

Interpretable Power Grid Overload Detection with Information Flow-based Fuzzy Cognitive Maps

Konstantinos Parginos[✉], Marios Tyrovolas[✉], *Member, IEEE*, Ricardo J. Bessa[✉], *Fellow, IEEE*, X. San Liang[✉], Chrysostomos Stylios[✉], *Senior Member, IEEE*, Simon Camal[✉], and George Kariniotakis[✉], *Senior Member, IEEE*

Abstract—The increasing integration of renewable energy sources (RESs) brings significant challenges, such as grid overloading, which can escalate operational costs and bring a risk of system-wide blackouts. Traditional AI-based solutions often fall short due to their “black-box” nature, while many existing eXplainable AI (XAI) methods either provide unreliable post-hoc explanations or compromise reliability and generalizability due to their vulnerability to spurious correlations. This study introduces a novel framework based on recently introduced Information Flow-based Fuzzy Cognitive Maps (IF-FCMs), a causal eXplainable Artificial Intelligence (CXAI) approach designed to overcome these limitations. By employing a two-layered architecture, this framework not only detects grid overloads but also identifies both local and global factors causing these events. Tested on the Medium Voltage Oberrhein network, IF-FCMs demonstrate predictive performance comparable to state-of-the-art models such as Explainable Boosting Machines (EBM) and Logistic Regression (LR), while offering superior interpretability. These strengths establish IF-FCMs as a reliable and powerful tool for future energy management systems.

Index Terms—Causality, explainable, fuzzy cognitive maps, grid overloading, interpretability, power grid, XAI.

I. INTRODUCTION

AS the energy sector progresses towards sustainability, integrating renewable energy sources (RESs), such as solar and wind, into power systems has become crucial for

decarbonization [1]. However, the inherent stochastic nature of RES poses significant challenges, including voltage fluctuations, grid congestion, and line overloading, which can escalate operational costs and increase the risk of system-wide blackouts [2], [3].

To ensure reliable power system operation, early detection of line overloads is critical [4]. Recent advances in Artificial Intelligence (AI) have shown promise in addressing this issue. For instance, cascade backpropagation neural networks (CBPNs) and genetic algorithm-optimized backpropagation networks have been successfully employed to predict line overloads and recommend corrective actions [5]–[7]. Additionally, Graph Neural Networks (GNNs) have demonstrated effectiveness in identifying critical lines, enhancing the efficiency of solving the Optimal Power Flow (OPF) problem by simplifying constraints, and reducing computation time [8].

Despite their promising features, these AI methods often function as “black boxes” that provide no insights into their decision-making processes. This lack of interpretability hinders operator trust, particularly in critical sectors such as power systems [9]. To bridge this gap, the field of Explainable Artificial Intelligence (XAI) has emerged, focusing on enhancing the transparency and interpretability of AI models [10].

A. State-of-the-art & Motivation

In power grid management and fault detection, XAI techniques are increasingly utilized to provide transparency in AI decision-making. These techniques enable operators to identify key factors behind predicted faults, facilitating the development of effective mitigation strategies [11]. For instance, [12] employed interpretable linear models, including Ridge Regression and Logistic Regression (LR), as well as non-linear Multi-layer Perceptrons (MLPs) combined with the post-hoc explanation method Integrated Gradients (IG), to predict and interpret power grid faults in a vulnerable Arctic grid in Northern Norway. Linear models provided a global view of the marginal contributions of power-related variables, while IG complemented this by explaining the features influencing specific predictions.

Other studies [13]–[15] have explored post-hoc methods, such as Shapley Additive Explanations (SHAP), Anchors, or Diverse Counterfactual Explanations (DiCE), to enhance the interpretability of black-box neural networks used for power system fault detection. These methods revealed how specific

Manuscript received June 25, 2024; revised January 24, 2025; accepted February 11, 2025. Date of online publication May 16, 2025; date of current version July 27, 2026. This work was partially supported by the Marie Skłodowska-Curie COFUND project AI4theSciences (Grant No. 945304) at PSL University. M. Tyrovolas is funded by Next Generation EU through the Greece 2.0 - National Recovery, and Resilience Plan, under the project “Greece4.0 - Network of Excellence for developing, disseminating, and implementing digital transformation technologies in Greek Industry” (Project Code: TAEDR-0535864).

K. Parginos (corresponding author, email: konstantinosparginos@gmail.com), S. Camal, and G. Kariniotakis are with the Center PERSEE, MINES Paris, PSL University, Sophia Antipolis 06560, France.

M. Tyrovolas and C. Stylios are with the Department of Informatics and Telecommunications, University of Ioannina, Arta 47150, Greece, and the Industrial Systems Institute, Athena Research Center, Patras 26504, Greece.

R. J. Bessa is with the Center of Power, and Energy Systems, INESC TEC, Porto 4200-465, Portugal, as well as with the Center PERSEE, MINES Paris, PSL University, Sophia Antipolis 06560, France.

X. S. Liang is with the Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai 200438, China, and with the Artificial Intelligence Group, Division of Frontier Research, Southern Marine Laboratory, Zhuhai 519099, China.

DOI: 10.17775/CSEEJPES.2024.04780

features influenced fault detection outcomes, aiding operators in identifying fault types, locations, and conditions affecting prediction stability.

From a technical viewpoint, post-hoc methods explain complex models by approximating their behavior locally using interpretable surrogate models or applying game-theoretic approaches to assess input variable importance. Despite their usefulness, these methods face significant limitations. Surrogate models may inaccurately represent the underlying black-box models, sometimes also relying on different variables, leading to misleading explanations [16]. Additionally, post-hoc explanations can be vulnerable to manipulation, potentially providing seemingly credible interpretations even for biased models [17]. Furthermore, the assumption of variable independence in many post-hoc methods often results in incorrect explanations when variables are interdependent [18].

These limitations have increased interest in intrinsically interpretable models, which are transparent by design and provide explanations alongside their predictions [19]. For instance, a recent study employed the Explainable Boosting Machine (EBM) to detect faults in electrical power transmission systems [20]. EBM combines the predictive accuracy of boosting algorithms with the interpretability of Generalized Additive Models. The study demonstrated EBM's superior performance compared to models like Random Forest (RF), Support Vector Classifier (SVC), and Long Short-Term Memory (LSTM) networks. Additionally, EBM's interpretability revealed key factors influencing fault detection, such as phase-line currents and voltage fluctuations, enhancing both user trust and practical applicability. However, current intrinsically interpretable models often rely on correlative-based learning algorithms, which struggle to distinguish spurious correlations from authentic causal relationships, potentially compromising model reliability and generalizability [21]. Consequently, there is a growing need for XAI systems that explicitly account for causality when explaining predictions [10], [22].

To address these challenges, causal eXplainable Artificial Intelligence (CXAI) has emerged as a promising approach. CXAI focuses on developing intrinsically interpretable models that incorporate causal relationships among variables and provide explanations grounded in causality rather than mere correlations [23]. By doing so, CXAI eliminates erroneous decisions and explanations caused by spurious correlations [24]. These models leverage causal discovery algorithms to identify causal structures in data, aligning their reasoning with human thought processes and advancing toward human-level artificial intelligence [25]. Recent research by Google DeepMind highlights the importance of causal reasoning for generalizing beyond initial datasets, positioning causal AI as a critical step toward achieving artificial general intelligence (AGI) [26].

In the existing literature, only one work has applied CXAI to the power systems domain. Specifically, [27] introduced the eXplainable Causal Graph Neural Network (X-CGNN) for electricity demand forecasting. In detail, X-CGNN integrated multivariate time-series causal inference with graph neural networks, achieving superior performance in forecasting electricity demand at household and distribution levels. Empirical

evaluation of its performance revealed that it outperformed traditional models, such as LSTMs, Random Forests, and transformer-based architectures, significantly reducing mean absolute error and root mean square error. Additionally, X-CGNN delivered both global and local causal explanations, offering valuable insights into the relationships within the data. Nevertheless, to the best of our knowledge, no CXAI applications have been reported for power grid overload detection and diagnosis, representing a critical gap in the field.

A recent advancement in CXAI is the introduction of *Information Flow-based Fuzzy Cognitive Maps* (IF-FCMs) by Tyrovolas *et al.* [28]. These models build on traditional Fuzzy Cognitive Maps (FCMs) [29], which can be viewed as interpretable recurrent neural networks represented as directed causal graphs. FCMs are valued for their intuitive structure, ability to integrate expert knowledge and data-driven insights, transferability, transparency, and inherent interpretability [30]. IF-FCMs extend FCMs by incorporating Liang-Kleeman Information Flow (L-K IF) analysis to identify authentic causal relationships from observational data and use them as constraints during learning to ultimately enhance their predictive and explanatory power by excluding spurious correlations. IF-FCMs have shown superior performance in applications such as fault detection and diagnosis in industrial systems and robotics [31].

