https://doi.org/10.56801/MME1027

Multivariate Analysis of the Main Operational Variables Involved in Steel Producing on BOF Using Time Series Tools

Luccas Esper Klotz[1], Guilherme Frederico Bernardo Lenz e Silva[2], Natalia Piedemonte Antoniassi[3], Ronaldo Adriano Alvarenga Borges[4], Thales Arantes Kerche Nunes[5]

[1]Metallurgical Engineer, Department of Metallurgy and Material Engineering, University of São Paulo, Brazil, E-Mail: luccas.klotz@usp.br
[2]Metallurgical, Safety and Environment Engineer, Department of Metallurgy and Material Engineering, University of São Paulo, Brazil, E-Mail: guilhermelenz@usp.br
[3]Chemical Engineer, Department of Metallurgy and Material Engineering, University of São Paulo, Brazil, E-Mail: nataliapiedemonte@gmail.com
[4]Metallurgical, Engineer, Department of Metallurgy and Material Engineering, University of São Paulo, Brazil, E-Mail: ronaldousp@usp.br
[5]Materials engineer, Department of Metallurgy and Material Engineering, University of São Paulo, Brazil, E-Mail: thalesaknunes22@gmail.com

Received: 04-11-2023
Accepted: 18-12-2023

Abstract: There is significant interest in accurately modeling the operational variables of the steel-making process in LD converters. Despite this, the task is challenging due to the complex interactions between process variables, which are not entirely comprehended. Often, decisions in the industry are grounded in experience. This study aims to introduce a robust model that can effectively guide engineers and technicians by forecasting the future behavior of steelmaking variables in the BOF furnace. We employed multivariate time series analysis to reach this goal, utilizing tools like Vector Autoregression models, ElasticNet, K-Nearest-Neighbors, Multiple Linear Regression, and Long Short-Term Memory Neural networks. These models were tested on data from three distinct steel production campaigns. A successful model was identified, predicting 35 out of the 42 chosen variables, demonstrating the potential to correlate a majority of the selected parameters.

Keywords: BOF, Time series analysis, AI, Modeling.

1. Introduction

The ongoing fourth industrial revolution, commonly known as Industry 4.0, signifies a rising wave of digital transformation sweeping the global market. At its core, this revolution is reshaping our lifestyles and professional realms, ushering in hopeful prospects for sustainability [1]. In this Industry 4.0 paradigm, interconnected computers, smart materials, and intelligent machinery work in tandem, interpreting environmental signals and executing decisions with limited human oversight. This digital shift in manufacturing and commerce, marked by the introduction of sophisticated machinery and tools, promises enhanced manufacturing productivity, heightened resource efficiency, and waste curtailment [2, 3].

However, there's a flip side. Industrial automation could amplify resource and energy use and exacerbate pollution [4, 5]. Socially, this wave of digital evolution and industrial overhaul could upheave job landscapes. Many surmise that while low-skill jobs may vanish (around 25% of all jobs), emergent technologies like smart robotics, autonomous transport, and cloud platforms could spawn numerous roles in areas like automation engineering, control systems, machine learning, and software design [6-8]. The realization of Industry 4.0 hinges on its technological viability and societal acceptance, given the shifting socio-economic constructs [9] that come with both opportunities and challenges [10]. Anticipated benefits include enhanced organizational control, real-time performance metrics [11], and a spike in global competitive standing [12, 13]. Furthermore, the investment costs for Industry 4.0 technology are predicted to wane over time [9, 14]. Real-time monitoring characteristic of Industry 4.0 is set to slash unexpected machinery breakdowns, boosting productivity and trimming expenses [15].

Figure 1 delineates the adoption rate of Industry 4.0's technological components within Brazil's steel sector, a rate determined by the pursuit of superior process efficiency, productivity, and enhanced production metrics.

