Granular Computing (2023) 8:2021–2038
https://doi.org/10.1007/s41066-023-00417-7

ORIGINAL PAPER

Information flow-based fuzzy cognitive maps with enhanced interpretability

Marios Tyrovolas[1] · X. San Liang[3,4] · Chrysostomos Stylios[1,2]

Received: 20 July 2023 / Accepted: 16 August 2023 / Published online: 7 September 2023

Abstract
Fuzzy Cognitive Maps (FCMs) are a graph-based methodology successfully applied for knowledge representation of complex systems modelled through an interactive structure of nodes connected with causal relationships. Due to their flexibility and inherent interpretability, FCMs have been used in various modelling and prediction tasks to support human decisions. However, a notable limitation of FCMs is their susceptibility to inadvertently capturing spurious correlations from data, undermining their prediction accuracy and interpretability. In addressing this challenge, our primary contribution is the introduction of a novel framework for constructing FCMs using the Liang-Kleeman Information Flow (L-K IF) analysis, a quantitative causality analysis rigorously derived from first principles. The novelty of the proposed approach lies in the identification of actual causal relationships from the data using an automatic causal search algorithm. These relationships are subsequently imposed as constraints in the FCM learning procedure to rule out spurious correlations and improve the aggregate predictive and explanatory power of the model. Numerical simulations validate the superiority of our method against state-of-the-art FCM-based models, thereby bolstering the reliability, accuracy, and interpretability of FCMs.

Keywords Fuzzy cognitive maps · Explainable AI · Quantitative causality · Information flow · Industry 4.0

1 Introduction

As the Industry 4.0 (I4.0) era approaches, factories come closer to advanced technologies, such as Artificial Intelligence (AI) and Industrial Internet of Things (IIoT), to significantly enhance their performance through innovative methods (Gilchrist 2016). For instance, through real-time data collection and processing, manufacturers can monitor the system condition, detect possible anomalies, and promptly inform supervisors to take action before the anomalies become severe and lead to production downtime (Wang et al. 2022). Several AI solutions, such as Support Vector Machines and Artificial Neural Networks, have been presented, demonstrating great accuracy in predicting production line malfunctions (Li et al. 2017). However, their black-box nature makes their outcome explanation challenging, which leads to supervisors' reduced trust in the AI models and, thus, hinders their deployment in critical applications where humans make the final judgment, for example, industrial anomaly detection (Rehse et al. 2019). Furthermore, the lack of transparency and interpretability hinders the efficient identification of weaknesses in AI algorithms and their subsequent improvement (Alfeo et al. 2023). Therefore, there is a need to develop AI models with transparent and interpretable behavior that can provide explanations, enabling end-users to understand

✉ Marios Tyrovolas
tirovolas@kic.uoi.gr

X. San Liang
xsliang@fudan.edu.cn

Chrysostomos Stylios
stylios@isi.gr

[1] Department of Informatics and Telecommunications, University of Ioannina, Kostaki Artas, Arta 47150, Greece

[2] Industrial Systems Institute, Athena Research Center, Patras Science Park building, Patras 26504, Greece

[3] Department of Atmospheric and Oceanic Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China

[4] The Artificial Intelligence Group, Division of Frontier Research, Southern Marine Laboratory, Tang Jia Wan, Zhuhai 519000, China

decision-making processes, explore impactful input factors, delve into model mechanics, and respond appropriately (Carletti et al. 2019).

Recently, a new research direction called eXplainable Artificial Intelligence (XAI) has emerged that deals with developing techniques, algorithms, and tools that produce human-comprehensible explanations of the decisions of AI-based systems (Adadi and Berrada 2018). These explanations can take various forms depending on the application, including *IF-THEN* rules that express input–output data relationships, visual highlighting, for example, the important parts of input images for model predictions, feature importance rankings, textual explanations, and counterfactuals, among others (Kök et al. 2023). The transition from AI to XAI is imperative for successfully integrating automated decision-making into production environments in which humans make supervision and final decisions.

In recent years, the research community has introduced two main categories of XAI methodologies based on their implementation. These categories are:

- **Post-hoc explanation methods**: External techniques that seek to explain black-box models by approximating them either globally or locally using interpretable surrogate models.
- **Intrinsic interpretable models**: Models that can explain their predictions by themselves.