B. Contribution

Given the current absence of CXAI-based methods for power grid overload detection and diagnosis, and the potential of recently introduced IF-FCMs, this paper proposes a novel two-layered IF-FCM architecture for such tasks. By leveraging the interpretability and causal reasoning of IF-FCMs, the proposed method not only identifies overload events but also pinpoints their root causes, thus allowing operators to make proactive decisions and implement targeted mitigation strategies. From the FCM perspective, this study extends the application of this new FCM model to broader systems and data, validating its potential. In addition, unlike previous works that focused on comparisons with pure FCM models or their extensions, this study provides a comprehensive empirical evaluation against state-of-the-art interpretable models, including EBM and LR. Using the generic medium-voltage Ober-rhein network, numerical simulations show that the proposed framework achieves predictive accuracy comparable to that of EBM and LR while offering superior interpretability.

The remainder of this paper is organized as follows. Section II outlines the theoretical and technical foundations of FCMs, L-K IF analysis, and IF-FCMs and introduces an innovative two-layered IF-FCM architecture for power grid overload detection and diagnosis. Section III presents the case study and experimental results for evaluating the proposed method. Lastly, Section IV summarizes the main findings and describes the future research directions.

II. METHODOLOGY

This section presents the theoretical and technical foundations of the proposed IF-FCM framework and then explains its two-layered architecture.

A. Fuzzy Cognitive Maps

Fuzzy Cognitive Maps (FCMs), introduced by Kosko [29], constitute a soft computing methodology for modeling and simulating dynamic complex systems and carrying out predictive tasks, such as pattern classification, multi-output regression, and time-series forecasting [32]. FCMs are represented as directed graphs with n nodes or concepts, C_i ($i \in \{1, 2, \dots, n\}$) denoting the system components, and weighted edges describing causal interactions (Fig. 1). The weights $w_{ij} \in [-1, 1]$ represent the strength and kind of causality between concepts C_i and C_j .

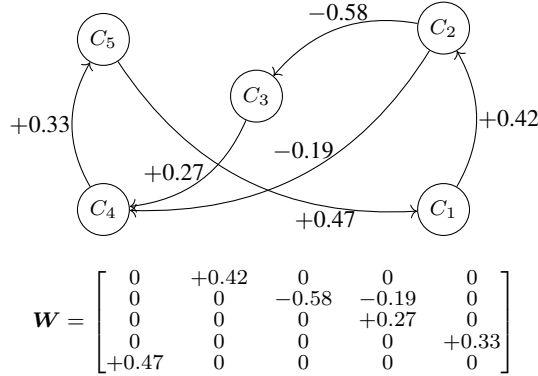


Fig. 1. A sample fuzzy cognitive map.

Each concept C_i in an FCM has an activation value $A_i^{(t)} \in [0, 1]$ or $[-1, 1]$ at iteration t of the FCM reasoning process, reflecting its current activation level. These values are iteratively updated by using a reasoning rule, with its generalized form being

$$A_i^{(t+1)} = \underbrace{\varphi f\left(\sum_{\substack{j=1 \\ j \neq i}}^n A_j^{(t)} w_{ji}\right)}_{\text{nonlinear component}} + \underbrace{(1 - \varphi) A_i^{(0)}}_{\text{linear component}} \quad (1)$$

where $f(\cdot)$ is the activation function that normalizes the activation values to the allowed interval, $\varphi \in [0, 1]$ controls the nonlinearity of the reasoning rule, and $A_i^{(0)}$ is the initial activation value of the i^{th} concept provided by domain experts or derived from data.

Common activation functions in the FCM domain include bivalent, trivalent, hyperbolic tangent, and sigmoid functions, which may affect FCM reliability [33]. To address these issues, Nápoles *et al.* [34] introduced a re-scaled activation function as follows:

$$f(\bar{A}_i^{(t)}) = \frac{\bar{A}_i^{(t)}}{\|\bar{A}^{(t)}\|_2}, \quad \bar{A}^{(t)} \neq \vec{0} \quad (2)$$

where $\bar{A}^{(t)} = [\bar{A}_1^{(t)}, \bar{A}_2^{(t)}, \dots, \bar{A}_n^{(t)}]$ is the raw state vector given by $\bar{A}^{(t)} = A^{(t)}W$, and $\|\cdot\|_2$ is the Euclidean norm.

The FCM's recurrent reasoning process starts with an initial state vector $A^{(0)} \in \mathbb{R}^n$, which contains the initial concept activation values. Subsequently, (1) is applied iteratively to update these values in each iteration t until either (i) the FCM converges to a fixed-point attractor, where changes in

activation values fall below a threshold ε (e.g., 10^{-5}); or (ii) the process reaches a predefined maximum number of iterations T , indicating potential cyclic or chaotic behavior.

B. Liang-Kleeman Information Flow Analysis

In modern power systems, identifying causal relationships between components is essential for developing AI models that help operators detect potential risks and faults, thereby facilitating targeted maintenance and improved mitigation strategies [35]. However, previous research has primarily focused on correlations between transmission lines, overlooking causal dynamics [36].

Recent studies have attempted to address this gap by exploring causal structures in power systems. For instance, Zhang *et al.* [37] used transfer entropy to empirically estimate the information flow (IF) between variables in converter-penetrated power systems, advancing reliability analysis. IF has been widely recognized as an effective measure of causality, quantifying both the magnitude and direction of interactions [38]. More recently, Liang and Kleeman further established IF as a real physical notion derived from first principles, introducing the Liang-Kleeman Information (L-K IF) analysis. This robust framework reliably assesses causality in multivariate time-series data for both linear and nonlinear systems, even under extreme measurement noise [39]–[41]. For instance, L-K IF has demonstrated high accuracy in identifying causal relationships in noisy time-series data, with noise-to-signal ratios exceeding 100, making it particularly suitable for power systems [42].

The analytical framework is based on a d -dimensional continuous-time stochastic system, defined as:

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}, t) + \mathbf{B}(\mathbf{X}, t)\dot{\mathbf{W}} \quad (3)$$

where $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ is the state variable vector, $\mathbf{F}$ represents arbitrary nonlinear functions of $\mathbf{X}$ and time t , $\dot{\mathbf{W}}$ is a white noise vector, and $\mathbf{B}$ is the stochastic perturbation amplitude matrix.

The IF rate $T_{j \rightarrow i}$ quantifies the causality from component X_j to X_i within the system, where a nonzero value indicates a direct causal link. In the case where only d time series $X_1, X_2, \dots, X_d$ generated by this system are available, assuming linearity, the maximum likelihood estimator (MLE) of $T_{j \rightarrow i}$ is defined as [42]:

$$\hat{T}_{2 \rightarrow 1} = \frac{1}{\det \mathbf{C}} \cdot \sum_{j=1}^d \Delta_{2j} C_{j,d1} \cdot \frac{C_{12}}{C_{11}} \quad (4)$$

where C_{ij} is the sample covariance between X_i and X_j , Δ_{ij} represents cofactors of the covariance matrix $\mathbf{C} = (C_{ij})$, and $C_{i,dj} = (X_i - \bar{X}_i)(\dot{X}_j - \bar{\dot{X}}_j)$ is the covariance between X_i and the derivative of X_j approximated via the Euler forward scheme: $\dot{X}_{j,n} = (X_{j,n+k} - X_{j,n})/(k\Delta t)$, with $k \geq 1$ being some integer. Given that (4) is an MLE, its statistical significance needs to be evaluated, e.g., via the Fisher information matrix, in order to validate causality [42].

C. Information Flow-based Fuzzy Cognitive Maps

The combination of IF, characterized by mathematical rigor, physical interpretability, and computational efficiency, with the strengths of FCMs, has led to the development of IF-FCMs. By preventing spuriousness during training, IF-FCMs achieve superior predictive and explanatory power compared to traditional FCMs on both synthetic and real-world datasets [31]. This section outlines the structure, training process, and interpretability of IF-FCMs.

As shown in Fig. 2, constructing an IF-FCM starts by mapping problem variables (e.g., power lines and overload targets) as distinct model concepts. Subsequently, the model is subjected to a two-phase learning process: 1) *Training Phase 1*, which involves identifying statistically significant IF rates through L-K IF analysis, revealing causal relationships between variables, and 2) *Training Phase 2*, which involves fine-tuning model weights and parameters of the activation function and reasoning rule to fit observed data. Unlike traditional approaches, IF-FCMs adjust only the weights of edges with statistically significant IF rates, setting others to zero *a priori*. This ensures the model captures genuine system interactions rather than mere correlations.

Originally, Particle Swarm Optimization (PSO) is used for Phase 2. However, due to PSO's limitations in high-dimensional problems and susceptibility to multiple local optima, this study adopts Social Learning PSO (SL-PSO) [43]. This variant enhances standard PSO by allowing particles (candidate solutions) to learn from multiple top-performing peers (demonstrators), promoting diverse exploration and avoiding premature convergence. Its dimension-dependent parameter control strategy adapts the influence of demonstrators based on problem dimensionality, maintaining solution diversity for large-scale problems and thereby balancing the exploration

and exploitation phases of the optimization algorithm. Additionally, SL-PSO requires minimal hyperparameter tuning and achieves lower time complexity across various dimensions.

