Big Data Analysis and Dimensionality Reduction to Predict Price Trends in the Brazilian Electricity Market Considering Interdisciplinary Phenomena

Everthon T. Sica, Larah F. R. Barboza, João V. R. Benedet, Karem V. P. de Lima, Vinicius V. L. Albani, Erinaldo Santos, Sérgio L. Avila, *Member, IEEE*

***Abstract*—This paper aims to reduce the dimensionality of features in the Brazilian electricity market when using the price trend forecasting model, thereby minimizing the model's complexity. This reduction must maintain the pertinence and relevance of the data and both quantitative and qualitative information. Our dimensionality reduction uses Spearman Correlation and Mutual Information for features coupled with time. Analysis guarantees the pertinence and relevance of the features, validated by evaluating the data selected in the study period through out-of-sample verification. Analysis of time-associated features with unique characteristics for application in the non-storable commodity market reduced the number of variables from 4,000 to 19, and the number of features from 13 to 6. Although dimensionality reduction of big data and the method used are familiar, we uniquely analyze time-coupled features with dynamic evaluation. Another contribution is the analysis and modeling of features with highly different characteristics and dimensions, including Hydrological, Energy, Climatic, Economic, and Geopolitical features. We obtained the features and database from 2015 to 2023, which are relevant to the Brazilian electricity market scenario. This method can be applied to other economic sectors and databases based on variation over time.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9771

***Index Terms*—Mutual Information, Big data, Database Coupled in Time, Price, Electricity Market Trend.**

I. Introduction

THE economic studies on energy planning show that microeconomic phenomena, such as supply and demand [1]-[4], and environmental factors [5]-[7], such as the utilization of natural capital, influence the behavior of the electricity market. Thus, the modeling and simulation strategy for long-term decision-making support in the future electricity market should be integrated and multivariate [4][5], correlating market trends and simulating multiple scenarios [8]-[10]. Decision-makers should have access to the largest possible number of variables, and the model must measure the qualitative differences in future scenarios [5][10]. Consequently, the predictions should be coherent with the organic behavior of natural and economic capital in the electric energy sector.

Electricity is a non-storable commodity, and its spot price in future markets exhibits pronounced seasonal patterns and is highly sensitive to real-time demand and supply in trading [11]. Therefore, it presents more volatile price developments compared to other energy sources, such as crude oil and natural gas. Multivariate long-term price forecasting models require organized data inputs, with well-defined categories and layers encompassing both qualitative and quantitative phenomena, to ensure precise forecasting. Thus, it is necessary to consider the characteristics of the power system, the electricity market, and the interdependence of decision variables to evaluate their mathematical and economic relevance for predicting price trends. Long-term multivariate forecasting is a complex big data problem due to the large number of variables and interdisciplinary factors, demanding high computational performance. Mathematical modeling evolves to reduce the cost of computational implementation, but the cost of information to make the model more accurate increases. Fig. 1 presents an outline of the information costs as the need for features and model inputs increases. This is relevant because integrated resource models represent a micro-reality that can be approached in two ways: top-down or bottom-up.

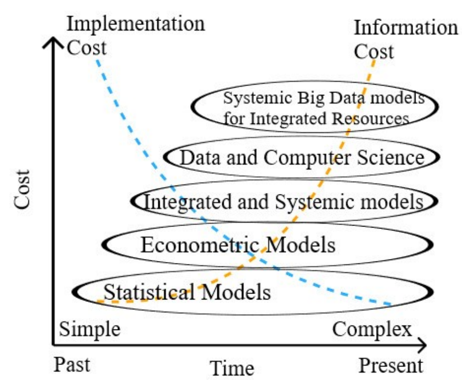

Fig. 1. Complexity and cost of information and implementation.

The approaches refer to strategies for how information is

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos Thomaz (C*orresponding author: Sérgio Luciano Avila*).

This research is supported in part by URCA, IFSC and EMBRAPII under grant PFLN-2211.0014.

E. T. Sica, L. F. R. Barboza, J. V. R. Benedet, K. V. P. de Lima, and Sérgio Luciano Avila are with the Federal Institute of Santa Catarina, Florianópolis, Brazil (e-mails: sica@ifsc.edu.br, larah.rb@aluno.ifsc.edu.br, joao.vrb@aluno.ifsc.edu.br, karem.v@aluno.ifsc.edu.br, and sergio.avila@ifsc.edu.br).

V. V. L. Albani is with the Federal University of Rio de Janeiro, Rio de Janeiro, Brazil (e-mail: vinicius.albani@ufrj.br).

E. Santos is with the URCA Trading Inc., São Paulo, Brazil (e-mail: erinaldo@urcaenergia.com).

structured and processed, whether in an aggregate or disaggregate manner. In the top-down approach, the modeled system is defined as a whole, with a simplified specification of the aspects and phenomena that comprise it. In the bottom-up approach, the system is modeled by providing a detailed specification of the elements that constitute it, which are progressively organized across hierarchical levels until a comprehensive system definition is achieved [12]-[15]. Regardless of the approach, challenges related to variable stratification, multidisciplinary representation, and the number of phenomena represented must be addressed to manage dimensionality.

This paper addresses the challenge of building a reduced and coherent database integrated with a bottom-up energy resource model to predict price trends in the Brazilian electricity market, whose energy system is large, interconnected, and predominantly hydro-thermal-wind. The key contribution lies in producing concise, relevant, real-time information for traders, promoting greater efficiency in decision-making. Fig. 2 describes the proposed approach, highlighting the multidisciplinary nature and complexity of the problem, and relating hydrological, energy, climatic, economic, and geopolitical phenomena. Therefore, several variables (time series) associated with phenomena relevant to price trend forecasting are modeled, such as Affluent Natural Energy, Stored Hydro Energy, Turbined Spillable Hydro Energy, Spot Price, Gross Domestic Product, Marginal Cost of Operation, Economic Activity Index, Load Projections, Available and Projected Power, Transmission Constraints, Oceanic Niño Index, Geopolitical Risk Index, Tropical Northern Atlantic Index, Hydraulic Constraints, and other economic and regulatory factors.

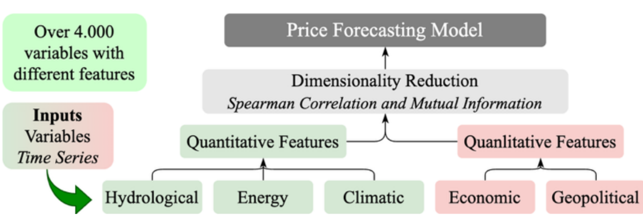

Fig. 2. Organized integrated resource model.

As shown in Fig. 2, the analysis aims to process the number of phenomena and their variables (time series) so that the integrated pricing trend model is both parsimonious and reliable. To determine which variables will be excluded from the model, a feature engineering process will be used to select the relevant phenomena and variables. Therefore, we suggest using feature engineering through Spearman Correlation and Mutual Information to reduce dimensionality, due to the peculiarities of Brazil's power system and electricity market.