However, while post-hoc methods have been valuable in interpreting complex decision processes, certain limitations have motivated a shift towards intrinsic interpretable models. These challenges arise from the fact that the surrogate model may not accurately reflect the actual behavior of the underlying black-box model, leading to misleading explanations (Rudin 2019). Furthermore, even if the surrogate model approximates well, it may rely on different features compared with the black-box model, further contributing to explanations inconsistent with the original model. Another disadvantage is that the explanations of these techniques can be easily manipulated to be acceptable through specific frameworks, even if the base model is highly biased (Slack et al. 2020). Finally, when the dataset includes interrelated features, the assumption of feature independence made by commonly used post-hoc methods such as permutation feature importance, Local Interpretable Model-agnostic Explanations (LIME) method, and SHapley Additive exPlanations (SHAP) method can result in misleading explanations (Aas et al. 2021). Thus, these disadvantages have heightened research interest in learning intrinsic interpretable models, whose decisions can be explained without additional techniques, representing assimilated knowledge in a manner consistent with human thought (Alonso et al. 2015).

In light of the challenges posed by post-hoc methods, the inherent adaptability and interpretability of fuzzy systems have emerged as promising solutions. Zadeh's foundational work on fuzzy sets has paved the way for developing fuzzy systems that can model complex systems using a higher level of abstraction in a human-understandable form (Zadeh 1965; Chen and Niou 2011). The interpretability of these systems is not merely incidental; it is a core feature rooted in their ability to capture and represent knowledge in a way that reflects human cognition and provides a detailed understanding of complex systems and their underlying dynamics (Chen and Chen 2002). This makes fuzzy systems essential in the shift from post-hoc methods to inherently interpretable models (Chen and Jian 2017). Furthermore, fuzzy systems have proven valuable in forecasting (Pant and Kumar 2022). In particular, fuzzy forecasting techniques, which leverage fuzzy logical relationships, present a novel method for predicting the behavior of complex systems (Chen and Wang 2010; Chen et al. 2006). In related work, Petri Nets, which are viewed as a tool for fuzzy modeling, face challenges such as the state explosion problem, reminiscent of the issues with black-box models (Shen et al. 2013). This issue arises when these nets become so large that their behavior becomes challenging to monitor, leading to inefficiency. Such challenges echo the problems observed with black-box models, highlighting the pressing need for models that balance complexity with clarity (Chen and Fang 2005). As the focus shifts towards intrinsic interpretable models, the fusion of fuzzy logic and advanced techniques has become evident. Merging fuzzy logic with techniques, such as neural networks and expert systems, results in neuro-fuzzy methods that provide a robust approach to knowledge representation. This fusion effectively bridges human cognition with sophisticated computational models, ensuring clarity and computational prowess (Chen et al. 2009).

Fuzzy Cognitive Maps (FCMs), a type of recurrent neural network, are widely used intrinsic interpretable models for knowledge representation. They typically integrate fuzzy logic features during development, classifying them as a neuro-fuzzy method (Kosko 1986). Specifically, FCMs are directed graphs consisting of nodes called concepts representing the components of the modeled system or conceptual entities, which can be seen as information granules, and incorporate weighted edges that describe the causal relations between them. This characteristic places FCMs within the area of granular computing, which focuses on the conceptualization and processing of information granules (Papageorgiou and Stylios 2008). FCMs find applications in modeling complex systems, including industrial systems, and in addressing prediction problems such as time-series forecasting and classification

(Wang et al. 2021; Song et al. 2011; Loia et al. 2016). FCMs offer several advantages due to their unique characteristics:

1. They can use experts' assessments when the collected data are insufficient,
2. Their graphical structure provides an intuitive representation where concepts and weights have a well-defined meaning for the system under analysis,
3. They provide feature-based explanations for their predictions, being inherently interpretable,
4. The inference process of FCMs is visually transparent, enabling users to comprehend the decision-making process leading to predictions,
5. Experts can modify the weights of FCMs to encode rules that have not yet been observed in data (e.g., a new type of fault in the manufacturing system), providing a level of flexibility unattainable in other intrinsic interpretable models,

Given the aforementioned advantages, FCMs have garnered significant interest from researchers and have proven to be extremely useful across various domains (Papageorgiou and Salmeron 2013). For instance, in the industry context, Lee et al. (1997) proposed an FCM-based model for fault diagnosis in a tank-pipeline system that successfully identified various simulated faults, whereas Stylios and Groumpos (1998) presented an FCM-based supervisor of manufacturing systems for failure detection and decision analysis. Lastly, Tirovolas and Stylios (2022) proposed FCMs as a health indicator prognostic method for engines' remaining useful life in the context of predictive maintenance. However, while the literature often highlights the interpretable nature of FCMs, this is primarily based on the clarity of their concepts and weights, rather than a demonstration of their explanatory performance. Therefore, thorough numerical simulations should be performed to determine the capabilities of FCMs to explain their decisions.