In IF-FCM training, SL-PSO optimizes the solution $x = [\varphi, w^{(1)}, w^{(2)}, \dots, w^{(\text{SIFs})}]$, where SIFs represents the number of statistically significant IF rates, to minimize the objective function:

$$\mathcal{E}(x) = \alpha_1 \underbrace{\frac{1}{K} \sum_{k=1}^K |Y_k - A_{nk}^{(l)}|}_{G(x)} + (1 - \alpha_1) \underbrace{\sum_{k=1}^K \sum_{i=1}^n \sum_{t=1}^l \frac{2\omega_t (A_{ik}^{(t)} - A_{ik}^{(t-1)})^2}{Kn(T-1)}}_{H(x)} \quad (5)$$

where $G(\cdot)$ measures the mean absolute prediction error, while $H(\cdot)$ evaluates model convergence to a fixed-point attractor by assessing the dissimilarity of successive state vectors $A_k^{(t)}$. Here, K and n denote the number of training instances and IF-FCM concepts, respectively, while $\omega_t = \frac{t}{T}$ linearly increases the importance of iterations, prioritizing convergence in later stages and enhancing model flexibility during early iterations. Finally, the coefficient $\alpha_1 \in [0, 1]$ balances the contributions of $G(\cdot)$ and $H(\cdot)$ in the objective function $\mathcal{E}$.

After training, IF-FCMs can perform prediction tasks, including binary classification (e.g., overload detection). In such problems, each k^{th} data instance is encoded as an initial state vector $A_k^{(0)} = [A_{1k}^{(0)}, \dots, A_{nk}^{(0)}] = 0$ which is fed into the model. The IF-FCM then executes its recurrent reasoning process, as detailed in Section II-A, with the final activation value of the output concept determining the predicted class. For this purpose, the activation interval is divided into m

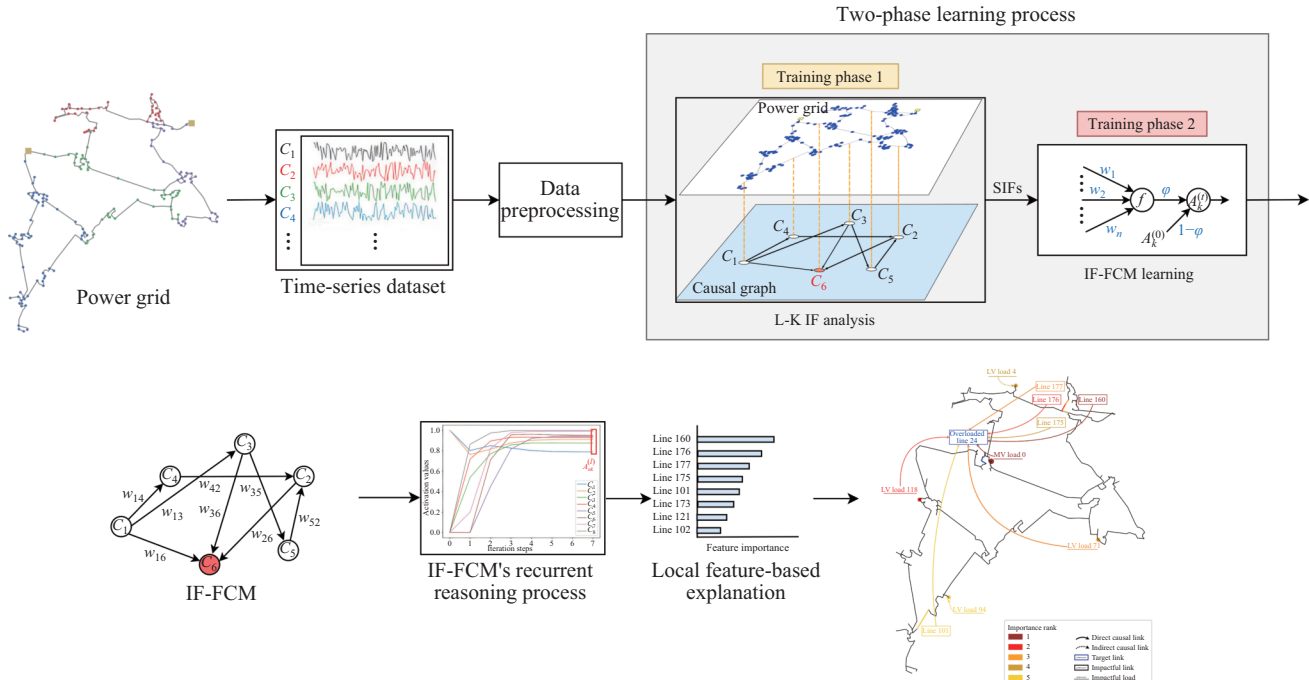


Fig. 2. Graphical methodology.

partitions corresponding to the number of classes, by means of $m - 1$ decision thresholds. For binary classification, a single threshold is optimized to balance true-positive and true-negative rates, maximizing the geometric mean.

$$g = \sqrt{\text{Precision} \times \text{Recall}} \quad (6)$$

Beyond predictions, IF-FCMs provide both global and local explanations due to their inherent interpretability. Global interpretability identifies the most important concepts for the whole model, whereas local interpretability focuses on individual predictions. For instance-level explanations, the final activation values of the concepts indicate their influence, with higher absolute values signifying greater importance. This makes it particularly useful for tasks such as overload detection, which enables the identification of input features responsible for specific outcomes. Figure 2 illustrates the development process of IF-FCMs, spanning from data collection to decision explanations.

D. Two-layered IF-FCM Architecture

This subsection introduces the primary contribution of this paper: a two-layered architecture based on IF-FCMs for detecting and diagnosing grid overloads. The model's strength lies in its ability to provide feature attributions for detected overloads, leveraging its built-in causal reasoning. Specifically, each layer of the framework is designed to address a distinct aspect of the problem:

- **Layer 1:** It detects overloads and identifies specific power lines contributing to the overload events.

- **Layer 2:** It detects overloads and analyzes interconnected grid components contribute to the detected overloads, such as renewable energy sources, nodal loads, and substations on the power lines.

These two layers offer complementary insights into grid overload dynamics. On the one hand, Layer 1 focuses on localized analysis, examining how the loading of specific power lines contributes to overloading events. This provides grid operators with an immediate understanding of grid vulnerabilities at the line level. On the other hand, Layer 2 broadens the perspective by including interconnected grid components, thereby contextualizing Layer 1's findings within a wider operational environment. It identifies upstream and indirect factors that may not be apparent when analyzing individual lines alone. Therefore, by combining these layers, the proposed architecture captures both local and global factors, enabling a comprehensive and interpretable framework for diagnosing and mitigating grid stress. In this way, the analysis granularity is enhanced, thus supporting more effective grid management.

As illustrated in Fig. 3, each layer employs a dedicated IF-FCM model and a corresponding dataset tailored to its specific focus. According to the previous description, for Layer 1, the dataset includes the loading status of the power lines and the associated overloaded target line. For Layer 2, the dataset comprises low- and medium-voltage loads as well as wind farm generator outputs. The development and processing of these datasets for the adopted case study are described in detail in the following section.

The development of both IF-FCM models follows the

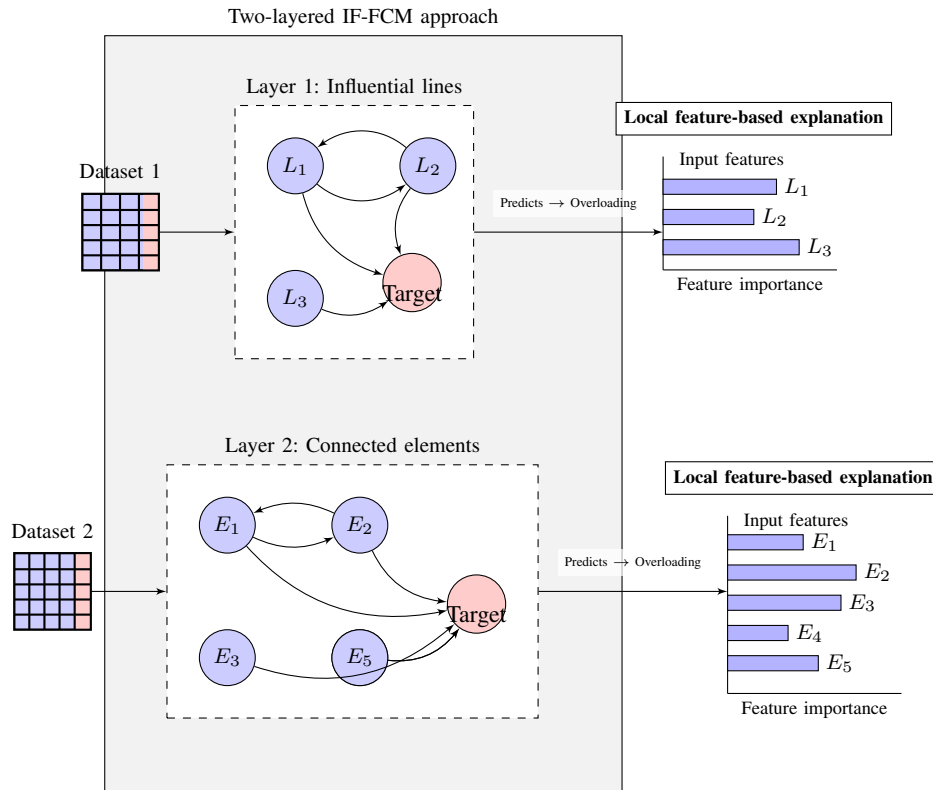


Fig. 3. Two-layered IF-FCM approach for grid overload detection and diagnosis. Layer 1 focuses on identifying key power lines, while Layer 2 examines interconnected grid components

methodology outlined in Section II-C. The process begins by mapping each problem variable to a distinct concept. Specifically, in Layer 1, each power line is represented as a unique concept, while in Layer 2, each interconnected element is assigned to a separate concept within the cognitive network. Subsequently, a two-phase learning process is performed, as detailed in Section II-C. In the first phase, L-K IF analysis is applied to perform causal inference, which identifies the causal links between power lines and the overloaded target line for Layer 1 and the causal links between interconnected elements and the overloaded line for Layer 2. The second phase involves training the IF-FCM models using the SL-PSO algorithm. This phase is guided by estimated statistically significant IF rates, ensuring robust and accurate model training, which rules out spurious correlations. Once trained, the IF-FCM models are used to detect overloading events and provide explanations based on feature attributions. In Layer 1, the explanations pinpoint the most critical power lines contributing to the overload, while in Layer 2, they highlight the interconnected grid components responsible for the detected overloads.