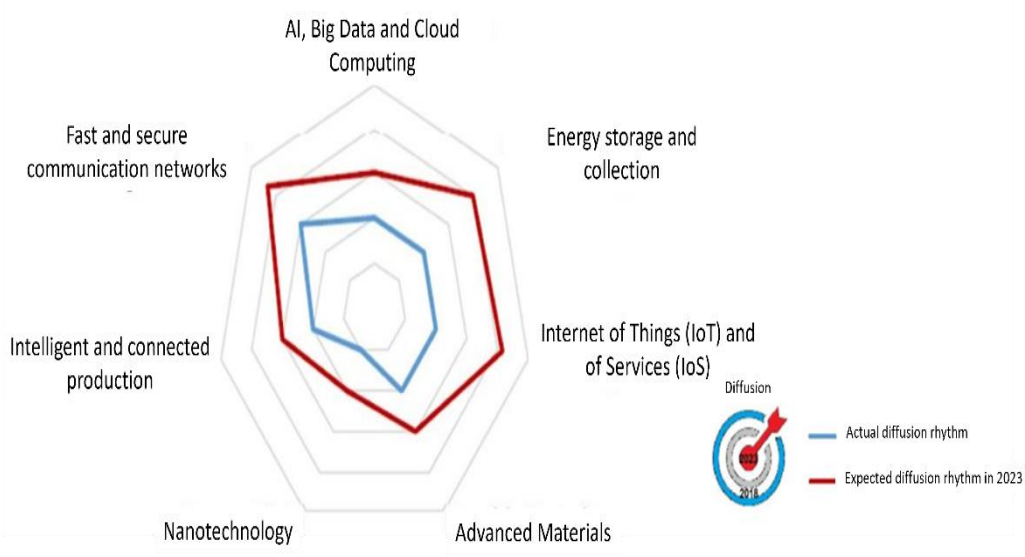

Figure 1. *Diffusion rhythm of technological clusters in the Brazilian steel industry [16].*

2. Literature Review

2.1 Steel Production in LD Converters

The steel sector experienced significant advancements in oxygen melting technology during the 1950s and 1960s. Notably, the inaugural LD Melting Plant began operations in Linz, Austria. In South America, Belgo Mineira Co. pioneered this technique in 1957. Over the decades, its evolution has been marked by substantial process automation [17]. The LD process introduction heralded numerous benefits like superior steel quality, consistent result reproducibility, and shorter operational durations. Originating from Linz-Donawitz and popularly known as LD, this process accounts for roughly 77% of Brazil's steel output. Its widespread adoption in Brazil and globally stems from its high efficiency, affordability, and metallurgical flexibility. Moreover, it can effectively produce an extensive range of steel types [17, 18]. While the core principles of the process have remained consistent, innovations have been witnessed in furnace capacity, design, support mechanisms, control of tilting devices, oxygen blowing nozzles, sub-lance enhancements, and advancements in operational and automation techniques.

2.2 Digitalization of the Brazilian steel industry

The vast amount of computational power and data required for training the models used in this work presents a challenge for the industry at its current stage in Brazil. Practical implementation of variable forecasting for ends of process control would depend upon online measurement of the variables, requiring full digitalization of the production facilities.

Brazilian steel industries are not significantly behind the global pace when it comes to the implementation of Industry 4.0 technologies, but such mature sectores tend to be more precautious when implementing disruptive innovations. The dissemination of digitalization technologies in brazilian steel mills is still held back by technical and economic barriers such as the uncertainty of financial return of projects, the absence of government incentives, the low levels of professional qualification in Latin America, the inadequate curriculum of local educational institutions with an relating to the challenges of Industry 4.0 and the reluctance of the corporations to share data [19].

2.3 Time Series Analysis

The modern era has witnessed an exponential growth in sensor deployment, whether for enhancing human convenience or monitoring industrial operations. This proliferation leads to vast data generation. Estimates from 2020 forecasted that 328.77 million terabytes of data are created every day in 2023 [20]. To derive meaningful insights from this vast data pool, advanced data analytical tools are imperative. A particular subset of this data, time series data, is of utmost significance. Continuous monitoring and data acquisition have become standard practices, necessitating the evolution of effective time series analysis employing statistical analysis, machine learning, or a hybrid of both [21]. Time series analysis, at its core, seeks to extract significant summaries and statistical patterns from chronologically arranged data points. Unlike traditional predictive models, time series emphasizes the significance of data order. Its primary goal is to analyze historical patterns and make informed future predictions, accounting for event sequences [21-23]. A landmark contribution to time series analysis is the Box-Jenkins method, first elucidated in 1970, applied to CO_2 emission levels from a gas blast furnace dataset [24].

2.4 Exploratory Time Series Methods

Any dataset's primary investigation involves an Exploratory Data Analysis (EDA) [25]. While confirmatory analysis leans on formal statistical models and inferences, EDA emphasizes comprehensive data understanding, error detection, hypothesis validation, variable correlation identification, and appropriate model selection [26].

2.5 Characterizing Time Series

Unique to time series data is the evaluation of values across different time points within a series. Stationarity is a crucial time series feature, indicating consistent data shifts. Assessing stationarity levels is vital as it reflects the system's long-term historical behavior vis-à-vis its anticipated future behavior [21].

2.6 Statistical Models for Time Series Evaluation

For efficacious characterizations and predictions, the appropriate model selection is crucial. Several statistical methods cater explicitly to time series data. Initially, one should grasp univariate time series data methods, progressing to more intricate multivariate models

(NIELSEN, 2021). An example of a non-stationary time series is defined in Figure 2, by a set of passenger data over the years.