The Spearman Correlation (SC) technique is applied to measure the statistical relationship between variables, while Mutual Information (MI) is used to verify the existence of mutual dependence between variables. The systemic model database comprises 13 phenomena, which resulted in over 4,000 different variables (time series), organized according to the principles of integrated resource planning. Due to the size of the database, analysis using SC and MI alone is not effective. Consequently, a linear technique for dimensionality reduction was implemented, which uses vector orthogonalization to transform correlated variables into a set of linearly uncorrelated values, thereby improving computational performance without compromising the accuracy and reliability of the results. Therefore, exploring the phenomena related to future electricity pricing trends can yield significant results in reducing the complexity of the problem and making long-term forecasting more accurate and agile.

II. Data Collection

The analysis of the aspects' interrelationships (see Fig. 3), along with the idiosyncrasies of the Brazilian electricity market, involves phenomena with dimensions that differ from, yet have causal effects on, the electricity market, and are therefore interdependent. It is advisable to consider variability and anthropogenic interference, as it is essentially both a multidisciplinary and an interdisciplinary problem [15]. The aspects form a set of data and information as an open system inserted into a systemic model with interdependent phenomena.

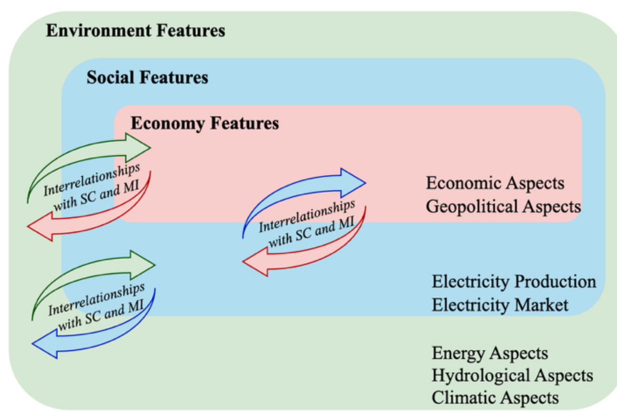

Fig. 3. Systemic model with interdependent phenomena.

Fig. 3 shows how the relationships among the energy, hydrological, climatic, economic, and geopolitical aspects of the power system and the electricity market can be understood through their dynamic interactions. This reinforces the complexity of the modeling problem and emphasizes the need for dimensionality reduction and big data analysis.

A. Essential Characteristics of the Brazilian Electricity Market

Predicting price trends involves forecasting the future behavior of a market product. This forecast is constructed with a certain granularity (frequency) according to a maturity period (discretization for delivery) and a historical series (observable variable). The product for the Brazilian electricity market is determined according to the operation of the National Interconnected System (SIN) by the National System Operator (ONS) and settled monthly by the Electricity Trading

Chamber (CCEE).

The National Interconnected System is a large hydro-thermal-wind system, with a predominance of hydroelectric plants, and has four subsystems: South, Southeast/Central-West, Northeast, and a large part of the North region [16]. The interconnection via transmission grid enables energy transfer between subsystems, allows synergistic by leveraging the diversity of hydrological regimes across basins, and ensures that the Brazilian market is supplied with energy security and economically.

In the Brazilian market, the product is defined according to the operation of the electricity system in a month, for example, product APR2024. The product results from the Monthly Operation Schedule (PMO), the Marginal Operation Cost (CMO), and the Difference Settlement Price (PLD). Products can be traded in two environments: the regulated market and the free market. In the regulated market, the product is commercialized according to the rules established by the National Electric Energy Agency (ANEEL), usually through auctions. In the free market, the product can be traded in two ways: the futures market or the spot market.

In the spot market, the settlement of traded products occurs with a very short delay. In the futures market, the settlement of traded products is only made in the future or even when the product is delivered (at the end of the month, for example). When trading in the on-the-spot market, the trader must pay a certain amount of money to acquire the product or receive it immediately when selling it. When trading in the futures market, the trader agrees to trade forward, i.e., the trader will buy or sell the product according to pre-agreed values for the day stipulated at the end of the contract - the delivery time and settlement of the product.

In the free market, the product can be traded in advance by accredited economic agents in an open trading environment. In Brazil, this free trading environment is called the Brazilian Electricity Trading Desk (BBCE). At BBCE, each product is traded in real-time, and it is possible to observe the historical series of prices, s, quantities traded, q, or even the quantities bought (demand), $q^{(d)}$, and sold (supply), $q^{(s)}$. However, a specific product, t, with maturity, h, can be traded several times, n, on the same market day, d, with different prices, $s_{h,t,n,d}$, and quantities, $q_{h,t,n,d} = q^{(s)}_{h,t,n,d} + q^{(d)}_{h,t,n,d}$. Thus, to reduce dimensionality and maintain the nature of the economic phenomenon by using a free interest rate, the market price is modeled using the Volume Weighted Average Price (VWAP) [17][18]. After manipulating the raw price data, a historical series is created in a single vector of information, according to:

$$s^{VWAP}_{h,t,d} = \frac{\sum_n^N \left(s_{h,t,n,d} \times q_{h,t,n,d}\right)}{\sum_n^N q_{h,t,n,d}} \tag{1}$$

where $s_{h,t,n,d}$ and $q_{h,t,n,d}$ are the price and the volume of each trade, n, of the same product, t, at the same maturity, h, and on same day, d, will be the VWAP price, $s_{h,t,d}$, with $n = 1, ..., N$, where $n = N$ refers to the last trade. In turn, the historical series of trading volume values is obtained according to:

$$q^{VWAP}_{h,t,d} = \sum_n^N q_{h,t,n,d} \tag{2}$$

where $q_{h,t,n,d}$ is the volume of each trade, n, of the same product, t, at the same maturity, h, will be the VWAP volume, $q_{h,t,d}$. Henceforth, the VWAP price in (1) and VWAP volume in (2) are denoted simply as $s_{h,t,d}$ and $q_{h,t,d}$, respectively, in the product information matrix:

$$t = \begin{bmatrix} \underset{\{d|d\in d(yy,mm,dd)+j\}\forall j=0,\dots,J}{d = d(yy,mm,dd)+0} & \underset{\{h|h\in H-j\}\forall j=0,\dots,J}{s_{h=H-j,t,d}} & q_{h=H-j,t,d} \\ d = d(yy,mm,dd)+1 & s_{h=H-1,t,d} & q_{h=H-1,t,d} \\ d = d(yy,mm,dd)+2 & s_{h=H-2,t,d} & q_{h=H-2,t,d} \\ \vdots & \vdots & \vdots \\ \underset{\{d:d=d(yy,mm,dd)+j\to d(YY,MM,DD)\}}{d = d(yy,mm,dd)+j} & \underset{\{h:h=H-j\to h=0\}}{s_{h=0,t,d}} & q_{h=0,t,d} \end{bmatrix} \tag{3}$$

where t is the product matrix obtained by month *[MM]* and year *[YY]*; d is trade date in the format *[yy, mm, dd]*; h is product maturity; $s_{h,t,d}$ is the price; and $q_{h,t,d}$ is the volume. Thus, the product historical series, t *(MM, YY)*, begins on the first trading day ($h = H - j \rightarrow j = 0$) at the highest maturity, H, until the last trade or delivery, d *(yy, mm, dd)* $+ j \rightarrow d$ *(YY, MM, DD)*, of the product. Thus, the price, $s_{h,t,d}$, and volume, $q_{h,t,d}$, are progressively associated ($j = 0, ..., J$) by the date of the first trade, $j = 0$, up to the last trading day, $j = J$, at the highest maturity, $h = H - j \rightarrow h = 0$. Likewise, each feature will have the same granularity and raw data information for the same day, d.