To evaluate the interpretability features of FCMs, it is crucial to understand their development processes. Currently, two fundamental methods for FCM construction are found in the literature: (a) expert-based and (b) data-driven approaches (Papageorgiou and Stylios 2008). In the expert-based method, FCM concepts and weights are determined solely based on domain experts' knowledge, which is incorporated into the model using fuzzy logic theory (Stylios and Groumpos 2004). However, this approach relies heavily on the expertise level of individuals, potentially leading to unsatisfactory performance, as experts may overlook essential aspects of the problem and assign inappropriate weight values (Song et al. 2009). On the other hand, the data-driven approaches automatically define FCM parameters from available data using learning algorithms or calculate interconnection weights as correlation coefficients between variables (Papageorgiou 2012; Czerwinski et al. 2021; Nápoles et al. 2020a). Specifically, when learning algorithms are employed without prior expert knowledge, the presence of all weights is usually assumed, leading to an over-parameterized model, or the interconnections between concepts are arbitrarily established (Nápoles et al. 2020b).

Nevertheless, the dataset may contain *spurious correlations* that can be unintentionally captured from the FCM and bias the learning process. This introduces fragility to the FCM, compromising its reliability, prediction accuracy, and interpretability (Forward 2022; Wang and Culotta 2021). Indeed, the employed learning algorithms, performing as black boxes, aim to fit the FCM to the available data based on passively observed historical correlations, which can indicate a predictive relationship among variables. However, these algorithms do not consider the semantics of the analyzed system and thus fail to distinguish between causal and spurious relationships. Therefore, without establishing appropriate constraints beforehand, these algorithms are fooled by illusory patterns and assign weight values to the corresponding edges, resulting in an FCM that does not represent the authentic system interactions; instead, it learns the correlational associations between features. Consequently, when the assimilated spurious correlations break down, the model's predictions inevitably fail, while the explanations are erroneous, as the FCM misconstrues the relationships between the problem variables. Similarly, relying on the correlation coefficient is equally unreliable because it has been shown that correlation does not necessarily imply causality (Rohrer 2018). In the domain of industrial anomaly detection, such an FCM proves ineffective, as its explanations misdirect plant supervisors to irrelevant parts of the manufacturing system, hindering the identification of the root causes of faults. Such gaps provide the motivation to develop new methods that identify authentic causal relationships between problem variables and rule out possible spurious correlations (Nápoles et al. 2020b). In this direction, Yosef et al. (2022) presented a method for removing spurious correlations by calculating the concepts' behavioral similarity through data and applying a set of defined rules from domain experts to discern the actual causal relationships. However, through this approach, an FCM can still contain spurious correlations that experts consider acceptable, while some actual causal associations can remain undetected as they can be beyond experts' knowledge. Finally, this expert-driven causality analysis is unfeasible for highly complex systems with many variables.

A potential solution to these limitations is to devise a method that identifies the real causal structure of an FCM from observational data, eliminating the need for domain

experts. By doing so, this method can offer a significant contribution in two ways: (a) preventing the injection of spuriousness into these cognitive networks, thereby enhancing their prediction accuracy and interpretability, and (b) providing a tool that can efficiently handle large-scale problems. Notably, a distinguishing contribution of this work is that, to the best of the authors' knowledge, no other data-driven causal discovery method has been suggested to rule out spurious correlations in FCMs, thereby elevating their performance and robustness.