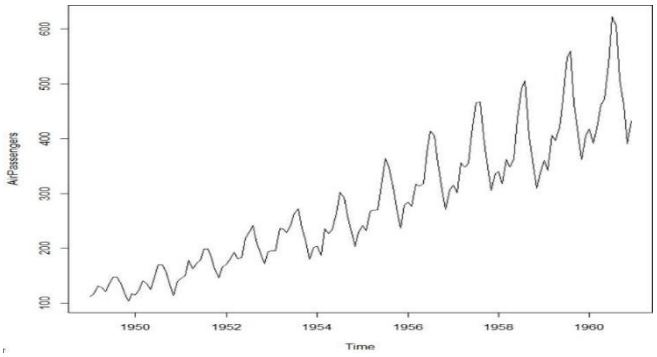

Figure 2. *Example of non-stationary time series [21].*

2.7 Deep Learning for Time Series

For certain data types, deep learning outperforms traditional machine learning, given its ability to process vast datasets without making assumptions. In this research's context, the Long Short-Term Memory neural network (LSTM), a leading deep learning technique for sequential data, proves invaluable. Initially crafted for natural language processing, LSTM's capabilities extend to effective time series modeling.

2.8 Long Short-Term Memory Neural Networks (LSTM)

LSTM, a subset of the Recurrent Neural Network (RNN), is renowned in artificial intelligence and deep learning realms. Its prowess lies in processing not just individual data points but entire data sequences, akin to time series [27].

3. Case and Methodology

This work consists of identifying forecasting models that can model the operational variables of an LD furnace as a function of its life age. It is considered three campaigns, named A, B, and C, with high, medium, and low performance, respectively. The age of the kiln is measured through the internal refractory coating. The wear rate of the refractories was calculated using laser measurement intervals, performed using the FERROTRON™ system (Figure 3). This equipment scans and determines the thickness of the refractories throughout the campaign, and the wear rate is defined as thickness reduction by the number of runs.

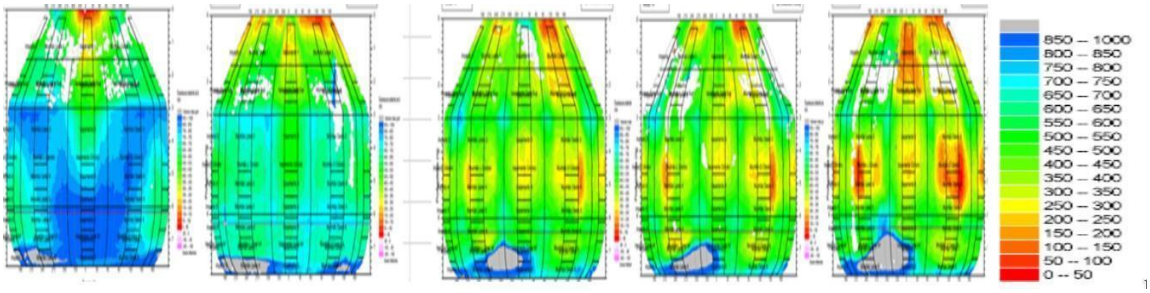

Figure 3. *Thickness (in mm) measured by laser scans over the lifetime of the LD using the FERROTRON™ equipment. Source: Own authorship.*

Along with the wear rate data, several operational variables were collected, taken from the databases of the production information systems of the steel shop of an integrated steel plant (via the Access database management program). The data deemed important in the performance process (life) of the melt shop's LD converter refractories were selected for the analysis of this work, according to the literature and experience of analysts and specialists in the operational area. The three campaigns were performed in different periods, between 2014 and 2018. The choice of the periods mentioned above considered the guarantee that the melt shop was at a production pace within certain stability parameters, ensuring that there were no large abnormal stoppage events and abnormal restrictions on the day-to-day activities of the plant.

The variables selected for the study were initially:

