Machine Learning Methods for Forecasting TRS Emissions in the Pulp and Paper Industry

Thommas K. S. Flores [*] **Willians P. Dutra** [**] **Ricardo L. Durski** [***]
Juan M. M. Villanueva [****] **André C. M. Cavalheiro** [†]

[*] *Universidade Federal do Rio Grande do Norte, RN, (e-mails: thommas.flores.101@ufrn.edu.br)*
[**] *Escola Superior de Agricultura Luiz de Queiroz, São Paulo-SP, (e-mails: willians.dutra@prosys.tec.br)*
[***] *Faculdade de Informática e Administração Paulista, São Paulo-SP, (e-mails: ricardo.durski@prosys.tec.br)*
[****] *Universidade Federal da Paraíba, João Pessoa-PB, (e-mail: jmauricio@cear.ufpb.br)*
[†] *Universidade de São Paulo, São Paulo-SP, (e-mail: ext.andre.cavalheiro@prosys.tec.br)*

Abstract: This study aimed to evaluate the performance of various time series forecasting algorithms for predicting Total Reduced Sulfur (TRS) emissions in the pulp and paper industry. The analysis was conducted using a real time series dataset, where 25 samples autocorrelation function was computed to identify the level of correlation among the variables lags. The results showed that the last four previous TRS values had a total autocorrelation above 0.8. Additionally, the partial correlation function indicated that only the last previous TRS value had a correlation value above 0.3. The comparison of the performance of different time series algorithms, including XGBRegressor, LSTM, CNN, MLP-MHA, MLPRegressor, and ARIAMA, was made based on two different horizons (1 and 8). The evaluation metrics used were MAE, MSE, R^2-score, and execution time. The results indicated that ARIAMA outperformed the other algorithms on horizon 1, with an MAE of 0.0565 and R^2-score of 0.7046. On the other hand, LSTM had the best performance on horizon 8, with an MAE of 0.0853 and R^2-score of 0.5241. These findings suggest that advanced time series prediction algorithms can provide more accurate models for predicting TRS emissions in the pulp and paper industry, contributing to environmental and health mitigation efforts. Finally, the use of these models can help the industry comply with regulatory standards and avoid penalties while also supporting its economic sustainability.

Keywords: Time series forecasting; TRS; Pulp and paper industry; Burning NCG; Environmental sustainability

1. INTRODUCTION

The production of pulp and paper involves various processes that consume significant amounts of energy and chemicals, resulting in the release of various Greenhouse Gas (GHG) emissions such as sulfur dioxide (SO_2), particulate matter (PM), carbon monoxide (CO), carbon dioxide (CO_2), ozone (O_3), nitrogen oxides (NOx), and hydrocarbons (HC) from different sources like combustion of fossil fuels and decomposition of organic matter.

According to a study by dos Santos et al. (2021), the pulp and paper industry in Brazil was responsible for 4.6% of the country's total GHG emissions in 2017. The study also shows that the combustion of fossil fuels for energy generation is a major contributor to GHG emissions in the industry. However, the Brazilian government has implemented several regulations to mitigate GHG emissions from the industry. For instance, the Brazilian National Climate Change Policy (PNMC) established a target to reduce GHG emissions from the industry by 5.8 million tonnes of CO_2 equivalent by 2020. Additionally, the Brazilian Forest Code requires the preservation and restoration of native vegetation, which can help to offset GHG emissions from the industry.

Several studies have shown that implementing energy-efficient technologies and using renewable energy sources can significantly reduce GHG emissions in the pulp and paper industry. For example, the study by Ferreira et al. (2019) found that using biomass-based energy sources can reduce GHG emissions by up to 96% compared to using fossil fuels. Moreover, the study by Silva et al. (2020) showed that implementing energy-efficient technologies in the pulp and paper industry can reduce GHG emissions by up to 43%.

Additionally, machine learning techniques have been used to predict and mitigate emissions from pulp mills, as shown by Jafari et al. (2019), who accurately predicted sulfur dioxide emissions and identified key influencing variables using these techniques. By combining these approaches,

the pulp and paper industry can take a multi-faceted approach to reduce their environmental impact and contribute to sustainability efforts.