B. Features Organization

According to (3), the features are organized in the matrix, based on the performance of each feature, $f = 1, ..., F,$ for each day, d, forming a complex and large matrix. Additionally, we implemented Python scripts for real-time data acquisition to database construction. Thus, the 5 groups in Fig. 2 is organize into 13 features: (i) Hydrological: Affluent Natural Energy (ENA); (ii) Energy: subsystems' Marginal Cost of Operation (CMO), Total Capacity of Transmission Lines [MVA], Total Size of Transmission Lines [Km], Electricity Load Projection, and subsystems' Average Stored Energy (EARm); (iii) Climatic: Climatology (Clim), Monthly Niño Ocean Index (ONIm), and Tropical North Atlantic Ocean Temperature indicator (TNA); (iv) Economic: Central Bank Economic Activity index (IBC), Gross Domestic Product (GDP), and Volume Weighted Average Price (VWAP); and (v) Geopolitical: Geopolitical Risk Index (GPRD).

The Hydrological group is composed of time series variations of historical Affluent Natural Energy (ENA) [19][20]. Each variation of the time series is formed with shifts, *sh*, of 0, 30, 60, 90, and 120 days for each Simple Moving Average (SMA) of 5, 7, 15, 30, 60, and 90 days. Shift (delay) is a way of identifying the temporal interrelationship between the variable (time series) of each feature, $f_{F,d}$, and the behavior of the VWAP, $s_{h,t,d}$. In other words, a shift helps determine whether a change in the behavior of a feature affects the price trend at the same time it occurs, or 30 days in advance, or 60 days in advance, and so on.

From the trader's perspective, changes in a feature's performance, $f_{F,d}$, can influence buying and selling decisions in different ways. For example, consider the trading of the APR2024 product in February 2024 (*M+2*): which ENA forecast time series has the greatest impact on the price of this product? The ENA forecast for April 2024 made in February (*sh = 0*), or the one made in January (*sh = 30)?* To increase the profit potential, the trader takes uncertainties into account, which, naturally, increases the risk in the futures market.

The Energy group comprises the Marginal Operating Cost (CMO) of the four subsystems, *sb*, (South, Southeast/Central-West, Northeast and North) of the SIN, the capacity and size of the SIN transmission network, the Electricity Load Projection (Proj_C), and the Average Stored Energy (EARm) of the four subsystems, *sb*, and the SIN [19]. The Average Stored Energy (EARm) refers to the amount of energy corresponding to the average volume of water stored in the reservoirs of hydroelectric plants, which can be used for power generation if the water is released through the turbines.

The Load Projection (Proj_C) considers the month under study and the four subsequent months, $M + i, i = 0, \dots, 4$. For example, regarding April 2024 (APR2024), *M+0* is the Electricity Load Projection for April 2024 based on data from April 2024; *M+1* is the projection for May 2024, based on the initial conditions of April 2024; *M+2* is the projection for June 2024, based on the initial conditions of April 2024, and so on.

The Climatic group comprises climatology (Clim), the El Niño Mean Ocean Index (ONIm), and the Tropical North Atlantic Ocean Water Temperature Index (TNA). Climatology refers to the country's climatological normal (the expected behavior for each year) or the long-term historical average of the ENA for the period analyzed [19][20].

The Economic group comprises the Gross Domestic Product (GDP) index, the Economic Activity Index (IBC) [21], and the Volume Weighted Average Price. Volume Weighted Average Price (VWAP) is defined by (1). GDP and IBC are considered qualitative due to the way information is collected and the political-regulatory influence, as evidenced in studies [22]-[28]. Therefore, it is imperative to incorporate qualitative variables to achieve long-term forecasts, whether related to prices or demand.

The Geopolitical group comprises the Geopolitical Risk Index (GPRD) [29]. This feature was chosen as a qualitative parameter because of the political-regulatory influence on commodity prices. For example, an analysis of trends, falls, and correlations in long-term commodity prices was performed by [30]. The adopted approach encompassed a dataset spanning 700 years (from 1300 to 2019), considering seven different commodity curves. The results showed that substantial price changes are often linked to significant historical events, often of a purely qualitative nature, such as sudden regulatory changes [21][30]. Political instability and conflicts in important international regions can disrupt the power supply from energy resources, such as renewables, and increase commodity prices, resulting in reduced purchasing power for energy consumers [31]. In [32], the GPRD was included in the model studied.

III. Feature Engineering Review and Method

Companies across various sectors encounter the challenge of managing large volumes of diverse data daily. When processed in real-time, this data has the potential to enhance decision-making. In big data, streaming feature selection has emerged as an effective technique for identifying the most relevant features in high-dimensional datasets, reducing learning complexity. This selection pertains to features that arrive sequentially over time, with a well-established number of instances but a variable number of features. Many researchers have proposed streaming resource selection algorithms to address this issue, as discussed in [33], which also highlights current challenges in big data research.

In data mining, the emphasis on feature selection has been growing, as a subset of quality data is crucial for building effective models. Ref. [34] summarizes data feature selection and underscores the need for careful consideration when choosing features for big data, where issues such as noise and data types are amplified. Challenges are classified according to scale, dimensionality, dynamicity, missing data, heterogeneity, and reliability.

Feature engineering is defined in two ways [35]: preparing input data into features understandable by machine learning algorithms and transforming variables into features to enhance performance and interpretability. The pair-wise verification process between features for exploration data analysis, as discussed in [36], is essential for prediction, allowing for the elimination of unnecessary variables. However, the high dimensionality of the feature space complicates this task.

The complexity and quantity of data hinder predictive outcomes, regardless of the machine learning technique employed [37]. However, machine learning models exhibit limited capacity to extract useful information as the number of features increases, which negatively affects prediction performance [38]. Simple models struggle to grasp deep information embedded within the data, leading to recommendations for combining deep learning models with feature engineering.

A comparison of machine learning models for prediction is conducted by [39], analyzing different features engineering approaches across 24 schemes. In our research, we employ weighted averages to address data gaps, like the methods described in [40], which implemented 12 machine learning models and created features based on prior domain knowledge. In the Brazilian power system, the increasing volume of stored data makes it necessary to filter and select the most relevant information for the problem at hand.

Computerized methods encountered significant resource and time consumption because of the number of variables associated with high-entropy alloy characterization [41]. A day-ahead forecast of electricity market prices using long-short term memory recurrent neural networks is done by [42], creating 44 features to enable the model to learn trends and volatilities.

A survey's data-driven approaches identifying various feature types that contribute to energy consumption prediction, such as meteorological data, building characteristics, and historical data, are presented by [43]. Our paper presents features that can yield favorable results for price trend prediction in the Brazilian electricity market.