- cal total: Total amount of quicklime (CaO) added
- cal after blow: Amount of quicklime added after blowing
- cdas with lime: Indicates if quicklime was added in the run
- total dolca: total calcined dolomite (CaOMgO) added
- dolca after blowing: Calcined dolomite added after blowing
- cdas with calcined dolomite: Indicates if calcined dolomite was added in the run
- total dolcr: Total amount of raw dolomite ($CaMgCO_3$) added
- dolcr before blowing: Total amount of raw dolomite added before blowing
- dolcr after blowing: Total amount of raw dolomite added after blowing
- cdas with raw dolomite: Indicates if raw dolomite was added in the run
- total calc: Amount of added limestone
- cdas with limestone: Indicates if limestone ($CaCO_3$) was added in the run
- cfs: End of blow carbon
- ubc: Indicates if the run is an ultra-low carbon (IF- Interstitial free) steel type
- mgo: Percentage of MgO in the slag
- fet: Percentage of total Fe in the slag
- cao: Percentage of CaO in the slag
- sio2: Percentage of SiO_2 in the slag
- bb: Binary basicity (CaO/SiO_2)
- rh: Indicates if the run was destined for the vacuum degasser (RH) treatment
- blow: Indicates if there was blow in the race
- scrap: Quantity of metallic scrap by weight
- scrap %: Percentage of scrap in the metallic charge loaded in the furnace
- pig iron %: percentage of pig iron of the total load loaded in the furnace
- return: Number of rides returned to BOF
- clean pig iron: Amount of pig iron after removing the slag formed in the desulfurization of the pig iron
- si pig iron: Amount of silicon (Si) in the pig iron
- steel: Amount of steel produced
- bath mat: Amount of material added to make the molten bath (steel/scrap…)
- slag splashing bool: Percentage of slag splashing performed
- tfs: End of blowing temperature
- fsop inivaz: Time between the end of blowing and the beginning of steel tapping
- t vaz: Tapping time
- vol o2: O_2 volume
- bas: Basicity
- hole life: Life of the tapping hole of runs
- c bs: Carbon from the BS measurement, near the end of the blowing
- c cs: Carbon from the CS measurement, at the end of the blowing
- ox3: Final heat oxidation level
- alt sole: Measured height of the sole in the bottom of the oven

• %txdp: Furnace dephosphorization rate.

More specific time series analyses were performed using autocorrelation, partial autocorrection, and stationarity tests. After data characterization and linking, some treatments were performed on them to meet the requirements of each model used, such as the application of the Kalman filter to smooth out outliers and the application of time series differentiation to make them stationary (see at figure 5). Then, modeling of the time series was carried out considering the variables that most influenced the time series. Finally, the performance of each constructed model was analyzed to predict which one would be more successful in evaluating the behavior of operational data throughout the considered life of the LD converter. The following metrics were used to measure the performance of the built models:

• Mean square error (MSE)
• Mean Absolute Error (MAE)
• Root Mean Square Error (RMSE)
• Mean absolute percentage error (MAPE) Coefficient of Determination R^2 (R2)

4. Results

Initially, all variable time series were analyzed individually, split by campaigns, to verify if there is any standard behavior between campaigns or any supposed correlation between the trends of the series. It is worth mentioning that the graphs were generated with groups of 24 points averaged between them, once the data is very irregular and this facilitates visualization. Thus, the following images demonstrate the value of the variable as a function of the lifetime of the converter. Figure 4 shows 9 graphs of the time series.

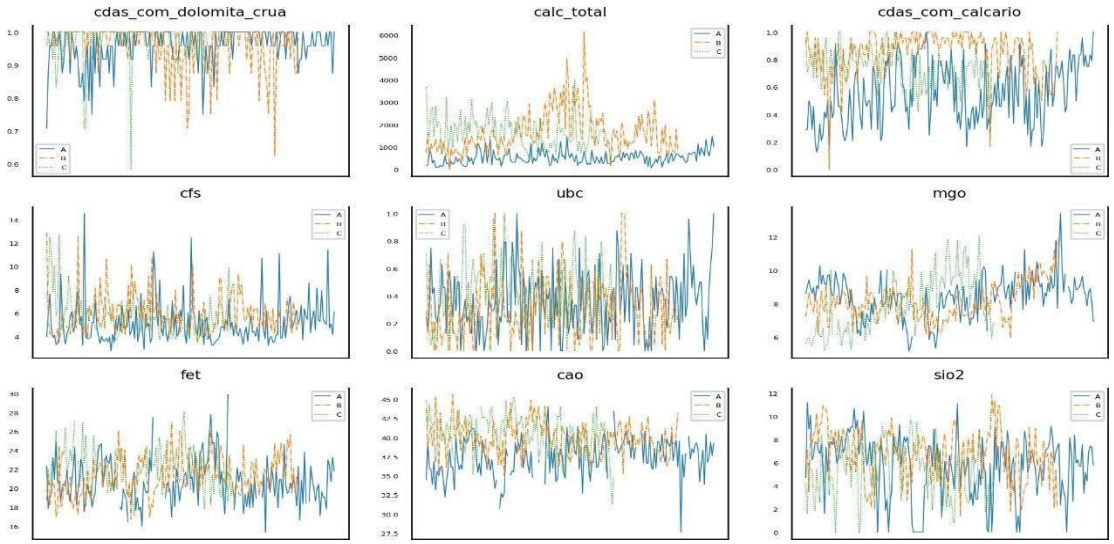

Figure 4. *Graphs of time series of 9 parameters chosen as a function of the life of the BOF.*

The next step was the stationarity analysis of all time series according to the Augmented Dickey-Fuller (ADF) Test, Kwiatkowskie-Phillips-Schmidt-Shin (KPSS) Test, and Phillips-Perron stationarity tests. Time series were considered stationary if, and only if, the three tests had results proving stationarity. Thus, if there is any result that indicates that the series is not stationary, it will be considered non-stationary. The stationarity tests performed considered a significance level of 5%. Using the smoothed data, stationarity tests were applied to each of the 3 different campaigns: A, B, and C.