III. EXPERIMENTAL RESULTS

This section presents the simulation results employed to evaluate the performance of the proposed IF-FCM-based framework in comparison with state-of-the-art models. First, the case study employed in the empirical investigation is introduced, along with the data preprocessing steps performed to prepare the dataset for model development. Second, the experiments conducted and their results are described in detail. These experiments are implemented by using MATLAB and Python and executed on a desktop computer equipped with a 12-core AMD Ryzen 9 5900X CPU and 128 GB of RAM.

A. Case Study

The Medium Voltage (MV) Oberrhein open-source network, based on a real German distribution grid and provided by Pandapower [44], is used to evaluate the proposed method through simulations. To increase RES penetration and introduce technical stress, the 20 kV network is modified to include two 25 MVA HV/MV substations. The network supplies 141 MV/LV (secondary) substations and 61.86 MW loads (peak power) through four MV feeders. Notably, while the network has a meshed topology, it operates as a radial grid.

To model realistic conditions, the network includes 147 consumers, 9 wind power plant (WPP) units, 4 combined heat and power (CHP) units, and 3 storage systems. The secondary substation load data are sourced from the Iowa Distribution Test System [45], spanning 4366 hourly (60 min) intervals (t), while wind power measurements are obtained from a WPP location in France, whose location remains confidential for privacy reasons. Furthermore, power flow results are derived from simulations in [46], with hourly market intervals. As a result, this study utilizes two datasets whose features are summarized in Table I:

1) Dataset 1: 186 features representing line loading percentages across the network elements

TABLE I
DESCRIPTION OF FEATURES FOR BOTH DATASETS

Category	No. of features	Description
Input data		
Line (Dataset 1)	186	Loading (%)
LV load (Dataset 2)	141	Low Voltage loads (MW)
MV load (Dataset 2)	6	Medium voltage loads (MW)
WF (Dataset 2)	9	Wind farm generators (MW)
Target data (Both Datasets)		
Line	1	Overload status of the inspected line

2) Dataset 2: 141 features for low voltage loads (in MW), 6 features for medium voltage loads (in MW), and 9 features for wind farm generator outputs (in MW)

The datasets encompass 4360 hourly intervals, offering detailed temporal granularity, and share a common binary target variable representing the overload status of a specific inspected line. In this study, we randomly focus on overloaded Line 24 as an illustrative example. This variable is constructed through data preprocessing, which involves several key steps to prepare the binary classification datasets for the empirical study. Initially, the datasets are checked for missing values, and none are found; therefore, no imputation is required. Subsequently, features with constant values are identified and removed because they lack statistical association with other varying features; such association is a necessary condition for establishing causality. Specifically, Line 57 and Line 124 are removed from the first dataset, while CHP1, CHP2, CHP3, and CHP4 are removed from the second dataset.

Next, the focus is on identifying lines with overloading events, defined as instances in which power lines operate at or above 100% capacity. From this process, 16 overloaded lines are identified. For each, a binary target variable is created, with “1” indicating an overloaded state and “0” indicating normal operation. This results in binary classification datasets for each overloaded line: one dataset includes power line loading data and the value of the binary target variable at each time step, while the other includes low-voltage and medium-voltage loads, along with wind farm data and the binary target variable value. As previously mentioned, only the datasets for Line 24 are employed in this study.

Since such datasets typically exhibit an imbalance, with overloading events less frequent than normal operations, in this paper, we apply the Synthetic Minority Oversampling Technique and Edited Nearest Neighbors (SMOTE-ENN) algorithm [47] to address this issue. Developed by Batista *et al.* [48], this approach combines SMOTE, which generates synthetic minority-class samples through interpolation, and ENN, which removes majority-class samples that disagree with their nearest neighbors’ class, thereby refining the class boundary and reducing noisy or overlapping points. Each balanced dataset is then partitioned for 10-fold stratified cross-validation to ensure robust model training and validation [49]. In this process, each dataset is split into ten folds, maintaining the balanced ratio of overloaded to non-overloaded cases. While SMOTE-ENN and k -fold cross-validation should be used cautiously with time-series data due to potential disruptions in temporal dependencies, their use is justified here, as the study focuses

on classifying patterns rather than analyzing sequential data.

Lastly, in the context of FCMs, data fuzzification is essential for normalizing data, computing initial concept activation values, and preparing inputs for the IF-FCM model. Here, the Generalized Logistic (GL) algorithm is employed to normalize both datasets to the $[0, 1]$ range [50]. This method uses a logistic function to approximate the cumulative density function (CDF) of each feature, making it robust to outliers and well-suited to classification tasks. Unlike min–max and z -score normalization, GL normalization adjusts the data density near the mean while reducing the influence of extreme values, enhancing the dataset stability and model performance. In summary, this preprocessing strategy produced datasets tailored for effective overload prediction across power lines.

B. Comparison with State-of-the-Art Methods

To evaluate the effectiveness of the proposed IF-FCM method, we conduct a comparative analysis with state-of-the-art intrinsically interpretable models, including an Explainable Boosting Machine (EBM) [51] and Logistic Regression (LR). These models are selected to avoid the limitations of post-hoc explanation methods, as discussed in Section I-A, and due to their strong performance in power grid fault detection and diagnosis in recent literature [12], [20]. This empirical study assess both the predictive and explanatory power within the context of our case study.

To ensure optimal performance in our numerical experiments, hyperparameter tuning is first performed. For LR, we optimize the regularization strength (C) over a range of 0 to 4 and select the best optimization solver from ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’. For EBM, we adopt the default parameter configuration recommended by its developers (5000 boosting rounds and a learning rate of 0.01) [52].

The IF-FCM training follows this setup. First, L-K IF analysis is applied to identify causal relationships in the datasets: 1524 causal links are detected among power lines and the overloaded target in the first dataset, and 3409 in the second dataset. For *Training Phase 2*, the SL-PSO is implemented by using MATLAB’s Parallel Computing Toolbox with `parfor` for parallel particle evaluation. The SL-PSO algorithm is sourced from GitHub (https://github.com/JaneChenJ/SL-PSO_Matlab) and used with default parameter settings, which have shown robust performance across diverse problem scales [43].

Finally, based on previous studies [53], the objective function parameter α_1 is set to 0.8.

First, to evaluate the predictive capabilities of the EBM, LR, and IF-FCM models, we use common metrics such as accuracy, area under the curve (AUC), kappa statistic, precision, recall, and F1-score. The results in Table II indicate that the IF-FCM model achieves a predictive performance comparable to EBM and LR models, with the EBM slightly outperforming in terms of all metrics. However, the IF-FCM model stands out for its superior interpretability, which is a crucial factor in deriving actionable insights into grid management. Below, we present the experiments conducted to assess the interpretability of the three models.

TABLE II
EVALUATION METRICS FOR THE FIRST AND SECOND LAYER

Metric	EBM	LR	IF-FCM
First layer			
Accuracy	1	0.99942	0.99815
AUC	1	0.99942	0.99992
Kappa	1	0.99884	0.99630
Precision	1	0.99885	0.99633
Recall	1	1	1
F1-Score	1	0.99942	0.99816
Second layer			
Accuracy	0.99893	0.99858	0.99858
AUC	0.99890	0.99854	0.99993
Kappa	0.99798	0.99716	0.99644
Precision	0.99793	0.99725	0.99678
Recall	1	1	0.99977
F1-Score	0.99896	0.99862	0.99827

To identify the power lines and connected elements responsible for overloading in Line 24, we analyze the models’ local explanations for true positive predictions—cases where the models correctly predicted the overload. Specifically, after calculating the importance of each input feature for these instances, we compute the average local importance score for each feature across these instances and cross-validation folds, deriving the average feature importance ranking for each model. As shown in Table III, the first layer results reveal that:

- The IF-FCM model highlights Line 160, Line 176, and Line 177 as the top three features.
- The LR model emphasizes Line 54, Line 121, and Line 119.
- The EBM prioritizes Line 25, Line 33, and Line 37.