In Wen et al. (2021), was used machine learning to model the impact of operating parameters on greenhouse gas emissions from pulp and paper mills. Their results showed that machine learning techniques could help optimize operating parameters to reduce emissions.

This paper proposes methods to improve the prediction of Total Reduced Sulfur emissions in the pulp and paper industry using advanced time series forecasting algorithms such as XGBRegressor, ARIMA, LSTM, CNN, MLP-MHA and MLPRegressor. The study compares the performance of these models on two different horizons and evaluates them based on several metrics, including MAE, MSE, and R^2-score.

Considering the contributions of the mentioned work, it is possible to highlight: 1) Analysis of the influence of lags values as parameters for estimating future values; 2) Comparison of the performance of different time series prediction algorithms; and 3) Discussion of the environmental and economic impact on the paper and pulp industry.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of some forecasting methods, while Section 3 describes the Recovery Boiler Incinerator. In Section 4, the experimental procedures are outlined, and Section 5 presents and discusses the results. Finally, Section 6 concludes the paper.

2. FORECASTING METHODS

Choosing the right model for time series forecasting is crucial for precise and valuable predictions. Thus, comparing diverse models becomes necessary to identify the best-suited one for a specific problem. In this article, we compared six methods for time series forecasting. This evaluation enables us to determine the most appropriate model for the given problem and optimize the forecasting outcomes.

2.1 XGBRegressor

XGBRegressor is a decision tree-based algorithm that uses Gradient Boosting Decision Tree (GBDT) as a machine learning technique. GBDT is an ensemble technique that combines multiple weak decision trees to form a stronger model. Each tree is trained to correct the errors of the previous trees Friedman (2001).

The XGBRegressor uses a stochastic gradient optimization method to optimize the loss function in order to improve model accuracy. It uses L1 and L2 regularization to prevent overfitting.

The goal of XGBRegressor is to minimize the loss function through a stochastic gradient optimization method. To do this, it uses the gradient descent algorithm to adjust the model parameters.

2.2 ARIMA

The autoregressive integrated moving average (ARIMA) model is a widely used time series forecasting technique. The ARIMA model consists of three components: the autoregressive component (AR), the moving average component (MA), and the differentiation component (I).

The AR component is responsible for modeling the serial dependence of the time series on its past values. The MA component is used to capture the random fluctuations of the time series. The I component is used to transform the time series into a stationary series, which facilitates the modeling and forecasting of the series.

The ARIMA model can be represented as ARIMA(p, d, q), where p is the order of the AR component, d is the number of differentiations required to make the time series stationary, and q is the order of the MA component. The ARIMA modeling process involves selecting the appropriate values of p, d, and q for the time series.

2.3 CNN

Convolutional Neural Networks (CNN) are a class of neural networks that use convolutions to process input data, primarily in computer vision applications. CNNs are composed of convolutional filter layers followed by pooling layers to reduce dimensionality, and finally output layers (LeCun et al., 2015).

The main difference between CNNs for computer vision and CNNs for time series is that CNNs for time series use one-dimensional convolutional filters instead of two-dimensional convolutional filters (Barino and dos Santos, 2020). In the convolutional layer, a set of convolutional filters is applied to the input time series. Each convolutional filter generates a new time series that represents the convolution of the input series with the filter. These new time series are then processed by an activation layer, such as the ReLU (Rectified Linear Unit), to introduce nonlinearity into the network (Goodfellow et al., 2016).

After the convolutional layer processes the input time series, the pooling layer is used to reduce the dimensionality of the data. The pooling layer operates on sliding windows of the time series and applies an aggregation function, such as average or maximum, to reduce the dimensionality of the data (Barino and dos Santos, 2020).

2.4 LSTM

The LSTM is a recurrent neural network consisting of several layers of memory cells, each with three gates: input gate, forget gate and the output gate. The input gate determines which information from the current input should be stored in the memory cell. The forget gate decides which information should be discarded from the memory cell. The output gate determines which information from the memory cell should be used in the current output (Van Houdt et al., 2020).