This paper's main contribution is introducing a novel approach for FCM construction, leveraging the Liang-Kleeman Information Flow (L-K IF) analysis for causal inference. In more detail, unlike the approach presented by Yosef et al. (2022), the proposed technique contributes by eliminating the necessity for expert involvement; it identifies the authentic causal relationships from the data using an automatic causal search algorithm. A pivotal part of our contribution is the imposition of the derived causal links as constraints during the FCM learning procedure. This strategic move is tailored to effectively remove spurious correlations and, in doing so, improve the FCM's aggregate predictive and explanatory power. The capabilities of the proposed method are demonstrated in the context of developing an XAI model for anomaly detection and root cause analysis in an industrial system. Finally, a comparative analysis is conducted between the developed FCM and state-of-the-art FCM-based models in terms of their predictive and explanatory power. It is worth noting that while the presented case study focuses on anomaly detection, the proposed method can be effectively employed in other prediction problems as well. For further details and implementation, the code for this study is available in Tyrovolas et al. (2023).

The rest of the paper is organized as follows. Section 2 presents the foundations of the classic FCM formalism and L-K IF analysis. Section 3 describes in detail the proposed methodology, including the model's development process and how to predict and interpret its results. Section 4 conducts extensive numerical simulations to compare the proposed model against state-of-the-art FCM-based models. Finally, Sect. 5 presents some concluding remarks.

2 Theoretical background

This section first presents some basic notions of FCMs regarding their structure and how they perform the simulations. Second, it describes the causal inference tool L-K IF analysis, used to determine the actual causal relationships between the analyzed system variables.

2.1 Fuzzy cognitive maps

As mentioned in Sect. 1, an FCM consists of n concepts $C_i, i \in \{1, 2, \cdots, n\}$, and weights $w_{ij} \in [-1, 1]$ that indicate the causal influence from C_i to C_j. In general, there are three kinds of causality:

- **Positive causality** ($w_{ij} > 0$): the affected variable (C_j) changes (increases or decreases) in the same direction as its cause variable (C_i) changes.
- **Negative causality** ($w_{ij} < 0$): the affected variable (C_j) changes in the opposite direction to its cause variable (C_i) change.
- **Zero causality** ($w_{ij} = 0$): there is no relation between the cause (C_i) and the affected (C_j) variable.

Each concept C_i has an activation value A_i, which is determined via a reasoning rule, where the most common is

$$A_i^{(t+1)} = f\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} A_j^{(t)} w_{ji}\right), \tag{1}$$

where t is the iteration step, $A_i^{(t+1)}$ denotes the activation value of the i-th concept at $(t+1)$th iteration step, $A_j^{(t)}$ denotes the activation value of the j-th concept at tth iteration step, w_{ji} denotes the causal weight from jth concept to i-th concept, and $f(\cdot)$ denotes the activation function that normalizes the concepts' activation values within a specified interval (Kosko 1986). The most known activation functions are bivalent, trivalent, hyperbolic tangent, and sigmoid, where depending on which is selected, $A_i^{(t+1)}$ receives values within the [0, 1] or $[-1, 1]$ intervals (Orang et al. 2022). The activation values of all concepts in each iteration step can be expressed as a state vector $\mathbf{A} \in \mathbb{R}^n$, while the values of the causal weights w_{ij} between each pair of concepts C_i and C_j, compose a weight matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, whose diagonal elements are equal to zero. Therefore, (1) can be rewritten as:

$$\mathbf{A}^{(t)} = f(\mathbf{A}^{(t-1)}\mathbf{W}). \tag{2}$$

Using (2), the activation values of the concepts in each iteration step are computed. An initial state vector $\mathbf{A}^{(0)}$, which includes input data (e.g., sensor data), triggers the FCM's iterative reasoning process (Falcon et al. 2019). Subsequently, a new state vector yields at each iteration step until the termination condition is satisfied, which can be either the FCM's convergence to an equilibrium point, leading to reliable results, or the completion of a maximum number of iterations, where the FCM exhibits cyclic or chaotic behavior (Kosko 1988).

2.2 Information flow

As mentioned above, accurately identifying authentic causal relationships between variables in a modeled system is crucial for developing efficient FCMs. Causality inference from data is a notoriously difficult problem that has been extensively studied for over half a decade (Egrioglu et al. 2022). Various methods (mostly statistical) have been proposed to address this challenge, but these often suffer from certain limitations (Eichler 2013; Hlavackovaschindler et al. 2007; Runge et al. 2012). Some of these approaches are qualitative, lacking the necessary quantitative information for this research's purpose, while other methods are empirically or half-empirically formulated, which, although successful in specific contexts, lack the desired universality for developing all-purpose algorithms. Recently, it has been realized that causality is actually a real physical notion called Information Flow (IF) and can be rigorously derived from first principles (Liang 2014, 2016). Specifically, IF describes the contribution of one variable's entropy per unit of time in increasing the marginal entropy of another variable and reflects the magnitude, kind, and direction of their cause-effect relationship. This offers a promising way to systematically formulate causality analysis in a quantitative sense based on a rigorous theoretical framework, enabling its universal applicability across different disciplines. The fundamental equations for calculating the IF between two or more system variables are as follows.