TABLE III
TOP 10 FEATURES ACROSS MODELS AND LAYERS

Local Interpretability						
Rank	First layer			Second layer		
	EBM	LR	IF-FCM	EBM	LR	IF-FCM
1	Line 25	Line 54	Line 160	MV Load 1	MV Load 0	MV Load 0
2	Line 33	Line 121	Line 176	MV Load 0	LV Load 118	LV Load 118
3	Line 37	Line 119	Line 177	LV Load 130	LV Load 36	LV Load 71
4	Line 32	Line 180	Line 175	LV Load 39	LV Load 4	LV Load 4
5	Line 50	Line 178	Line 101	LV Load 42	LV Load 61	LV Load 94
6	Line 151	Line 185	Line 173	LV Load 115	LV Load 50	LV Load 104
7	Line 48	Line 181	Line 121	LV Load 33	LV Load 24	LV Load 92
8	Line 133	Line 34	Line 102	LV Load 52	LV Load 92	LV Load 81
9	Line 143	Line 170	Line 114	LV Load 5	LV Load 136	LV Load 30
10	Line 53	Line 150	Line 161	LV Load 15	LV Load 81	LV Load 73

For the second layer:

- The IF-FCM model identifies MV Load 0, LV Load 118, and LV Load 71 as the most important.
- The LR model similarly highlights MV Load 0 and LV Load 118 but ranks LV Load 36 third.
- The EBM prioritizes MV Load 1, MV Load 0, and LV Load 130.

Figure 4 provides a visual representation of the grid topology, highlighting the top five power lines and loads that contribute to the overload in Line 24. Solid lines represent direct causal links, while dashed lines indicate indirect causal links. The color scheme reflects the importance ranking, with darker shades corresponding to higher importance.

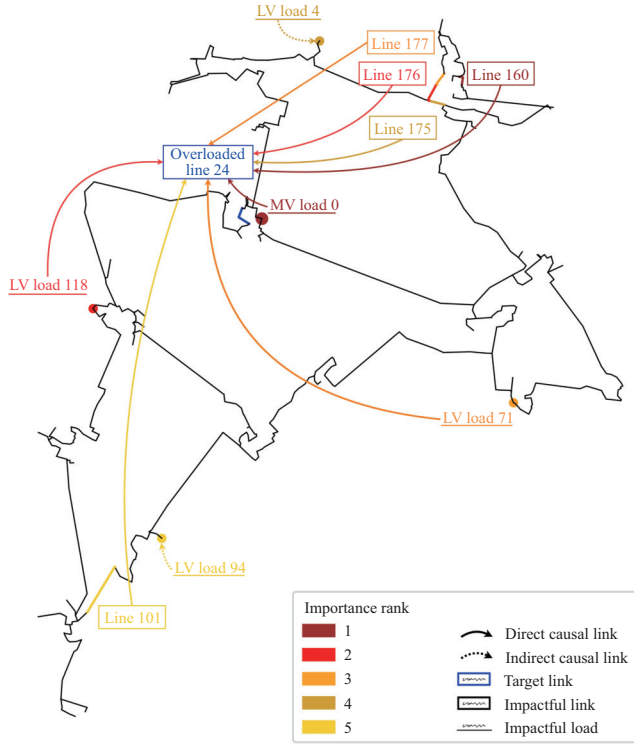


Fig. 4. Grid topology highlighting the top five power lines and loads contributing to the overload in Line 24.

To assess the first-layer results concerning power lines responsible for overloading events, an empirical data-driven analysis is initially conducted grounded in power system principles. This analysis focuses on identifying lines experiencing overloads during grid operation as they can contribute to grid stress and potentially trigger cascading failures.

The findings reveal that some lines (e.g., Lines 52 and 41) consistently operate within the load limits, while Line 24 remains persistently overloaded (Fig. 5). These patterns suggest that these lines are unlikely to be primary sources of system overload. Similarly, some lines (e.g., Lines 49, 50, and 58) display persistently high loads, indicating that they may not be major contributors to sudden overloading events. Consequently, attention is shifted towards lines that exhibit abrupt transitions from an underloaded state (below 100%) to an overloaded condition (above 100%) because such shifts are more likely to trigger cascading failures.

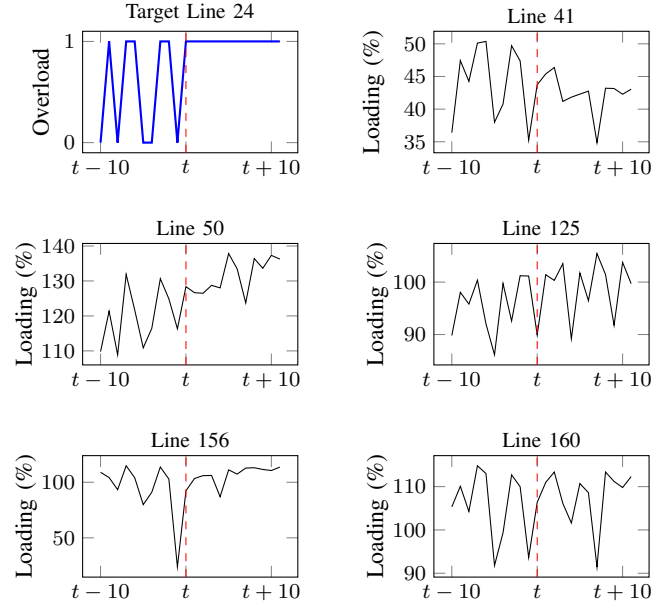


Fig. 5. Overloading event at time step #4346. The x -axis shows time steps from $t - 10$ to $t + 10$, where t denotes the time step of the overload event.

To quantify this behavior, the frequency of each line's transition from underloaded to overloaded conditions during overloading events is recorded. Model explanatory power is assessed by measuring the overlap between the lines identified in this empirical analysis and the average feature importance rankings generated by each model.

As shown in Fig. 6 and Table III, the IF-FCM model is aligned with the empirical analysis for two lines (Lines 160 and 161), with Line 161 exhibiting the most abrupt load changes during overloading events, closely followed by Line 160. In contrast, the EBM and LR models identify only one common line each (Line 50 and Line 34, respectively), both with the minimal sudden load changes. Notably, the LR model highlights Line 54 as the most significant, despite it consistently operating below 3% capacity, rendering it a non-congested and non-fluctuating line.

To further evaluate the quality of explanations provided by the XAI models in the first layer, we employ a systematic

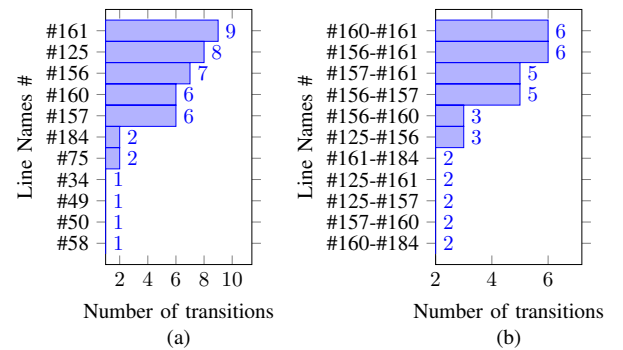


Fig. 6. Line loadability during overloading events. (a) Transitions for individual lines. (b) Transitions for line combinations. The x -axis shows the number of transitions from underloaded (below 100%) to overloaded (above 100%) states.

method inspired by the functionality-grounded evaluation approach for XAI methods, as described in [54]. Specifically, this approach evaluates the consistency between the average feature importance rankings of the interpretable models across dataset folds and the rankings produced by several feature selection methods, which served as a “ground truth”. To avoid bias towards any specific machine learning model, we select model-independent feature selection techniques. These methods also vary in their underlying mechanisms, allowing for a comprehensive and unbiased evaluation of the average feature importance rankings. The feature selection methods applied are as follows:

- **Minimal Redundancy Maximal Relevance (mRMR):** This filter-based method ranks input features based on their relevance to the target variable while minimizing redundancy among themselves, leveraging mutual information [55].
- **Univariate ROC-AUC Ranking:** This method evaluates each feature individually by calculating its AUC-ROC score, which measures its capacity to distinguish between classes. A higher AUC-ROC score indicates a stronger discriminative ability [56].
- **SVD-based Feature Ranking:** This uses Singular Value Decomposition (SVD) to perform dimensionality reduction, highlighting the features that contribute most to data variance by identifying significant principal components [57].

We exclude other filter methods, such as the Chi-Squared test and ANOVA, due to limitations, including their failure to account for associations between problem variables

and reliance on strict assumptions about data distribution and homoscedasticity [58]. Furthermore, methods such as Information Gain and Chi-Squared are often biased toward features with more unique values, while others require data discretization, introducing complexity and potential errors [59]. Our selected methods avoid these issues by capturing feature interactions, avoiding strict statistical assumptions, and revealing latent patterns in data. Thus, we extract a robust ground truth for evaluating the quality and consistency of explanations from interpretable models.