To train the LSTM model, the error backpropagation technique is used. For time series prediction, LSTM uses a sliding window, where for each prediction the window is shifted one position forward, and the next observation is added to the window. This process is repeated until all observations in the time series are predicted.

2.5 MLP Multi-head attention

The architecture of MLP-MHA consists of two main components: a Multi-Head Attention (MHA) layer and an Multi-Layer Perceptron (MLP) layer. The Multi-Head Attention layer allows the model to attend to different parts of the input sequence simultaneously and capture complex relationships between the input and output. The MLP layer processes the output of the Multi-Head Attention layer and learns to predict the target variable Vaswani et al. (2017).

At its core, Multi-Head Attention addresses the imperative of targeted allocation of focus to specific segments within input sequences during computational processing. This attribute gains paramount importance in time series forecasting, where prolonged sequences and intricate temporal dependencies demand nuanced analytical dissection. Diverging from the conventional single-attention counterparts, the Multi-Head Attention paradigm employs a collective of attention heads, each dedicated to discerning distinct temporal relationships encapsulated within the data stream.

Multi-Head Attention orchestrates through phases. It starts with linear transformations, assigning distinct query, key, and value attributes for each attention head. The scaled dot-product attention operates in isolation across these parts, computing weighted summations based on query-key correlations. This generates an attention score matrix that encodes temporal relationships. Subsequent stages involve concatenation, linear projection to harmonize attention outputs. The synthesized result enters later neural network layers, amalgamating diverse representations from various attention heads. Multi-Head Attention enhances time series prediction by adeptly harmonizing temporal perspectives for improved efficacy.

The attention computation follows, where attention scores are calculated between the queries and keys. The attention scores measure the relevance or importance of each historical observation for predicting the future values. The attention scores are computed using a dot-product similarity measure, scaled by the square root of the dimension of the key vectors.

2.6 MLPRegressor

MLPRegressor is an MLP (Multilayer Perceptron) type artificial neural network model used for regression problems, where the goal is to predict a continuous value based on a set of inputs Mehmood et al. (2019). This algorithm uses the error backpropagation technique to adjust its parameters, which involves calculating the error between the output predicted by the model and the actual value of the time series. This error is then propagated back through the neural network to adjust the synaptic weights. The back-propagation process is repeated several times until the error is minimized.

3. RECOVERY BOILER INCINERATOR

In the pulp and paper industry, Methanol (CH_3OH), Dissolved Non-Condensable Gases (DNCG), Concentrated Non-Condensable Gases (CNCG) and Liquefied Petroleum Gas (LPG) in a Non-Condensable Gas (NCG) incinerator in the pulp and paper industry is a common practice to treat the gases generated in the production process, as illustrated in Figure 1.

DNCG and CNCG gases are generated during the boiling of wood in pulp and paper production, while methanol is obtained from the recovery of these gases by distillation and condensation. LPG is obtained externally. The burning of these gases can result in the emission of various air pollutants, such as nitrogen oxides (NOx), sulfur dioxide (SO_2), carbon monoxide (CO), carbon dioxide (CO_2), and reduced sulfur compounds (TRS), among others Wang et al. (2019).

Several factors can influence TRS emissions from CNCG incineration. One important factor is the incineration temperature, as studies have shown that increasing the temperature can reduce TRS emissions. For instance, Yu et al. (2015) found that increasing the CNCG incineration temperature from 800°C to 900°C resulted in a 50% reduction in TRS emissions. The sulfur content of the CNCGs being burned is another important factor that can influence TRS emissions, with higher sulfur content leading to higher emissions. Therefore, predicting to controller the sulfur content of the gases being fed into the incinerator is essential to minimize TRS emissions.

In this paper, 6 algorithms for time series prediction (XGBRegressor, ARIMA, LSTM, CNN, MLP-MHA and MLPRegressor) are analyzed to predict TRS emission time series in pulp and paper industry. For this, the following steps are considered: data collection, data preprocessing, correlation analysis, model training and evaluation.

4. METHOD

This section provides a detailed account of the experimental procedures. It is organized into three parts. The first part offers an in-depth overview of the dataset, including its preprocessing and the evaluation metrics used in the study. The second part outlines the models' architecture and the experimental parameter settings adopted for the comparative analysis. The third part presents the experiment's outcomes through visual representations of the prediction results obtained from the proposed models and the comparison results.