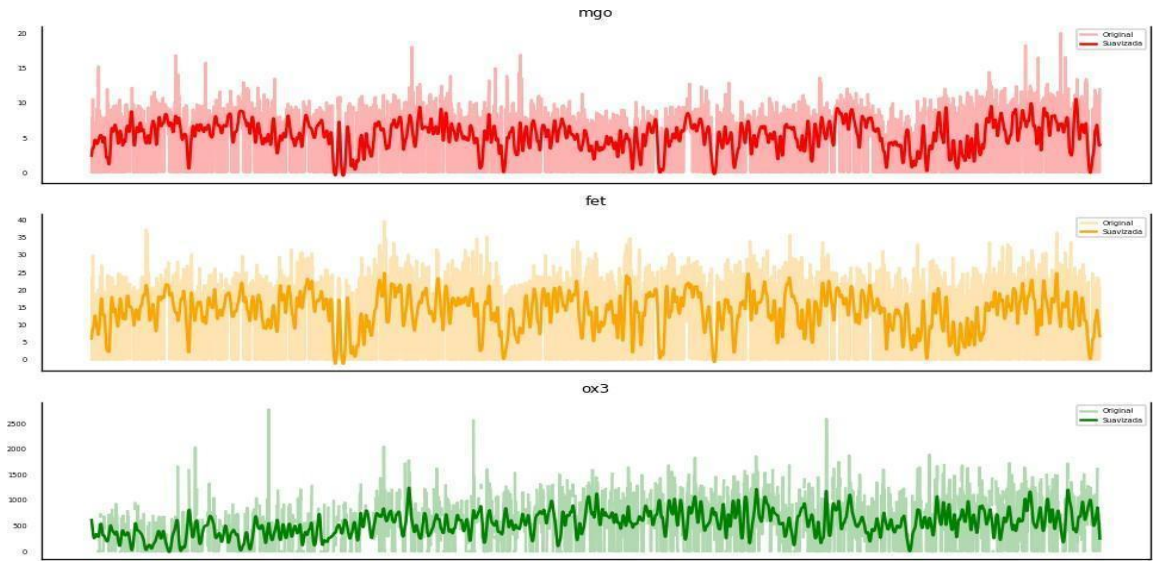

Figure 5. *Example of Kalman smoothing for the "mgo", "fet" and "ox3" time series.*

It was possible to reach stationarity for all time series from the data differentiation. Therefore, it was not necessary to differentiate the data again. With the data properly differentiated, cleaned, and prepared, the data modeling is initiated to obtain a model that describes the behavior of these time series. Initially, the vectorial autoregression model is used, as it is the simple and most efficient model in real situations of variable relationships. To obtain the best VAR model, a p-lag order selection algorithm must be used first. Thus, an available algorithm with a maximum lag parameter equal to 30 was applied to obtain the best coefficient p within this range through criteria evaluation of the AKAIKE Information Criterion (AIC), BAYESIANA Information Criterion (BIC), Final Prediction Error (FPE) or Complete Percentile Error, and HANNA-QUINN Information Criterion (HQIC) of each series in each campaign.

After adjusting the data, two LSTM models were built and named LSTM1 and LSTM2. Figure 6 describes the LSTM models using time series for 3 fluxes variables, such as: quick lime, total dolomite amount and dolomite add after oxygen blowing.