To implement this approach, we apply feature selection methods to each data fold and calculate the average feature ranking across ten folds for each method. Subsequently, for each of them, we incrementally select the top N features, increasing by 20 at each step, and calculate the overlap between these selected features and the important features identified by each XAI model (EBM, LR, and IF-FCM). This process continues until all features are included, generating a curve for each model that demonstrates how consistency improves as more features are considered.

The area under each curve (AUC) is used as a quantitative measure for comparison, with the model achieving the highest AUC considered the most consistent and, therefore, superior in terms of explanation quality. The results, illustrated in Fig. 7(a), (b), and (c), show that the proposed IF-FCM model outperforms both EBM and LR, achieving the highest AUC across all feature selection methods in the first layer. For instance, the SVD-based feature selection method shares 4 out of the top 10 ranked features with IF-FCM, 3 with LR, and none with EBM.

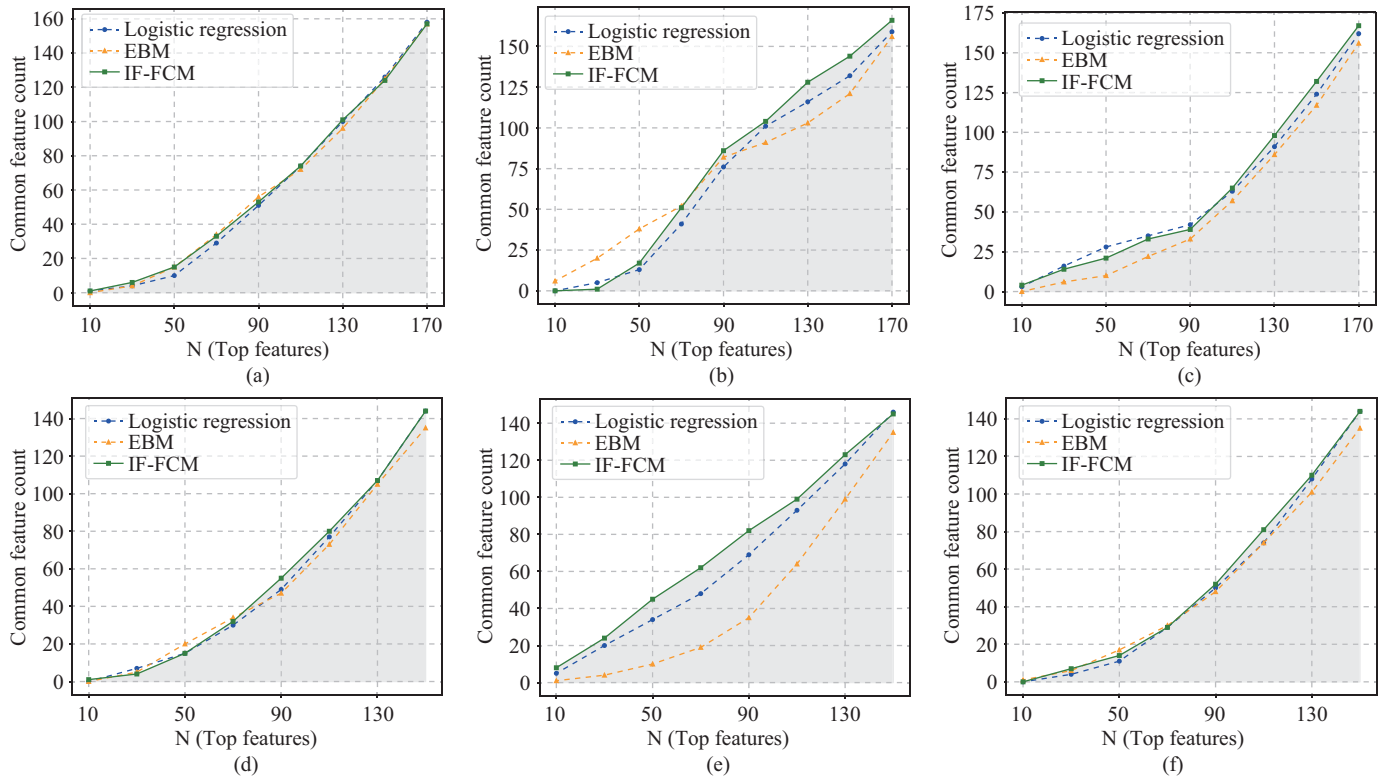


Fig. 7. Feature importance consistency of XAI models. First layer: (a) mRMR. (b) Univariate ROC-AUC. (c) SVD-based feature ranking. Second layer: (d) mRMR. (e) Univariate ROC-AUC. (f) SVD-based feature ranking.

In the second layer, the wind farm uncertainty is suspected to be the congestion source of the targeted line. IF-FCM results confirm WF6, which is closely connected to Target Line 24, as the most influential for overloading. Notably, the lowest fluctuating WF3, with a peak of 3.39 MW and a standard deviation of 0.85 MW, is deemed the least impactful by IF-FCM.

Analyzing the grid-connected loads, Fig. 8 shows the mean values in kW of grid elements when the target line is not overloaded and when it is. On the right side, MV Load 0 exhibits the highest positive deviation between non-congested and congested sets, indicating that this is the most critical element, as correctly identified by IF-FCM. Conversely, LV Load 130 showcases a significant drop, likely alleviating the target line's stress. However, EBM incorrectly ranks LV Load 130 as the most influential, conflicting with expert insights. Further analysis of LV Load 118, MV Load 1, and LV Load 4 supports these findings.

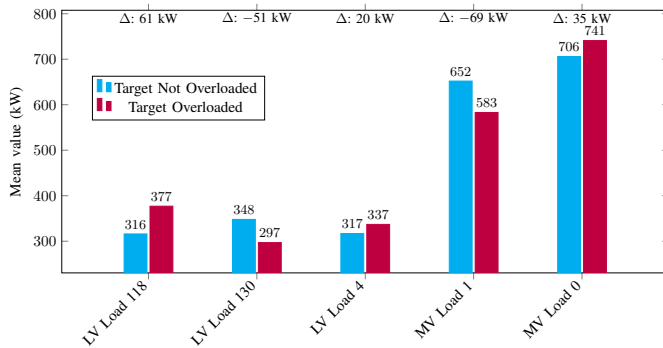


Fig. 8. Influential load identification.

Similar to the first layer, we assess the consistency between the average feature importance rankings of the compared models and the rankings produced by the feature selection methods. The IF-FCM model again demonstrates superior performance over both EBM and LR, achieving the highest AUC across all the evaluated feature selection methods (Fig. 7(d), (e), and (f)). This is especially evident with the Univariate ROC-AUC Ranking method, where the IF-FCM model shares 8 of the top 10 features with this method, indicating a high degree of alignment, while LR shares only 3 features and EBM just one.

Overall, the IF-FCM model demonstrates high predictive performance, differing from the other two models only at the third decimal place, which is negligible. However, its key advantage lies in superior interpretability. Identifying overload sources accurately is crucial for efficient grid management, as predictions alone may leave errors and inefficiencies unresolved. To validate the reliability of IF-FCM explanations, we employ empirical validation aligned with domain-specific knowledge and quantify consistency using well-established model-independent feature selection techniques, which avoid algorithm-specific biases and provide an objective benchmark, thus providing an objective benchmark. Unlike other models, IF-FCM combines high predictive accuracy with causal reasoning, enabling the efficient identification of potential

overload sources while offering an intuitive representation of system interactions. This integration makes IF-FCM a comprehensive and practical tool for power system operators.

IV. CONCLUSION

This paper introduces a novel approach for power grid overload detection using IF-FCMs. By leveraging IF-FCMs' intrinsic interpretability and incorporated causality, the proposed framework bridges the gap between traditional black-box AI models and XAI techniques, aligning with the operator's need for actionable and reliable insights. The proposed two-layered architecture not only detects grid overload events, but also identifies their root causes, such as transmission lines, renewable energy sources, and nodal loads.

The results of the Medium Voltage Oberrhein network case study underscore the robustness and effectiveness of the IF-FCM model compared to existing methods, such as EBM and LR. Specifically, IF-FCM outperforms these models in terms of interpretability, while maintaining competitive prediction accuracy. This enhanced explanatory power is critical for building operator trust and enabling targeted interventions, such as adjusting power generation schedules, shedding non-critical loads, or reconfiguring network topology to alleviate stress on overloaded lines.

From a practical perspective, IF-FCM outputs can be seamlessly fed into SCADA systems to trigger automated alarms or to provide actionable insights on operator dashboards. In Energy Management Systems (EMSs), the model can support optimization workflows by prioritizing corrective actions, such as load balancing or redispatching generation, making it a valuable tool for improving grid resilience and operational efficiency.

To build on the proposed CXAI method, future efforts will address key challenges, including efficient detection and diagnosis of overloads during unpredictable events, such as rapid fluctuations in wind or solar power generation. We will also evaluate the framework's performance under varying noise conditions, leveraging the robustness of L-K IF analysis in highly noisy environments. Moreover, to address the dynamic nature of power networks, where topology and causal relationships evolve over time, we plan to integrate a time-varying L-K formalism [60] with online FCM learning algorithms [61]. This combination will enable real-time adaptation to streaming data, immediate overload detection, and diagnosis. Finally, further optimization for sub-minute prediction times will ensure responsiveness for real-time operations.