Ref. [44] discusses a feature engineering process to extract and select the most important features, addressing overfitting issues. The authors noted that extraction aims to reduce the original feature space while selection focuses on choosing relevant information for efficient machine learning model training. Our study applies this principle by employing feature engineering for the same purpose. Similar to [45][46], we use knowledge-driven feature engineering, which combines domain knowledge with feature engineering to refine the variables related to the problem.

The feature engineering processing method can employ an attention mechanism because of the highly fluctuating nature of the data, allowing the model to learn more effectively [47]. Ref. [48] categorizes feature engineering techniques into three groups: filter, wrapper, and hybrid. Filter techniques eliminate insignificant features but may not address redundancy, while wrapper techniques analyze features and incorporate a selection algorithm within an induction algorithm, albeit with high computational costs. Hybrid techniques combine these approaches in a single model, balancing computational power and efficiency.

Ref. [49] uses hybrid feature engineering techniques, such as instance-based Relief-F (wrapper) and MI, for variable selection and dimensionality reduction. Ref. [50] states that dimensionality reduction methods are commonly used in supervised machine learning, highlighting the effectiveness of feature selection and extraction techniques. Feature selection reduces dimensionality and identifies relevant variables, with MI being a key tool. Ref. [51] discusses electricity market price forecasting that uses MI for feature selection, redundancy identification, and assessment of individual input relevance.

Ref. [52] employs a multi-label concept for feature selection, categorizing features into distinct labels through mutual information analysis. The input variables are typically selected based on user experience and trial-and-error methods [53]. Given many non-linear variables in electricity price forecasting, methods such as MI are valuable. Ref. [54] applies MI for feature selection but notes that the algorithm suffers from the curse of dimensionality, incurring high computational burdens. To improve this, it suggests normalizing inputs and using joint probability for MI computation. We implement MI to analyze nonlinear variables, finding it valuable for investigating the interrelationships between features and price. The research conducted by [55] investigates various feature engineering approaches, emphasizing that optimal techniques vary across machine learning models.

For feature visualization, analyzing correlations between feature pairs to create a correlation matrix is recommended. For feature selection, the Random Forest technique is suggested to be applied to the decision tree model in Fig. 2, while Principal Component Analysis is regarded as a leading method for feature exclusion.

Our research paper aligns with these approaches, employing SC and MI matrices for visualization and implementing a decision model for selection, focusing on the utility among the variables. The methodological steps and procedures applied to reduce dimensionality are summarized in Fig. 4. We implement SC to eliminate redundant features and MI to filter out insignificant ones, carefully evaluating variable importance.

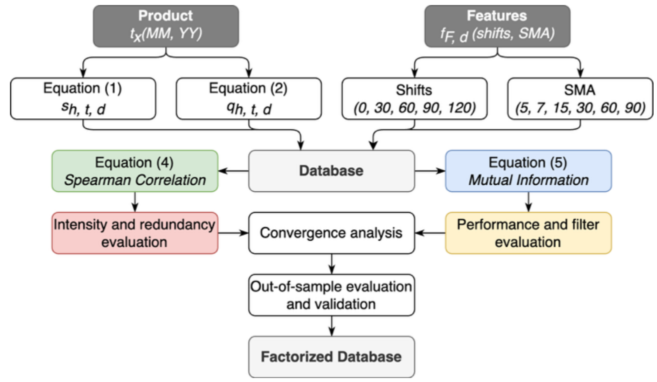

Fig. 4. Methodological procedure.

A. Spearman Correlation

Spearman Correlation (SC) is a non-parametric measure of statistical dependence between two variables, where the relationship is described by a monotonic function. SC works best for non-linear information, which is characteristic of the features and variables that comprise the integrated resource model (Fig. 2). This method normalizes its result, returning values between -1 and 1 [56], as shown next:

$$\rho_d = 1 - \frac{6}{F_d(F_d^2 - 1)} \sum_{f_d=1}^{F_d} \kappa_{f_d}^2 \tag{4}$$

where ρ_d is the Spearman Correlation coefficient on day, d; κ_{f_d} is the difference between two feature performances on day, d; and F_d is the number of feature observations on day, d. The SC result is used to evaluate the correlation between random features, $f_d = 1, \ldots, F_d$, on the same day, d, to determine whether the feature can be excluded from price trend forecasting. We adapted the interpretation of the SC result using a qualitative intensity scale applying it according to the specificities of the problem, as follows: $\rho = 0$ means no correlation; $0 < |\rho| \leqslant 0.20$ is indifferent; $0.20 < |\rho| \leqslant 0.40$ is weak; $0.40 < |\rho| \leqslant 0.60$ is moderate; $0.60 < |\rho| \leqslant 0.80$ is strong; $0.80 < |\rho| < 1.00$ is absolute; and 1 is monotonic correlation [57].

In our study, for example, the Total Line Capacity [MVA] feature has seasonal behavior over time, with some periods showing a positive correlation between features and others showing a negative correlation. However, this feature was excluded due to low relevance and weak intensity in both MI and SC. Furthermore, the capacity of transmission lines involves discrete events over time, as their capacity increases occur at specific points.

We implemented a performance matrix to visualize the time-coupled behavior of each variable in each feature. This matrix supports validation and specific studies. Nevertheless, including them here is impractical, so they are available in the supplementary material on the GitHub link. However, in this paper, we present the main analyses in the following.

The Climatology (Clim) feature (see Fig. 5) shows a strong positive correlation between the variables with a 30-day shift and the ENA_hist feature. However, the VWAP feature shows weak correlation. In this sense, it can be concluded that SC between Clim and ENA presents a redundancy, given that climatology is the country's climatological normal or even the historical average of ENA over a long period (1991-2020).

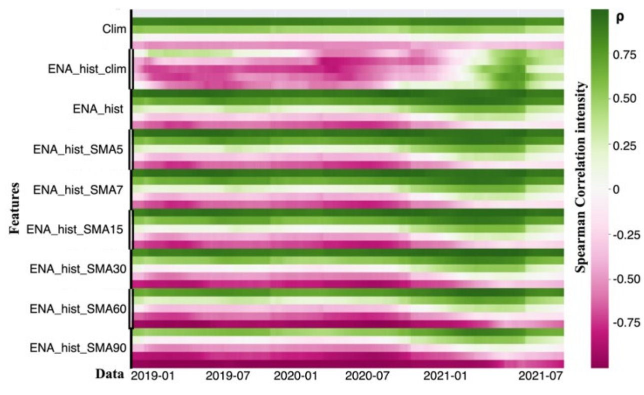

Fig. 5. Spearman correlation for Climatology feature.

The Clim features (with 30-, 60-, 90-, and 120-day shifts) show strong positive correlations with each other, with the ENA_hist feature, and with the Electricity Load Projection (Proj_C) feature. However, the ENA_hist feature has a weak correlation with the VWAP feature; therefore, it is excluded. The results show that the greater the shift for the Clim feature, the stronger the positive correlation with the subsystems' Average Stored Energy (EARm), and the stronger the negative correlation with Electricity Load Projections. The subsystems' EARm feature shows a strong positive correlation, except in the South subsystem. There is a strong negative correlation between the subsystems' CMO feature with most of the Clim and EARm features, which weakens with increased shifts. However, a slightly negative correlation is observed under certain 2022 VWAP, while a positive correlation remains with Electricity Load Projections. The features not mentioned have a weak correlation with the VWAP feature and are likely to be excluded. However, whether they can be refuted requires further analysis using the Mutual Information technique.