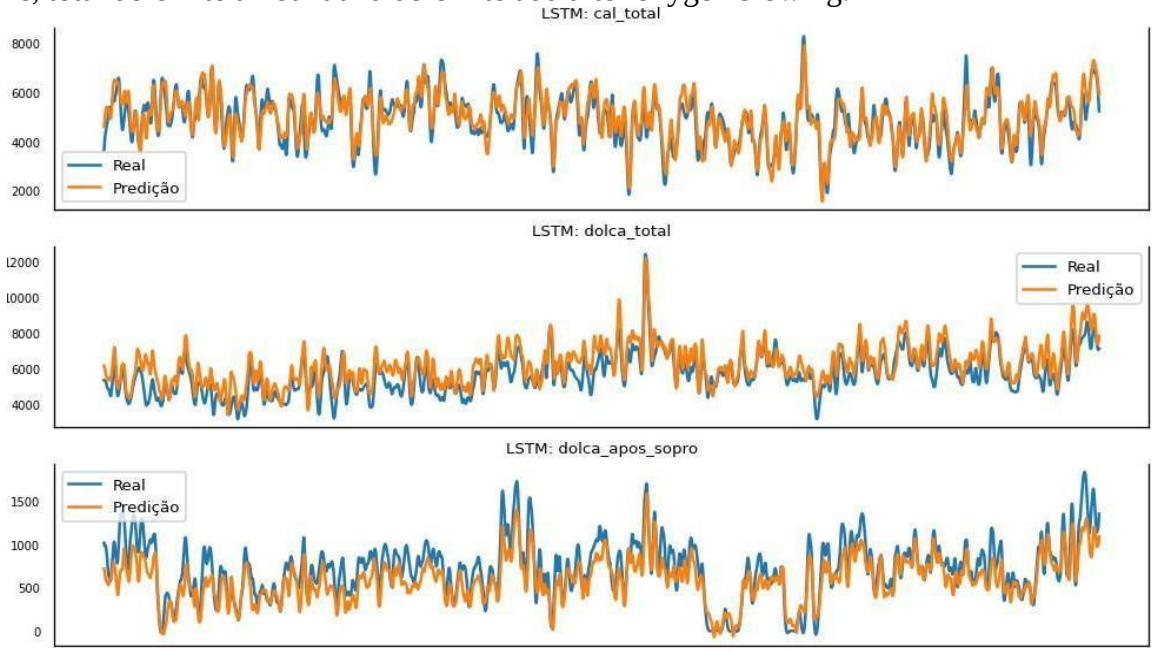

Figure 6. *Results of the models using time series for the variables: Total amount of quicklime added (total quicklime), total amount of dolomite added (total dolca), and amount of dolomite used after blowing (dolca after blowing).*

The difference between the models is that the first one contains two layers of LSTMs, while the second one contains four layers, one layer being the encoder and the decoder for the first case and two layers for each function in the second model. Finally, additional layers are stacked in the encoder part and decoder part for the sequence model. By stacking LSTMs, it is possible to increase the model's ability to understand more complex representations of the data. During the training of the model, the Adam optimizer and the Huber penalty function were used, with data from two campaigns used for training and one for test/prediction. Figure 7 presents the tabulated results of the LSTM1 model after adjusting the variables and the distribution of the R2 values of the different models studied, in which the LSTM1 model, with adjustment of the variables, showed the best R2 adjustment.