4.1 Data collection

For the present case study, one year of data from a Brazilian pulp and paper industry was considered. The TRS values were collected every hour through instrumentation located in the tower, as shown in Figure 1.

4.2 Data segregation

During this phase, preprocessed and validated data is partitioned into distinct training and test sets. A fraction of the training data is allocated for model validation. In this study, data was randomly divided, with 30% for testing and 70% for training. Within the 70%, 20% served as validation and 80% for training the model. It's essential to maintain consistent class proportions in resulting subsets.

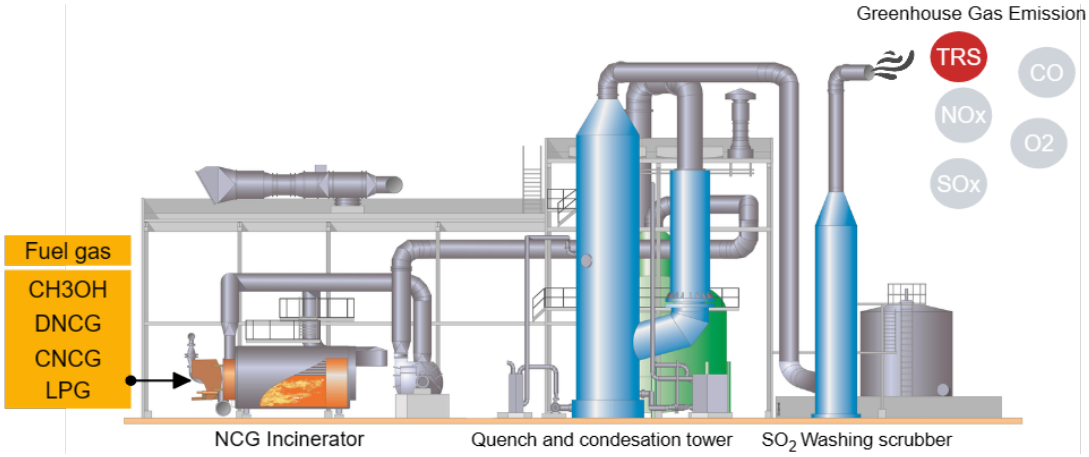


Figure 1. Flue gas scrubbing system focusing on TRS emissions for this study (adapted from Valmet (2019)).

4.3 Model Evaluation Indicators

We used Mean Square Error (MSE) as the evaluation indicators of the model. At the same time, Mean Absolute Error (MAE), MSE and R^2-Score were also used as evaluation indicators for comparison experiments with other models.

4.4 Correlation Analysis

Autocorrelation is a statistical measure that indicates the degree of correlation between a time series and its lagged copies, where values close to 1 or -1 indicate strong positive or negative correlation, respectively. In this study was consider 25 lags.

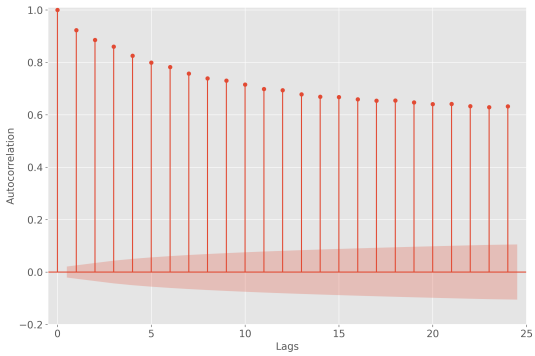
Figure 2a) shows the last 4 past values have total autocorrelation greater than 0.8. While the partial correlation, the last 3 values have values smaller than 0.3, except the $TRS[n-1]$ value (see Figure 2b).

The total autocorrelation can help identify repetitive patterns in the series, while partial autocorrelation can help identify the relationship between the current series and the lagged copies, controlling for the effect of other lagged copies.