B. Mutual Information

Mutual Information (MI) is a relevant metric for filtering and ranking features. With the obtained MI indicator, it is possible to evaluate the associations between VWAP feature and various Hydrological, Energy, Climatic, Economic, and Geopolitical features implemented in the database, thereby reducing its size. Unlike the Spearman Correlation, MI can detect any type of dependency between variables:

$$\mathrm{MI}(f_{X,d}, f_{Y,d}) = \sum_{\substack{x=1 \\ x\epsilon X}}^{X} \sum_{\substack{y=1 \\ y\epsilon Y}}^{Y} p(f_{x,d}, f_{y,d}) log_2\left(\frac{p(f_{x,d}, f_{y,d})}{p(f_{x,d}) \cdot p(f_{y,d})}\right) \quad (5)$$

where $f_{X,d}$ and $f_{Y,d}$ are the pair of feature observations analyzed on day, d; p($f_{x,d}$, $f_{y,d}$) is the joint probability distribution; and $p(f_{x,d})$ and $p(f_{y,d})$ are the marginal probability distribution functions. The minimum value of MI is 0, due to its logarithmic. When MI equals 0, the variables are independent – that is, one does not influence the other [54][58]. MI provides a more comprehensive analysis of features and their dependencies, assessing the ambiguity of each feature and its relevance to the price forecasting.

As in the previous section, figures were generated for all features and are available in the supplementary material on the GitHub link. In this sense, some conjectures are presented. Fig. 6 shows that the Clim feature has high MI with several ENA variations, except for ENA_hist_clim, whose MI decreases as the shift increases.

The subsystems' CMO (except for the South subsystem) and Clim features show high MI with each other, as well as with the Electricity Load Projection, EARm, TNA, and VWAP features. However, both have low MI with GPRD, GDP, and IBC features.

The subsystems' EARm features show high MI both with each other and with the Electricity Load Projection and ENA_hist_SMA features. The GPRD, GDP, and IBC features are excluded, as they show low to moderate MI in some parts of the year. The ONIm and TNA features show high MI with all the EARm, Electricity Load Projection, VWAP, and Clim features.

The VWAP feature has a high MI with several features from the Hydrological and Climatic groups, which enables an accurate analysis by highlighting features that can be refuted.

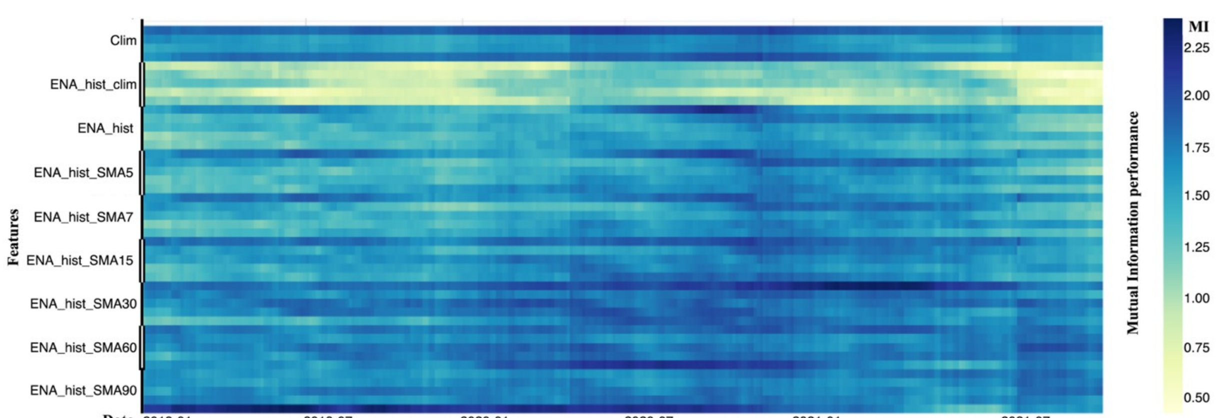

Fig. 6. Mutual Information for Climatology feature.

IV. Analysis of results

The analysis investigates the influence and relevance of Hydrological and Climatic group features on market price trends and their impact on other group features. For the features in the Economics group, Spearman Correlation (SC) shows that the GDP feature is weakly influenced by the behavior of the Hydrological and Climatic group features with a 30-day forward forecast. The IBC feature is moderately influenced by the behavior of the Hydrological and Climatic group features, with a 0-day-ahead forecast. The GPRD feature is not influenced by the behavior of the Hydrological and Climatic group features, so it should be excluded. According to Mutual Information (MI), the GDP and IBC features show low dependence on the behavior of the Climatology features, especially with the 30- and 60-day forward forecasts. The GPRD feature shows moderate dependence on the behavior of Hydrological and Climatic group features but has the lowest conceptual relevance and should therefore be excluded.

For the Energy features, Spearman Correlation (SC) shows that the Total Line Capacity and the subsystems' CMO features are moderately influenced by the behavior of the Hydrological and Climatic group features in the 30-day-ahead forecast. The Total Line Size feature is not influenced by the behavior of the Hydrological and Climatic group features and should therefore be excluded. According to Mutual Information (MI), features such as Total Line Size and Total Line Capacity show no MI. The result was inconclusive and should therefore be excluded. The subsystems' CMO feature shows regional variation in dependence on Climatology features - low in the South and Southeast/Central-West, but high in the North and Northeast for 30- and 60-day-ahead forecasts.

The Electricity Load Projection feature, according to the SC, is strongly influenced by the behavior of the Clim feature, with a 0-day-ahead forecast and moderately with a 30-day-ahead forecast. The Electricity Load Projections (*M+0 to M+4*) are influenced by the ENA_hist feature, with moderate dependence at shorter horizons (*M+0 and M+1)* and stronger dependence at longer horizons (*M+2 to M+4*) for a 0-day-ahead forecast, decreasing to moderate at 30 days. According to MI, the Electricity Load Projection features (*M+1 to M+4)* show moderate dependence on Clim and ENA_hist (*SMA=15*) with 60- and 90-day-ahead forecasts, while *M+0* shows high dependence on both. All projections (*M+0 to M+4*) show high dependence on the TNA and ONIm features.

According to SC, the EARm features of the SIN, North, and Northeast subsystems are weakly influenced by the Clim and several ENA_hist features - particularly with a 90-day-ahead forecast for shorter SMs (5 and 7) and a 60-day-ahead forecast for longer SMA (15, 30, and 90). The EARm features of the South and Southeast/Central-West subsystems are weakly influenced by the Clim and ENA_hist features with 60-, 90-, and 120-day-ahead forecasts. According to MI, the subsystems' EARm features show moderate dependence on the behavior of the Clim features with 0-, 30-, 60-, and 90-day-ahead forecasts, as well as on the behavior of the TNA and ONIm features.