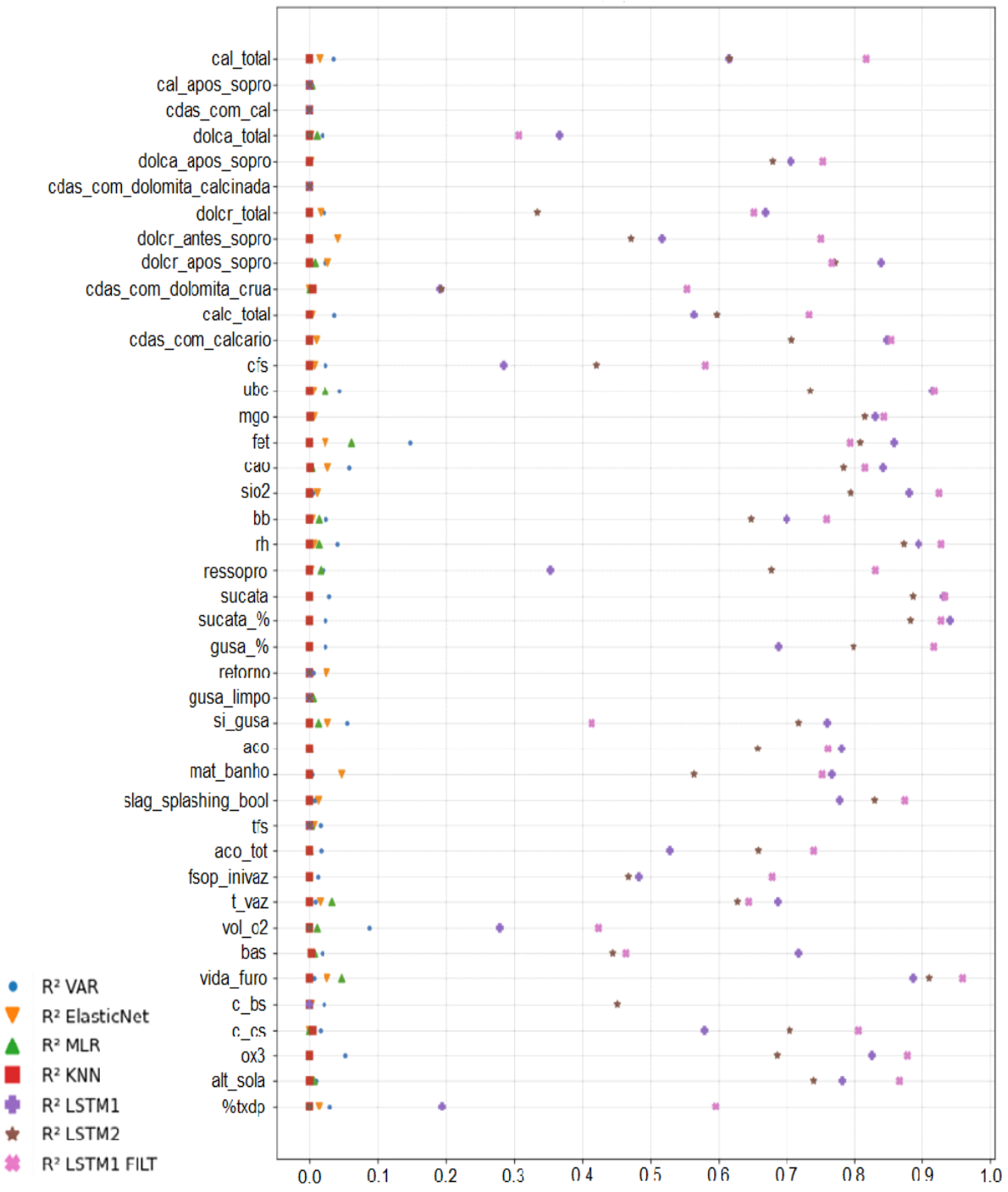

Figure 7. *Comparative chart of the R² of the variables between models.*

5. Discussion

The simple visual analysis of the data over time did not bring conclusions about the refractory wear process since several variables present non-linear and irregular behavior over the period. The first model built (multivariate time series model), the VAR model, was also not sufficient to represent the data due to the complexity of interactions between variables, even though this is one of the most used models for multivariate prediction of time series. The model did not manage to capture any of the intricate interactions between the variables, performing very poorly for all campaigns, with no significant difference between them (the highest R2 between all three campaigns was 0.09). This, although frustrating at the beginning, can be explained by the linear assumption done for the VAR model and also the size of the data sample, which is still small in order to predict the high amount of variables involved in the process. The following three models evaluated: ElasticNet, MLR, and KNN, with applications not necessarily specific to time series, also did not obtain adequate R2 correlation coefficients. The ElasticNet model, based on Lasso and Ridge regressions and merging their penalty functions has diverse applications. This model has a good ability to eliminate less important variables, managing to receive various information as input and discern what is relevant for a good prediction. It, however, proved to be ineffective in the studied high-dimensional scenario, most likely due to the highly covariant relationships between the variables in the system. The MLR model was chosen to test if the data set could be modeled using multiple linear regression, even without apparent linearity. Analogous to the previous case, the unsatisfactory performance of the MLR model was noticed to predict the data set studied. The KNN model is based on the distance between observations and groups them, making predictions based on these groupings of similar observations. This model struggled to represent the data due to data variability, which affected distance based algorithms. In this way, it was no longer possible to model the presented data set adequately, being necessary to start with more complex and robust models. Thus, the most suitable model for modeling sequential data, such as time series with data from different refractory campaigns, was the LSTM model, a type of recurrent neural network composed of cells with input, output, and forgetting ports. Because it is a model capable of memorizing information in the long term, it may have obtained better performance by being able to store singular correlations for longer periods, thus, obtaining better information for more accurate predictions. LSTM can produce better results than parameterizable models and common Recurrent Neural Networks (RNNs) when dealing with complex autocorrelation sequences (long memory), large data sets, and when the probability distribution of the basilar process is unknown or not replicable using methods standard parameters such as Autoregressive Integrated Moving Average (ARIMA).

6. Conclusion

In this work, it was possible to achieve a model with satisfactory performance to predict the behavior of multiple time series. With the help of data preparation techniques, it was possible to transform a data set of 42 highly irregular and volatile variables into a set with a greater chance of being modeled. Thus, the Vectorial Autoregression, ElasticNet, K-Nearest-Neighbors, Multiple Linear Regression, and long-term and short-term memory neural networks were applied, where only the latter one had considerable performance in predicting the time series. The model with the best performance achieved a good degree of predictive reliability for 35 variables, seven variables with R2 above 0.9, nine variables with R2 between 0.8 and 0.9, nine variables with R2 between 0.7 and 0 .8, and all others above 0.3.

It is recommended for future work to include new important process variables, such as the refractory wear rate, which was not possible in the present work due to a lack of data. It is also recommended to apply new models to the data set, aiming to improve the prediction of

variables that did not perform well, such as GARCH models and other types of neural networks. Furthermore, to obtain models closer to reality, it is always recommended to use a larger amount of data for training, also including datasets with different behaviors.