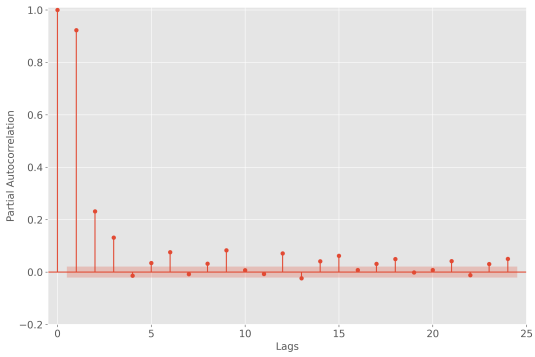
In view of these results, 3 delays were adopted as input to the time series forecasting algorithms for different prediction horizons. This choice is justified due to the statistical significance that these lags have, that is, if the value of the partial or total correlation is less than a threshold value (filled area of the graph), the null hypothesis is rejected and the result is considered statistically significant. This means that the probability of getting the observed result due to chance is very low, and therefore there is likely to be a real difference between the samples.

4.5 Model Architecture Configuration

In this case study the architecture in Figure 3 will be considered, where the inputs to the forecasting algorithm are the past values of the Total Reduced Sulphur (TRS) gas, where d is the maximum delay number and h is the prediction horizon.



(a) Total autocorrelation.



(b) Partial autocorrelation.

Figure 2. Autocorrelation.

Figure 3. Architecture.

In addition, to form the forecasting block in the figure above, the algorithms listed in Table 1, whose parameters were empirically defined, are adopted. In addition, it was adopted that the dimensionality of the algorithms' input depends on the number of lags adopted and the output of the prediction horizon. For the LSTM, CNN and Self-Attention algorithms, 100 epochs, Adam optimizer, mean

square error as loss function and learning rate equal to 0.001 were adopted.

Table 1. ML algorithms and parameters.

Algorithm	Parameters
XGBRegressor	Estimators=1000 Max depth=7 Learning rate=0.001 Random state=0
ARIAMA	p=1, d=0, q=0
LSTM	LSTM(neuron = 64, act=relu,return seq=True) LSTM(neuron=64, act=relu,return seq=True) LSTM(neuron=64, act=relu) Dense(horizon)
CNN	Conv1D(filters=64, kernel size=3, act=relu) MaxPooling1D(pool=1, strides=2, padding=valid) Flatten Dense(horizon)
MLP-MHA	MultiHeadAttention(num heads=5, key_dim=1) Flatten Dense(neuron = 30, act=relu) Dense(Horizon)
MLPRegressor	Hidden layer sizes=(50, 50) Act=relu Solver=adam Max iter=5000

5. RESULTS AND DISCUSSIONS

In this section, the results of the work will be exposed in detail. This way, the time series prediction algorithms proposed in the previous section will be taken into account for the prediction of TRS emissions in an industry of pulp and paper. The algorithm performances were evaluated in the test stage, considering MAE, MSE, R^2 and forecast execution time as evaluation metrics.

5.1 TRS forecasting for horizon equal 1

Figure 4 shows the result for test values for different time series forecasting methods, where 3 lags were considered to predict the next future value of TRS emission.

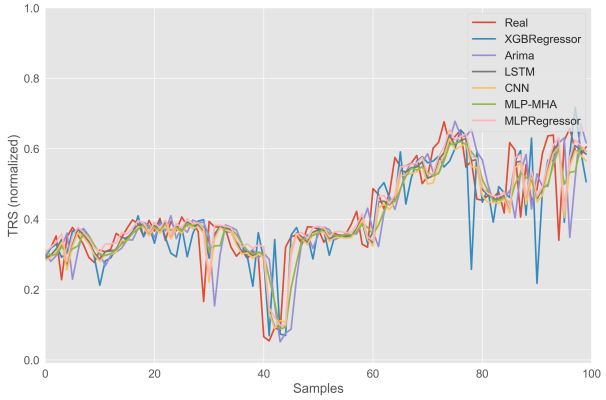


Figure 4. Performance for a lag of 3 with a horizon of 1.

Analyzing the results displayed in Table 2, it can be seen that CNN and LSTM showed the lowest value of MAE and MSE, with 0.046 and 0.006 for CNN and 0.047 and 0.006 for LSTM, respectively. These values indicate that these models were able to predict the time series with the least number of error, which is an important advantage in practical applications.