According to the SC, the VWAP feature for products *M+0*, *M+1*, *M+2*, and *M+3* is weakly influenced by the behavior of the ONIm feature, as well as by the Clim and ENA_hist features, with 90- and 120-day-ahead forecasts. The same interpretation applies, but with less adherence to the variations of ENA_hist (*SMA=5, ..., 90*), always with 90- and 120-day-ahead forecasts. According to MI, the VWAP feature for products *M+0*, *M+1*, *M+2*, and *M+3* shows moderate dependence on the behavior of Climatology features with 90- and 120-day-ahead forecasts. A similar interpretation applies to the variations ENA_hist, but with low dependence at a 90-day-ahead forecast. The VWAP feature for all products shows moderate dependence on the TNA and ONIm features, as well as on the behavior of the subsystems' CMO feature - with high dependence in the North and Northeast and low in the South and Southeast/Central-West subsystems. The VWAP feature for all products shows high dependence on the behavior of the Electricity Load Projection feature, but with lower dependence for longer-term products (*M+2 and M+3*). The VWAP feature for all products shows moderate dependence on the behavior of the subsystems' EARm features.

V. Conclusion

We aim to reduce the dimensionality of the features in the Brazilian electricity market when using the price trend forecasting model, thereby minimizing both the model's complexity and the cost of computational implementation. It can be concluded that the most relevant and useful features for the price trend model in the real-time electricity market are:

- For Gross Domestic Product (GDP) and Economic Activity Index (IBC) - the Climatology features (with 30- and 60-day-ahead forecasts) are the most relevant.
- For Marginal Cost of Operation (CMO) - the ONIm, Climatology (with 30- and 60-day-ahead forecasts), and TNA features are the most relevant.
- For Electricity Load Projection - the TNA and Climatology features (with 30- and 60-day-ahead forecasts) are the most relevant.
- For Average Stored Energy (EARm) - the Climatology (with a 90-day-ahead forecast), TNA and ONIm features are the most relevant. However, the TNA has greater influence on the SIN, South, and Southeast/Central-West subsystems and less influence on the North and Northeast subsystems.
- For VWAP - the ONIm, TNA, and Climatology (with 30-, 0-, and 60-day-ahead forecasts) features are the most relevant. However, the ONIm feature has greater influence, and the TNA feature has moderate influence, as do the Electricity Load Projection (M +i) and Marginal Operating Cost (CMO) features on the North and Northeast subsystems. In turn, Average Stored Energy (EARm) has a moderate influence in the SIN, Southeast/Central-West and Northeast subsystems, a greater influence in the SIN, Northeast, and Southeast/Central-West subsystems, and a lower intensity and relevance in the South subsystem.

Therefore, the five groups and 13 features, totaling 4,000 variables (time series), can be reduced to 6 features and 19 variables without loss of information and relevance: (i)

Climatology (with 0-, 30-, and 60-day-ahead forecasts); (ii) ONIm; (iii) TNA; (iv) Electricity Load Projection; (v) Marginal Cost of Operation (CMO) for the Southeast/Central-West, Northeast and North subsystems; and (vi) Average Stored Energy (EARm) for the Southeast/Central-West, Northeast, North, and SIN subsystems.

Our study has an original aspect that lies in the analysis of time-coupled features through a dynamic evaluation. Another contribution is the analysis and modeling of features with very different characteristics and dimensions, as this paper considers Hydrological, Energy, Climatic, Economic, and Geopolitical features. The features and database cover the period from 2015 to 2023 and are specifically relevant to the Brazilian electricity market and power system. However, the method can be applied to other economic sectors and databases based on variation over time.

References

[1] M. Ben Tahar, S. Ben Slimane, and M. Ali Houfi, “Commodity prices and economic growth in commodity-dependent countries: New evidence from nonlinear and asymmetric analysis”, *Resources Policy*, vol. 72, p. 102043, Aug. 2021, doi: 10.1016/J.RESOURPOL.2021.102043.

[2] F. Zhou, L. Page, R. K. Perrons, Z. Zheng, and S. Washington, “Long-term forecasts for energy commodities price: What the experts think”, *Energy Econ*, vol. 84, p. 104484, Oct. 2019, doi: 10.1016/J.ENECO.2019.104484.

[3] L. G. T. Carpio, “Cointegration to estimate long-term electricity prices in periods of rationing: The case of the brazilian hydrothermal system”, *Electric Power Systems Research*, vol. 121, p. 351–356, Apr. 2015, doi: 10.1016/J.EPSR.2014.11.022.

[4] F. Ziel and R. Weron, “Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks”, *Energy Econ*, vol. 70, p. 396–420, Feb. 2018, doi: 10.1016/J.ENECO.2017.12.016.

[5] F. Ziel and R. Steinert, “Probabilistic mid- and long-term electricity price forecasting”, *Ren and Sust Energy Reviews*, vol. 94, p. 251–266, Oct. 2018, doi: 10.1016/J.RSER.2018.05.038.

[6] E. O. Jåstad, I. M. Trotter, and T. F. Bolkesjø, “Long term power prices and renewable energy market values in Norway – A probabilistic approach”, *Energy Econ,* vol. 112, p. 106182, Aug. 2022, doi: 10.1016/J.ENECO.2022.106182.

[7] R. Huisman, “The influence of temperature on spike probability in day-ahead power prices”, *Energy Econ*, vol. 30, p. 2697–2704, Sept. 2008, doi: 10.1016/J.ENECO.2008.05.007.

[8] A. Livas-García, O. May Tzuc, E. Cruz May, R. Tariq, M. Jimenez Torres, and A. Bassam, “Forecasting of locational marginal price components with artificial intelligence and sensitivity analysis: A study under tropical weather and renewable power for the Mexican Southeast”, *Electric Power Systems Research*, vol. 206, p. 107793, May 2022, doi: 10.1016/J.EPSR.2022.107793.

[9] H. Lu, X. Ma, M. Ma, and S. Zhu, “Energy price prediction using data-driven models: A decade review”, *Comput Sci Rev*, vol. 39, p. 100356, Feb. 2021, doi: 10.1016/J.COSREV.2020.100356.

[10] R. Weron, “Electricity price forecasting: A review of the state-of-the-art with a look into the future”, *Int J Forecast,* vol. 30, p. 1030–1081, Oct. 2014, doi: 10.1016/J.IJFORECAST.2014.08.008.

[11] J. Nowotarski, J. Tomczyk, and R. Weron, “Robust estimation and forecasting of the long-term seasonal component of electricity spot prices”, *Energy Econ,* vol. 39, p. 13–27, Sept. 2013, doi: 10.1016/J.ENECO.2013.04.004.

[12] J. N. P. Swisher, G. M. Jannuzzi, and R. Y. Redlinger, *Tools and Methods for Integrated Resource Planning: Improving Energy Efficiency and Protecting Environment*. Riso National Lab, Denmark: UNEP Collaborating Centre on Energy and Environment, 1997.

[13] J. C. Piai, R. D. M. Gomes, and G. D. M. Jannuzzi, “Integrated resources planning as a tool to address energy poverty in Brazil”, *Energy Build*, vol. 214, p. 109817, May 2020, doi: 10.1016/J.ENBUILD.2020.109817.

