, L., Cao, Y., et al. (2020). Collecting agent–mineral interactions in the reverse flotation of iron ore: A brief review. *Minerals*, 10(8).

Filippov, L., Severov, V., and Filippova, I. (2014). An overview of the beneficiation of iron ores via reverse cationic flotation. *International Journal of Mineral Processing*, 127:62–69.

Guo, X., Gao, Y., Li, Y., et al. (2021). Short-term household load forecasting based on long- and short-term time-series network. *Energy Reports*, 7:58–64.

Hewamalage, H., Ackermann, K., and Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2):788–832.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

McCoy, J. and Auret, L. (2019). Machine learning applications in minerals processing: A review. *Minerals Engineering*, 132:95–109.

Nakhaei, F., Rahimi, S., and Fathi, M. (2022). Prediction of sulfur removal from iron concentrate using column flotation froth features: Comparison of k-means clustering, regression, backpropagation neural network, and convolutional neural network. *Minerals*, 12(11):1434.

Oliveira, E. M. (2017). Quality prediction in a mining process. [Online].

Perera, Y. S., Ratnaweera, D., Dasanayaka, C. H., et al. (2023). The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review. *Engineering Applications of Artificial Intelligence*, 121:105988.

Pu, Y., Szmigiel, A., and Apel, D. B. (2021). Purities prediction in a manufacturing froth flotation plant: the deep learning techniques. *Neural Computing and Applications*, 33:12217–12227.

Pu, Y., Szmigiel, A., Chen, J., and Apel, D. B. (2020). Flotationnet: A hierarchical deep learning network for froth flotation recovery prediction. *Powder Technology*, 375:317–326.

Pural, Y. E. (2023). Developing a data-driven soft sensor to predict silicate impurity in iron ore flotation concentrate. *Physicochemical Problems of Mineral Processing*, 59(5).

Suhasini, R., Mallick, A. K., Vasumathi, N., et al. (2015). Evaluation of flotation collectors in developing zero waste technology for processing iron ore tailings. *International Journal of Engineering Research*, 4(11):604–608.

Szmigiel, A., Apel, D. B., Skrzypkowski, K., et al. (2024). Advancements in machine learning for optimal performance in flotation processes: A review. *Minerals*, 14(4):57.

Warne, K., Prasad, G., Rezvani, S., et al. (2004). Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. *Engineering Applications of Artificial Intelligence*, 17(8):871–885.