Table 2. Evaluating metrics for a lag of 3 with a horizon of 1.

	MAE	MSE	R^2-Score	Time (s)
XGBRegressor	0.0541	0.0091	0.7513	0.0687
ARIAMA	0.0565	0.0109	0.7046	0.0536
LSTM	0.0474	0.0067	0.8174	0.3766
CNN	0.0464	0.0066	0.8205	0.1639
MLP-MHA	0.0529	0.0075	0.7959	0.2078
MLPRegressor	0.0478	0.0069	0.8103	0.0059

In terms of R^2-Score, LSTM achieved the highest value (0.81), closely followed by CNN with a score of 0.82. This indicates that both models effectively capture a significant portion of the variability present in the time series data, presenting a substantial advantage.

Despite their competitive performance, CNN demonstrated a shorter execution time compared to LSTM. This aspect positions CNN as a preferable option when swift computation is imperative.

While XGBRegressor and MLPRegressor showcased marginally higher values for MAE, MSE, and R^2-Score when compared to CNN and LSTM, their noteworthy efficiency advantage over neural network models makes them notable choices. ARIAMA and MLP-MHA exhibited intermediate performance across the evaluation metrics.

To synthesize, Table 2 underscores CNN and LSTM as robust contenders for time series prediction, offering commendable accuracy and elevated R^2-Scores.

5.2 TRS forecasting for horizon equal 8

In order to evaluate the range of the prediction horizon for the lag value equal to 3, a horizon equal to 8 was adopted. The graph of this test is illustrated in Figure 5 and the evaluation metrics are shown in Table 3.

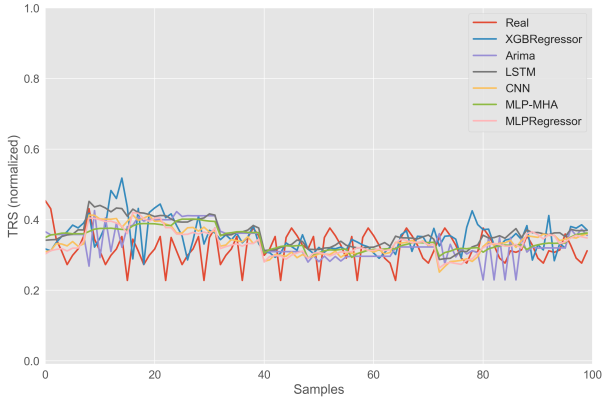


Figure 5. Performance for a lag of 3 with a horizon of 8.

Observing the results obtained, we can see that the CNN, MLP-MHA, LSTM and MLPRegressor algorithms had a similar performance regarding MAE, with values around 0.08. The XGBRegressor and ARIAMA had higher MAE values, with 0.09 and 0.07, respectively. However, it is important to note that the reference value for MAE can vary depending on the context of the problem, so it is important to evaluate other metrics as well.

In terms of MSE, the algorithms exhibited consistent results, all below 0.02. Among them, CNN achieved the

Table 3. Evaluating metrics for a lag of 3 with a horizon of 8.

	MAE	MSE	R^2-Score	Time (s)
XGBRegressor	0.0942	0.0203	0.4495	0.3887
ARIAMA	0.0785	0.0179	0.5145	0.0545
LSTM	0.0853	0.0175	0.5241	0.3835
CNN	0.0827	0.0159	0.5688	0.2198
MLP-MHA	0.0836	0.0158	0.5713	0.2526
MLPRegressor	0.0789	0.0153	0.5830	0.0020

lowest value, while ARIAMA and XGBRegressor had relatively higher MSE scores.

The R^2-Score quantifies the explained variance between dependent and independent variables. Here, CNN, MLP-MHA, and MLPRegressor performed comparably, each attaining a score of approximately 0.57, closely followed by MLPRegressor with 0.58. Conversely, ARIAMA and XGBRegressor demonstrated lower performance concerning this metric.

In execution time, MLPRegressor was swift at 0.002 seconds, whereas LSTM was the slowest at 0.38 seconds. The remaining algorithms exhibited similar execution times.