[14] G. M. Jannuzzi, J. N. P. Swisher, and R. Y. Redlinger, *Planejamento Integrado de Recursos Energéticos: Oferta, Demanda e suas Interfaces*, 2th ed. Campinas: IEI, 2018. Accessed: Jun. 16, 2024. [Online]. Available: https://iei-brasil.org/livro-pir/

[15] E. T. Sica, “Integrated planning of water resources for electricity production: a decision support system by multicriteria and system dynamics”, Ph.D. dissertation, Department of Electrical Engineering, Federal University of Santa Catarina, Florianópolis, 2009.

[16] ONS, “Map of the Interconnected Power Transmission and Production System - Horizon 2027”, National Electrical System Operator. Accessed: Jun. 17, 2024. [Online]. Available: https://www.ons.org.br/paginas/sobre-o-sin/mapas

[17] C. Kath and F. Ziel, “The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts”, *Energy Econ*, vol. 76, p. 411–423, Oct. 2018, doi: 10.1016/J.ENECO.2018.10.005.

[18] T. Serafin, G. Marcjasz, and R. Weron, “Trading on short-term path forecasts of intraday electricity prices”, *Energy Econ,* vol. 112, p. 106125, Ago. 2022, doi: 10.1016/J.ENECO.2022.106125.

[19] ONS, “General Data: operation results”, National Electrical System Operator. Accessed: Jun. 17, 2024. [Online]. Available: https://www.ons.org.br/paginas/resultados-da-operacao/historico-da-operacao/dados-gerais

[20] AMPERE, “Ampere Consultoria”, Energy planning, development and meteorology. Accessed: Jun. 17, 2024. [Online]. Available: https://ampereconsultoria.com.br/

[21] BC, “Selected Economic Indicators”, Brazilian Central Bank. Accessed: Jun. 17, 2024. [Online]. Available: https://www.bcb.gov.br/en/statistics/selectedindicators

[22] J. Wu, N. Levi, R. Araujo, and Y. G. Wang, “An evaluation of the impact of COVID-19 lockdowns on electricity demand”, *Electric Power Systems Research,* vol. 216, p. 109015, Mar. 2023, doi: 10.1016/J.EPSR.2022.109015.

[23] M. Carvalho, D. Bandeira de Mello Delgado, K. M. de Lima, M. de Camargo Cancela, C. A. dos Siqueira, and D. L. B. de Souza, “Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns*”, Int J Energy Res*, vol. 45, p. 3358–3364, Feb. 2021, doi: 10.1002/ER.5877.

[24] V. Mahler, R. Girard, and G. Kariniotakis, “Data-driven structural modeling of electricity price dynamics”, *Energy Econ,* vol. 107, p. 105811, Mar. 2022, doi: 10.1016/J.ENECO.2022.105811.

[25] L. Mitridati and P. Pinson, “A Bayesian Inference Approach to Unveil Supply Curves in Electricity Markets”, *IEEE Transactions on Power Systems,* vol. 33, p. 2610–2620, May 2018, doi: 10.1109/TPWRS.2017.2757980.

[26] C. Ruiz, A. J. Conejo, and D. J. Bertsimas, “Revealing Rival Marginal Offer Prices Via Inverse Optimization”, *IEEE Transactions on Power Systems,* vol. 28, p. 3056–3064, Aug. 2013, doi: 10.1109/TPWRS.2012.2234144.

[27] R. Chen, I. Ch. Paschalidis, M. C. Caramanis, and P. Andrianesis, “Learning from Past Bids to Participate Strategically in Day-Ahead Electricity Markets”, *IEEE Trans*

Smart Grid, vol. 10, p. 5794–5806, Sep. 2019, doi: 10.1109/TSG.2019.2891747.

[28] F. Saâdaoui, S. Mefteh-Wali, and S. Ben Jabeur, "Multiresolutional statistical machine learning for testing interdependence of power markets: A Variational Mode Decomposition-based approach", *Expert Syst Appl,* vol. 208, p. 118161, Dec. 2022, doi: 10.1016/J.ESWA.2022.118161.

[29] EPU, "Geopolitical Risk Index", Economic Policy Uncertainty Index. Accessed: Jun. 17, 2024. [Online]. Available: https://www.policyuncertainty.com/gpr.html

[30] S. Awaworyi-Churchill, J. Inekwe, K. Ivanovski, and R. Smyth, "Breaks, trends and correlations in commodity prices in the very long-run", *Energy Econ,* vol. 108, p. 105933, Apr. 2022, doi: 10.1016/J.ENECO.2022.105933.

[31] F. Saâdaoui and S. Ben Jabeur, "Analyzing the influence of geopolitical risks on European power prices using a multiresolution causal neural network", *Energy Econ,* vol. 124, p. 106793, Aug. 2023, doi: 10.1016/J.ENECO.2023.106793.

[32] O. A. Adeosun, M. I. Tabash, X. V. Vo, and S. Anagreh, "Uncertainty measures and inflation dynamics in selected global players: a wavelet approach", *Qual Quant,* vol. 57, p. 3389–3424, Aug. 2023, doi: 10.1007/S11135-022-01513-7/TABLES/10.

[33] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, "Streaming feature selection algorithms for big data: A survey", *Applied Computing and Informatics,* vol. 18, p. 113–135, Jan. 2022, doi: 10.1016/J.ACI.2019.01.001/FULL/PDF.

[34] M. Rong, D. Gong, and X. Gao, "Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends", *IEEE Access,* vol. 7, p. 19709–19725, Jan. 2019, doi: 10.1109/ACCESS.2019.2894366.

[35] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial", *Mach Learn,* vol. 113, p. 3917–3928, Jul. 2021, doi: 10.1007/S10994-021-06042-2/METRICS.

[36] L. Santos Coelho, H. V. Hultmann Ayala, and V. Cocco Mariani, "CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering", *Fuel*, vol. 355, p. 129366, Jan. 2024, doi: 10.1016/J.FUEL.2023.129366.

[37] Z. Wang, L. Xia, H. Yuan, R. S. Srinivasan, and X. Song, "Principles, research status, and prospects of feature engineering for data-driven building energy prediction: A comprehensive review", *Journal of Building Engineering,* vol. 58, p. 105028, Out. 2022, doi: 10.1016/J.JOBE.2022.105028.

[38] J. Wang, Y. Dong, and J. Liu, "A novel multifactor clustering integration paradigm based on two-stage feature engineering and improved bidirectional deep neural networks for exchange rate forecasting", *Digit Signal Process,* vol. 143, p. 104258, Nov. 2023, doi: 10.1016/J.DSP.2023.104258.

[39] G. Yan *et al*., "A comparative study of machine learning models for respiration rate prediction in dairy cows: Exploring algorithms, feature engineering, and model interpretation", *Biosyst Eng,* vol. 239, p. 207–230, Mar. 2024, doi: 10.1016/J.BIOSYSTEMSENG.2024.01.010.

[40] Y. Q. Wang *et al*., "Machine learning framework for intelligent aeration control in wastewater treatment plants: Automatic feature engineering based on variation sliding layer", *Water Res,* vol. 246, p. 120676, Nov. 2023, doi: 10.1016/J.WATRES.2023.120676.

[41] D. Dai *et al*., "Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys", *Comput Mater Sci*, vol. 175, p. 109618, Apr. 2020, doi: 10.1016/J.COMMATSCI.2020.109618.