In conclusion, the IF-FCM framework establishes a new standard for integrating CXAI into power grid analysis, providing a robust, practical, and user-friendly tool for next-generation power systems.

REFERENCES

- [1] Q. H. Wu, A. Bose, C. Singh, J. H. Chow, G. Mu, Y. Z. Sun, Z. X. Liu, Z. G. Li, and Y. Liu, "Control and stability of large-scale power system with highly distributed renewable energy generation: viewpoints from six aspects," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 1, pp. 8–14, Jan. 2023.

- [2] J. Y. Wu, B. Q. Li, L. L. Hao, F. S. Shi, and P. J. Zhao, "Adaptive emergency control of power systems based on deep belief network," *CSEE Journal of Power and Energy Systems*, vol. 10, no. 4, pp. 1618–1631, Jul. 2024.
- [3] A. Marot, A. Kelly, M. Naglic, V. Barbesant, J. Cremer, A. Stefanov, and J. Viebahn, "Perspectives on future power system control centers for energy transition," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 2, pp. 328–344, Mar. 2022.
- [4] M. Batool, F. Shahnia, and S. M. Islam, "Multi-level supervisory emergency control for operation of remote area microgrid clusters," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 5, pp. 1210–1228, Jan. 2019.
- [5] S. Balaraman and N. Kamaraj, "Cascade BPN based transmission line overload prediction and preventive action by generation rescheduling," *Neurocomputing*, vol. 94, pp. 1–12, Oct. 2012.
- [6] L. Yang, Z. M. Luo, W. Q. Lin, and S. Z. Li, "Research on line overload identification of power system based on improved neural network algorithm," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 24, pp. e5933, Dec. 2020.
- [7] R. Fainti, M. Alamaniotis, and L. H. Tsoukalas, "Three-phase line overloading predictive monitoring utilizing artificial neural networks," in *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, San Antonio, TX, USA, 2017, pp. 1–6.
- [8] T. Pham and X. P. Li, "Reduced optimal power flow using graph neural network," in *2022 North American Power Symposium (NAPS)*, Salt Lake City, UT, USA, 2022, pp. 1–6.
- [9] J. L. Cremer, I. Konstantelos, and G. Strbac, "From optimization-based machine learning to interpretable security rules for operation," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3826–3836, Sept. 2019.
- [10] K. J. Chen, Z. Y. He, S. X. Wang, J. Hu, L. C. Li, and J. L. He, "Learning-based data analytics: moving towards transparent power grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 1, pp. 67–82, Mar. 2018.
- [11] E. Sarmas, D. P. Panagoulas, G. A. Tsihrantzis, V. Marinakis, and H. Doukas, "An explainable AI-based framework for supporting decisions in energy management," in *Machine Learning Applications for Intelligent Energy Management: Invited Chapters from Experts on the Energy Field*, H. Doukas, V. Marinakis, and E. Sarmas, Eds. Cham: Springer, 2024, pp. 1–27.
- [12] O. F. Eikeland, I. S. Holmstrand, S. Bakkejord, M. Chiesa, and F. M. Bianchi, "Detecting and interpreting faults in vulnerable power grids with machine learning," *IEEE Access*, vol. 9, pp. 150686–150699, 2021.
- [13] H. N. Noura, Z. Allal, O. Salman, and K. Chahine, "Explainable artificial intelligence of tree-based algorithms for fault detection and diagnosis in grid-connected photovoltaic systems," *Engineering Applications of Artificial Intelligence*, vol. 139, pp. 109503, Jan. 2025.
- [14] C. Utama, C. Meske, J. Schneider, R. Schlatmann, and C. Ulbrich, "Explainable artificial intelligence for photovoltaic fault detection: a comparison of instruments," *Solar Energy*, vol. 249, pp. 139–151, Jan. 2023.
- [15] A. M. Reyes, A. Chengu, N. Gatsis, S. Ahmed, and M. Alamaniotis, "Model explainable AI method for fault detection in inverter-based distribution systems," in *2024 IEEE Texas Power and Energy Conference (TPEC)*, College Station, TX, USA, 2024, pp. 1–6.
- [16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019.
- [17] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 2020, pp. 180–186.
- [18] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: more accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, pp. 103502, Sept. 2021.
- [19] P. Mahya and J. Fürnkranz, "An empirical comparison of interpretable models to post-hoc explanations," *AI*, vol. 4, no. 2, pp. 426–436, May 2023.
- [20] I. Akhtar, S. Atiq, M. U. Shahid, A. Raza, N. A. Samee, and M. Abdulhafith, "Novel glassbox based explainable boosting machine for fault detection in electrical power transmission system," *PLoS One*, vol. 19, no. 8, pp. e0309459, Aug. 2024.
- [21] G. D. Xu, T. D. Duong, Q. Li, S. W. Liu, and X. Z. Wang, "Causality learning: a new perspective for interpretable machine learning," arXiv preprint arXiv: 2006.16789, 2020.
- [22] A. Marot, A. Rozier, M. Dussartre, L. Crochepierre, and B. Donnot, "Towards an AI assistant for power grid operators," in *The First International Conference on Hybrid Human Artificial Intelligence (HHAI 2022)*, 2022, pp. 1–19.
- [23] G. Carloni, A. Berti, and S. Colantonio, "The role of causality in explainable artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. e70015, Jun. 2025.
- [24] S. Y. Xu, Y. Q. Ge, and Y. F. Zhang, "Causal explainable AI," in *Machine Learning for Causal Inference*, S. Li, Z. X. Chu, Eds. Cham: Springer, 2023, pp. 137–159.
- [25] J. Pearl and D. MacKenzie, *The Book of Why*, New York: Basic Civitas Books, 2017.
- [26] J. Richens and T. Everitt, "Robust agents learn causal world models," in *The Twelfth International Conference on Learning Representations*, Vienna, Austria, 2024.
- [27] A. Miraki, P. Parviainen, and R. Arghandeh, "Electricity demand forecasting at distribution and household levels using explainable causal graph neural network," *Energy and AI*, vol. 16, pp. 100368, May 2024.
- [28] M. Tyrovolas, X. S. Liang, and C. Stylios, "Information flow-based fuzzy cognitive maps with enhanced interpretability," *Granular Computing*, vol. 8, no. 6, pp. 2021–2038, Sept. 2023.
- [29] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1, pp. 65–75, Jan. 1986.
- [30] I. D. Apostolopoulos and P. P. Groumpos, "Fuzzy cognitive maps: their role in explainable artificial intelligence," *Applied Sciences*, vol. 13, no. 6, pp. 3412, Mar. 2023.
- [31] M. Tyrovolas, C. Stylios, K. Aliev, and D. Antonelli, "Leveraging information flow-based fuzzy cognitive maps for interpretable fault diagnosis in industrial robotics," in *15th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing*, Caparica, Portugal, 2024, pp. 98–110.
- [32] M. Tyrovolas and C. Stylios, "Introducing fuzzy cognitive map for predicting engine's health status," *IFAC-PapersOnLine*, vol. 55, no. 2, pp. 246–251, Mar. 2022.
- [33] G. Nápoles, N. Ranković, and Y. Salgueiro, "On the interpretability of fuzzy cognitive maps," *Knowledge-Based Systems*, vol. 281, pp. 111078, Dec. 2023.
- [34] G. Nápoles, I. Grau, L. Concepción, L. Koutsoviti Koumeri, and J. P. Papa, "Modeling implicit bias with fuzzy cognitive maps," *Neurocomputing*, vol. 481, pp. 33–45, Apr. 2022.
- [35] I. A. Tøndel, J. Foros, S. S. Kilskar, P. Hokstad, and M. G. Jaatun, "Interdependencies and reliability in the combined ICT and power system: an overview of current research," *Applied Computing and Informatics*, vol. 14, no. 1, pp. 17–27, Jan. 2018.
- [36] S. B. Gao, J. Y. Lei, J. Shi, X. G. Wei, M. Dong, and Z. Han, "Assessment of overloading correlations among transmission lines under load redistribution attacks," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 1570–1581, Mar. 2022.
- [37] B. W. Zhang, M. Q. Wang, and W. C. Su, "Reliability interdependencies and causality assessment for a converter-penetrated power system," *IET Generation, Transmission & Distribution*, vol. 16, no. 13, pp. 2547–2558, Jul. 2022.
- [38] A. Korzeniewska, M. Mańczak, M. Kamiński, K. J. Blinowska, and S. Kasicki, "Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method," *Journal of Neuroscience Methods*, vol. 125, no. 1–2, pp. 195–207, May 2003.
- [39] X. S. Liang, "Unraveling the cause-effect relation between time series," *Physical Review E*, vol. 90, pp. 052150, Nov. 2014.
- [40] X. S. Liang, "Information flow and causality as rigorous notions *ab initio*," *Physical Review E*, vol. 94, pp. 052201, Nov. 2016.
- [41] X. S. Liang, "Causation and information flow with respect to relative entropy," *Chaos*, vol. 28, no. 7, pp. 075311, Jul. 2018.
- [42] X. S. Liang, "Normalized multivariate time series causality analysis and causal graph reconstruction," *Entropy*, vol. 23, no. 6, pp. 679, May 2021.
- [43] R. Cheng and Y. C. Jin, "A social learning particle swarm optimization algorithm for scalable optimization," *Information Sciences*, vol. 291, pp. 43–60, Jan. 2015.
- [44] L. Thurner, A. Scheidler, F. Schafer, J. H. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun, "Pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, Nov. 2018.
- [45] Z. Y. Wang. (2023, Sep.). Iowa distribution test systems. [Online] Available: <http://wzy.ece.iastate.edu/Testsystem.html>.
- [46] R. J. Bessa, F. Moaidi, J. Viana, and J. R. Andrade, "Uncertainty-aware procurement of flexibilities for electrical grid operational planning," *IEEE Transactions on Sustainable Energy*, vol. 15, no. 2, pp. 789–802, Apr. 2024.