Regarding performance metrics (MAE, MSE, R^2-Score), CNN, MLP-MHA, LSTM, and MLPRegressor showed similar outcomes. However, ARIAMA and XGBRegressor fell short. Considering execution time is crucial; MLPRegressor's efficiency is an asset, but LSTM's prolonged duration is a drawback

6. CONCLUSION

Upon analyzing the outcomes, it can be observed that the LSTM and CNN models yielded the best results for the horizon of 1, with an MAE of 0.047 and 0.046, respectively. These models outperformed the XGBRegressor, which had the third-best performance, with an MAE of 0.054. On the other hand, for the horizon equal to 8, the MLPRegressor model had the best performance, with an MAE of 0.078, followed by ARIAMA, with an MAE of 0.078. The XGBRegressor had the worst performance, with an MAE of 0.094.

The use of these models for Total Reduced Sulfur prediction in the paper and pulp industry can bring significant benefits in terms of reducing environmental pollution and improving the health of the population. Total Reduced Sulfur is a highly toxic gas that can cause severe respiratory problems and has a pungent odor that can cause nuisance to the surrounding population. Therefore, predicting its concentration accurately can help industries take appropriate measures to control their emissions and reduce their impact on the environment and public health.

In addition to the environmental and health benefits, the use of these models can also bring significant economic advantages to the industry.

ACKNOWLEDGMENTS

We would like to express our gratitude to PROSYS for their valuable support and assistance during this research. Their expertise and resources were instrumental in helping us achieve our research goals. We also want to thank all the participants who generously gave their time and contributed to this study.

REFERENCES

Barino, F.O. and dos Santos, A.B. (2020). Rede neural convolucional 1d aplicada à previsão da vazão no rio madeira. *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*.

dos Santos, R.B., de Oliveira, C.V., de Oliveira, L.A., and Fagundes, L.A.A. (2021). Greenhouse gas emissions in the brazilian pulp and paper industry. *Journal of Cleaner Production*, 312, 127649.

Ferreira, L.C., Bizzo, W.A., Moraes, L.F.D., and Neto, J.D.M. (2019). Comparative analysis of the greenhouse gases emissions in the use of fossil fuels and biomass in a brazilian pulp and paper industry. *Journal of Cleaner Production*, 210, 1090–1097.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 29(5), 1189–1232.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Jafari, R., Shahsavari, S., and Taherzadeh, M. (2019). Prediction of sulfur dioxide emission from kraft pulp mills using machine learning techniques. *Journal of Cleaner Production*, 233, 983–992.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Mehmood, A., Lee, Y.H., and Kang, B.H. (2019). Short-term traffic flow prediction using multilayer perceptron regressor with principal component analysis. *Journal of Advanced Transportation*, 2019, 1–14. doi:10.1155/2019/2706826.

Silva, R.A., Ferreira, L.C., Fernandes, A.L., and Neto, J.D.M. (2020). Energy efficiency and ghg emissions mitigation opportunities in the brazilian pulp and paper industry. *Energy*, 191, 116580.

Valmet (2019). Vpsulf acid plant. `https://www.valmet.com/insights/articles/up-and-running/new-technology/VPSulfAcidPlant/`. Accessed on April 03, 2023.

Van Houdt, G., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53, 5929–5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. URL `https://arxiv.org/pdf/1706.03762.pdf`.

Wang, Y., Chen, J., Zhang, J., Lu, H., Zou, B., and Liu, Y. (2019). Air pollutant emissions from the combustion of pulping and papermaking wastes in china: Based on an industrial survey. *Journal of Cleaner Production*, 227, 864–874.

Wen, Z., Zhang, L., Li, H., Fang, G., and Li, W. (2021). Modeling the impact of operating parameters on greenhouse gas emissions from pulp and paper mills using machine learning. *Journal of Cleaner Production*, 279, 123383.

Yu, Y., Li, S., Luo, Z., Yang, J., Zhang, Y., and Wang, C. (2015). Study on the emission characteristics of trs from the incineration of cncg. *Journal of Hazardous Materials*, 297, 240–246. doi:10.1016/j.jhazmat.2015.04.076.