[42] L. Jiang and G. Hu, "Day-Ahead Price Forecasting for Electricity Market using Long-Short Term Memory Recurrent Neural Network", in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Nov. 2018, p. 949–954. doi: 10.1109/ICARCV.2018.8581235.

[43] Y. Sun, F. Haghighat, and B. C. M. Fung, "A review of the-state-of-the-art in data-driven approaches for building energy prediction", *Energy Build*, vol. 221, p. 110022, Aug. 2020, doi: 10.1016/J.ENBUILD.2020.110022.

[44] N. Kim, H. Shin, and K. Lee, "Feature engineering process on well log data for machine learning-based SAGD performance prediction", *Geoenergy Science and Engineering*, vol. 229, p. 212057, Out. 2023, doi: 10.1016/J.GEOEN.2023.212057.

[45] Y. Li *et al*., "A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering", *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103269, Apr. 2023, doi: 10.1016/J.JAG.2023.103269.

[46] O. Björneld, M. Carlsson, and W. Löwe, "Case study - Feature engineering inspired by domain experts on real world medical data", *Intell Based Med*, vol. 8, p. 100110, Jan. 2023, doi: 10.1016/J.IBMED.2023.100110.

[47] J. Zheng, J. Zhu, and H. Xi, "Short-term energy consumption prediction of electric vehicle charging station using attentional feature engineering and multi-sequence stacked Gated Recurrent Unit", *Computers and Elect Eng,* vol. 108, p. 108694, May 2023, doi: 10.1016/J.COMPELECENG.2023.108694.

[48] D. Elavarasan, P. M. Durai Raj Vincent, K. Srinivasan, and C. Y. Chang, "A Hybrid CFS Filter and RF-RFE Wrapper-Based Feature Extraction for Enhanced Agricultural Crop Yield Prediction Modeling", *Agriculture 2020,* Vol. 10, Page 400, vol. 10, p. 400, Set. 2020, doi: 10.3390/AGRICULTURE10090400.

[49] M. Zulfiqar, M. Kamran, M. B. Rasheed, T. Alquthami, and A. H. Milyani, "A hybrid framework for short term load forecasting with a navel feature engineering and adaptive grasshopper optimization in smart grid", *Appl Energy,* vol. 338, p. 120829, May 2023, doi: 10.1016/J.APENERGY.2023.120829.

[50] A. Forootani, M. Rastegar, and A. Sami, "Short-term individual residential load forecasting using an enhanced machine learning-based approach based on a feature engineering framework", *Electric Power Systems Research*, vol. 210, p. 108119, Sept. 2022, doi: 10.1016/J.EPSR.2022.108119.

[51] H. Ebrahimian, S. Barmayoon, M. Mohammadi, and N. Ghadimi, "The price prediction for the energy market based on a new method", *Economic Research*, vol. 31, p. 313–337, Jan. 2018, doi: 10.1080/1331677X.2018.1429291.

[52] W. Gao and H. Pan, "Multi-Label Feature Selection Based on Min-Relevance Label", *IEEE Access,* vol. 11, p. 410–420, Dec. 2023, doi: 10.1109/ACCESS.2022.3231871.

[53] O. Abedinia, N. Amjady, and H. Zareipour, "A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems", *IEEE Transactions on Power Systems,* vol. 32, p. 62–74, Jan. 2017, doi: 10.1109/TPWRS.2016.2556620.

[54] N. Amjady and F. Keynia, "Day-Ahead Price Forecasting of Electricity Markets by Mutual Information Technique and Cascaded Neuro-Evolutionary Algorithm", *IEEE Transactions on Power Systems*, vol. 24, p. 306–318, Feb. 2009, doi: 10.1109/TPWRS.2008.2006997.

[55] C. Zhang, L. Cao, and A. Romagnoli, "On the feature engineering of building energy data mining", *Sustain Cities Soc,* vol. 39, p. 508–518, May 2018, doi: 10.1016/J.SCS.2018.02.016.

[56] D. Amengual, E. Sentana, and Z. Tian, "Gaussian Rank Correlation and Regression", in *Essays in Honor of M. Hashem Pesaran: Panel Modeling, Micro Applications, and Econometric Methodology*, vol. 43B, A. Chudik, C. Hsiao, and A. Timmermann, Eds., Emerald Publishing Limited, p. 269–306, Jan. 2022, doi: 10.1108/S0731-90532021000043B012.

[57] B. Everitt and A. Skrondal, *The Cambridge Dictionary of Statistics*, 4th ed. Cambridge, UK; New York: Cambridge University Press, 2010.
[58] S. Li, P. Wang, and L. Goel, "A Novel Wavelet-Based Ensemble Method for Short-Term Load Forecasting with Hybrid Neural Networks and Feature Selection", *IEEE Transactions on Power Systems,* vol. 31, p. 1788–1798, May 2016, doi: 10.1109/TPWRS.2015.2438322.

Everthon Taghori Sica received his B. Sc. Ins Electrical Engineering from the Federal Technological University of Paraná. He is a Doctor in Electrical Engineering from the Federal University of Santa Catarina. Currently, Everthon is an Associate Professor at the Electrical Engineering Department of the Federal Institute of Santa Catarina. His research lines are integrated energy resources planning and economic theory applied to the electric sector.

Larah Farias Rodrigues Barboza is a graduate in Electronics Engineering at the Federal Institute of Santa Catarina. Currently, Larah is a researcher at the Intelligent Energy Systems Unit of the Innovation Hub linked to the Federal Institute of Santa Catarina. Her research interests include scientific computing, big data and power systems.

João Vitor Russi Benedet is a graduate in Electrical Engineering at the Federal Institute of Santa Catarina. Currently, João is a researcher at the Intelligent Energy Systems Unit of the Innovation Hub linked to the Federal Institute of Santa Catarina. His research interests include scientific computing, electricity market pricing, and power system planning.

Karem Vieira Paes de Lima is a graduate in Electrical Engineering at the Federal Institute of Santa Catarina. Currently, Karem is a researcher at the Intelligent Energy Systems Unit of the Innovation Hub linked to the Federal Institute of Santa Catarina. Her research interests include scientific computing, electricity market pricing, and power system planning.

Vinicius V. L. Albani received the B. Math. degree from the Federal University of Rio de Janeiro and the M.Sc. degree from the Institute for Pure and Applied Mathematics (IMPA). He is a Doctor in Mathematics also from IMPA. Currently, Vinicius is an Assistant Professor in the Mathematics Department at the Federal University of Rio de Janeiro. His research lines include mathematical methods in quantitative finance and inverse problems.

Erinaldo Santos received his B. Sc. and M. Sc. degrees in Electrical Engineering from the State University of Campinas. Currently, Erinaldo is the Head of the electricity marketing team at Urca Trading Inc. His interests include economic policy, purchasing, natural gas, electricity market and scientific computing.

Sérgio Luciano Avila. received his B. Sc. in Electrical Engineering from the University of Blumenau. He is a Doctor in Electrical Engineering with a double degree from the Ecole Centrale de Lyon/France and the Federal University of Santa Catarina. Currently, Sérgio is an Associate Professor at the Electrical Engineering Department of the Federal Institute of Santa Catarina. His research lines are scientific computing for industry and power systems.