References

1. Ghobakhloo, M., *Industry 4.0, digitization, and opportunities for sustainability.* Journal of cleaner production, 2020. **252**: p. 119869.
2. Tortorella, G.L. and D. Fettermann, *Implementation of Industry 4.0 and lean production in Brazilian manufacturing companies.* International Journal of Production Research, 2018. **56**(8): p. 2975-2987.
3. Contador, J.C., et al., *Flexibility in the Brazilian industry 4.0: Challenges and opportunities.* Global Journal of Flexible Systems Management, 2020. **21**(Suppl 1): p. 15-31.
4. Beier, G., et al., *Sustainability aspects of a digitalized industry–A comparative study from China and Germany.* International journal of precision engineering and manufacturing-green technology, 2017. **4**: p. 227-234.
5. Liu, X. and J. Bae, *Urbanization and industrialization impact of CO2 emissions in China.* Journal of cleaner production, 2018. **172**: p. 178-186.
6. Brougham, D. and J. Haar, *Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace.* Journal of Management & Organization, 2018. **24**(2): p. 239-257.
7. Frey, C.B. and M.A. Osborne, *The future of employment: How susceptible are jobs to computerisation?* Technological forecasting and social change, 2017. **114**: p. 254-280.
8. *WEF (World Economic Forum) (2023). Future of Jobs Report 2023. .*
9. Kovacs, O., *The dark corners of industry 4.0–Grounding economic governance 2.0.* Technology in society, 2018. **55**: p. 140-145.
10. Weber, K.M., N. Gudowsky, and G. Aichholzer, *Foresight and technology assessment for the Austrian parliament—Finding new ways of debating the future of industry 4.0.* Futures, 2019. **109**: p. 240-251.
11. Horváth, D. and R.Z. Szabó, *Driving forces and barriers of Industry 4.0: Do multinational and small and medium-sized companies have equal opportunities?* Technological forecasting and social change, 2019. **146**: p. 119-132.
12. Theorin, A., et al., *An event-driven manufacturing information system architecture for Industry 4.0.* International journal of production research, 2017. **55**(5): p. 1297-1311.
13. Gružauskas, V., S. Baskutis, and V. Navickas, *Minimizing the trade-off between sustainability and cost effective performance by using autonomous vehicles.* Journal of Cleaner Production, 2018. **184**: p. 709-717.
14. Cunha, A., et al. *Sustainable manufacturing: the impact of collaboration on SMEs*. in *2018 International Conference on Intelligent Systems (IS)*. 2018. IEEE.
15. Long, F., P. Zeiler, and B. Bertsche, *Modelling the flexibility of production systems in Industry 4.0 for analysing their productivity and availability with high-level Petri nets.* IFAC-PapersOnLine, 2017. **50**(1): p. 5680-5687.
16. Teixeira, R.L.P., et al., *Os discursos acerca dos desafios da siderurgia na indústria 4.0 no Brasil.* Brazilian Journal of Development, 2019. **5**(12): p. 28290-28309.
17. Chaves, A.J.M., *Avaliação do desempenho operacional de um convertedor LD através do desenvolvimento do processo de sopro com lança de quatro furos.* 2006.
18. *IAB (Instituto Aço Brasil) (2023). Estatísticas da Siderurgia 2023 3º Trimestre*
19. Martins, M.S., G.M. de Paula, and M.d.R.A. Botelho, *Inovações tecnológicas e indústria 4.0 na siderurgia: difusão, estrutura de mercado e heterogeneidade intrassetorial.* Revista Brasileira de Inovação, 2021. **20**: p. e021006-e021006.

20. *STATISTA (2023). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*
21. NIELSEN, A., *Análise Prática de Séries Temporais–Predição com Estatística e Aprendizado de Máquina*. 2021, Rio de Janeiro: Alta Books.
22. Ehlers, R., *Análise de séries temporais. Laboratório de Estatística e Geoinformação.* Universidade Federal do Paraná, 2007.
23. Morettin, P.A. and C.M. Toloi, *Análise de séries temporais: modelos lineares univariados*. 2018: Editora Blucher.
24. Box, G.E., et al., *Time series analysis: forecasting and control*. 2015: John Wiley & Sons.
25. Monteiro, A.C.P., *Análise de Séries Temporais de Dados Meteorológicos da Cidade do Porto.* 2021.
26. Data, M.C., et al., *Exploratory data analysis.* Secondary analysis of electronic health records, 2016: p. 185-203.
27. Sherstinsky, A., *Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.* Physica D: Nonlinear Phenomena, 2020. **404**: p. 132306.