- [47] M. Lamari, N. Azizi, N. E. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, and N. E. Benzebouchi, "SMOTE-ENN-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification," in *Advances on Smart and Soft Computing*, F. Saeed, T. Al-Hadhrani, F. Mohammed, and E. Mohammed, Eds. Singapore: Springer, 2021, pp. 37–49.
- [48] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004.
- [49] S. Purushotham and B. K. Tripathy, "Evaluation of classifier models using stratified tenfold cross validation techniques," in *4th International Conference on Global Trends in Information Systems and Software Applications*, Vellore, TN, India, 2011, pp. 680–690.
- [50] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinformatics*, vol. 17, no. 1, pp. 359, Sept. 2016.
- [51] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: a unified framework for machine learning interpretability," arXiv preprint arXiv: 1909.09223, 2019.
- [52] H. Nori, R. Caruana, Z. Q. Bu, J. H. Shen, and J. Kulkarni, "Accuracy, interpretability, and differential privacy via explainable boosting," in *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, 2021, pp. 8227–8237.
- [53] G. Nápoles, E. Papageorgiou, R. Bello, and K. Vanhoof, "On the convergence of sigmoid Fuzzy Cognitive Maps," *Information Sciences*, vol. 349–350, pp. 154–171, Jul. 2016.
- [54] G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, "Evaluating explainable artificial intelligence methods based on feature elimination: a functionality-grounded approach," *Electronics*, vol. 12, no. 7, pp. 1670, Mar. 2023.
- [55] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- [56] I. J. Johnraja, P. G. J. Leelipushpam, J. Jebaveerasingh and S. Salaja, "Experimental analysis of filtering-based feature selection techniques for fetal health classification," *Serbian Journal of Electrical Engineering*, vol. 19, no. 2, pp. 207–224, Jun. 2022.
- [57] A. D'Addabbo, M. Papale, S. Paolo, S. Magaldi, R. Colella, V. D'Onofrio, A. Palma, E. Ranieri, L. Gesualdo, and N. Ancona, "SVD based feature selection and sample classification of proteomic data," in *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*, Zagreb Croatia, 2008, pp. 556–563.
- [58] Mahto A K. (2023, May 14). Filtering out irrelevant features: a comprehensive survey of filter-based techniques for feature selection. [Online]. Available: <https://medium.com/@dancerworld60/filtering-out-irrelevant-features-a-comprehensive-survey-of-filter-based-techniques-for-feature-re-9275d86996ab>.
- [59] M. Cherrington, F. Thabtah, J. Lu, and Q. Xu, "Feature selection: filter methods performance challenges," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2019, pp. 1–4.
- [60] D. F. T. Hagan, G. J. Wang, X. S. Liang, and H. A. J. Dolman, "A time-varying causality formalism based on the Liang-Kleeman information flow for analyzing directed interactions in nonstationary climate systems," *Journal of Climate*, vol. 32, no. 22, pp. 7521–7537, Nov. 2019.
- [61] K. Wu, J. Liu, P. H. Liu, and F. Shen, "Online fuzzy cognitive map learning," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 7, pp. 1885–1898, Jul. 2021.



Konstantinos Parginos received his Diploma (5-year) degree from the Department of Electrical and Computer Engineering, University of Patras, Greece. He is currently a Ph.D. Researcher at MINES ParisTech, granted Marie Curie Scholarship for the Horizon 2020 project "AI 4 the Sciences". He is an Adjunct Professor at the International University of Monaco. In 2023, he was a Visiting Researcher at the Center for Power and Energy Systems at INESC TEC, Porto, Portugal. His research aims to enhance the interpretability of AI tools in the energy sector.



Marios Tyrovolas received the Diploma degree in Electrical and Computer Engineering from the Department of Electrical and Computer Engineering, University of Patras, Patras, Greece, in 2020, and the Ph.D. degree in Informatics and Telecommunications from the Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece, in 2025. He is currently a Post-doctoral Research Fellow with the Knowledge and Intelligent Computing (KIC) Laboratory, University of Ioannina, and a Collaborating Researcher with the Industrial Systems Institute, Athena Research Center, Patras, and with the Computer Technology Institute and Press "DIOPHANTUS," Patras. In 2024, he was a Visiting Researcher at the Artificial Intelligence Oceanography Group, Division of Frontier Research, Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China. His current research interests include fuzzy cognitive maps, causal and explainable artificial intelligence, modeling and analysis of complex systems, soft computing, large language models for causal inference and model interpretability, intelligent manufacturing, and industry 4.0 applications.



Excellence Award in 2022.

Ricardo J. Bessa received a B.Sc. degree in Electrical and Computer Engineering, an M.Sc. degree in Data Analysis and Decision Support Systems, and a Ph.D. degree in Sustainable Energy Systems from the University of Porto. He is the coordinator of the Center for Power and Energy Systems, INESC TEC. His research interests include renewable energy, energy analytics, smart grids, and electricity markets. He is Associate Editor of the *Journal of Modern Power Systems and Clean Energy*, and received the Energy Systems Integration Group (ESIG)



with the Artificial Intelligence Departments of Southern Marine Laboratory and Sun Yat-Sen University, Zhuhai, China. He is interested in a variety of interdisciplinary fields, including complex systems, information flow, causality analysis, deep learning, multi-scale modeling and simulation, predictability, turbulence, atmosphere-ocean-climate science, brain-computer interface, etc.

X. San Liang received his Ph.D. in Applied Mathematics from Harvard University, Massachusetts, USA. He has worked at Harvard University, Courant Institute, MIT, the Second Institute of Oceanography of the State Oceanic Administration, China Institute for Advanced Study, Central University of Finance and Economics, Nanjing University of Information Science and Technology, etc., and was a Team Leader of the Ninth Chinese National Antarctic Research Expedition. Currently, he is a Chair Professor at Fudan University, Shanghai, and is also affiliated



Chrysostomos Stylios is Director of the Industrial Systems Institute, Athena Research Center, Professor at the Department of Informatics and Telecommunication, University of Ioannina, Greece, and was the Founder and Director of the Laboratory of Knowledge & Intelligent Computing. Based on his pioneering research on Fuzzy Cognitive Maps (FCMs), FCMs, an XAI methodology, has been applied in discipline research areas. He has published over 200 journals, conference papers and book chapters. His research work has been acknowledged

by the scientific community, while he has received more than 8,000 citations and has an h-index of 44. His main scientific interests include: artificial intelligence, big data, data mining, soft computing, computational intelligence techniques, Industry 4.0 technologies, biosignal processing & analysis, modeling and decision support systems, knowledge-based systems, open innovation, technology transfer, educational methodologies and tools, virtual and augmentative reality. He has been a scientific coordinator in more than 40 R&D projects funded by EU and national funds. He is a senior member of IEEE and a member of IFAC's TC 8.2 and TC 5.4.



Simon Camal received his Engineering degree in Energy and Environmental Engineering from Mines Nancy, France in 2010 and obtained his Ph.D. degree at MINES ParisTech - PSL University in 2020 on forecasting and optimization of ancillary service provision by renewable energy power plants. He is currently on a tenure track at MINES Paris Center for Processes, Renewable Energies and Energy Systems (PERSEE) in Sophia Antipolis, France. His scientific interests include forecasting, optimization, and control in power system applications.



Georges Kariniotakis received his Engineering and M.Sc. degrees from Greece, in 1990 and 1992, respectively, and his Ph.D. from Ecole des Mines de Paris, France, in 1996. He is currently a Full Professor with Mines Paris, PSL University. He is the Head of the Renewable Energies and Smart Energy Systems Group, Centre PERSEE, Sophia-Antipolis, France. He has authored more than 350 scientific publications in journals and conferences. He has been involved as a participant or coordinator in more than 50 R&D projects in the fields of renewable energies and smart grids. Among them, he coordinated three major EU projects in the field of wind power forecasting and integration: Anemos (FP5), Anemos.plus (FP6), SafeWind (FP7), and the Smart4RES (H2020) project on renewable energy sources forecasting and integration. His scientific interests include forecasting, renewables integration in power systems and markets, modeling, management and planning of power systems, artificial intelligence, and data science for energy.