Let be a two-dimensional (2-D) dynamic system:

$$d\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, t)dt + \boldsymbol{B}(\boldsymbol{x}, t)d\boldsymbol{w}, \tag{3}$$

where $\boldsymbol{F} = (F_1, F_2)$ is the deterministic components, $x = (x_1, x_2) \in \mathbb{R}^2$ is the state variables, $\boldsymbol{w} = (w_1, w_2)$ is a standard 2-D Wiener process, and $\boldsymbol{B} = (b_{ij})$ is the matrix of perturbation amplitude (Liang 2008). For the aforementioned system, the IF from x_2 to x_1 is

$$T_{2\to 1} = -E\left(\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right) + \frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2 \mathrm{g}_{11}\rho_1}{\partial x_1^2}\right), \tag{4}$$

where $\rho(t; x_1, x_2)$ is the joint probability density function, $\rho_1(t; x_1) = \int_{\mathbb{R}} \rho dx_2$ is the marginal density of x_1, $\mathrm{g}_{11} = \sum_{\mathrm{k}=1}^{2} \mathrm{b}_{1\mathrm{k}}^2$, and E is the expectation with respect to ρ. An important property of (4) is the satisfaction of the *nil causality* principle, according to which x_2 is not causal to x_1 ($T_{2\to 1} = 0$) if the evolution of the latter is independent of the former (neither F_1 nor g_{11} depends on x_2) (Liang 2016).

As a further step, Liang (2014) established that under a linearity assumption, the IF of two system variables can be estimated from only two time series, say, X_1 and X_2, using the following maximum-likelihood estimator of (4):

$$T_{2\to 1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^2 C_{1,d1}}{C_{11}^2 C_{22} - C_{11}C_{12}^2}, \tag{5}$$

where C_{ij} is the sample covariance between $X_i$and X_j, and $C_{i,dj} = \overline{(X_i - \overline{X_i})(\dot{X}_j - \overline{\dot{X}_j})}$ is the sample covariance between X_i and the difference approximation of $\frac{\mathrm{d}X_j}{\mathrm{d}t}$, which is computed using the Euler forward scheme: $\dot{X}_{j,n} = \left(X_{j,n+k} - X_{j,n}\right)/(k\Delta t)$, with $k \geq 1$ some integer. The IF in the opposite direction, i.e., $T_{1\to 2}$, is obtained by swapping indices 1 and 2. Besides, writing (5) as a function of correlation and/or correlation-like quantities gives

$$T_{2\to 1} = \frac{r}{1 - r^2}(\acute{r}_{2,d1} - r\,\acute{r}_{1,d1}), \tag{6}$$

where $r = C_{12}/\sqrt{C_{11}C_{22}}$ is the sample correlation coefficient between X_1 and X_2, and $\acute{r}_{i,dj} = C_{i,dj}/\sqrt{C_{ii}C_{jj}}\,(i, j = 1, 2)$ is the "correlation" between X_i and $\dot{X}_j$ but normalized with the variances of X_i and X_j. According to (6), when two variables are causally related ($T_{2\to 1} \neq 0$), they are correlated ($r \neq 0$). However, the opposite does not hold. This property helps distinguish authentic causal relationships from spurious correlations.

Recently, (5) was generalized, resulting in a simple formula for causality analysis among multiple variables (Liang 2021). In detail, given a dataset of d time-series variables, the IF from X_2 to X_1 is

$$\hat{T}_{2\to 1} = \frac{1}{detC} \cdot \sum_{j=1}^{d} \Delta_{2j} C_{j,d1} \cdot \frac{C_{12}}{C_{11}}, \tag{7}$$

where $C_{j,d1}$ is the sample covariance between X_j and $\dot{X}_1$, and Δ_{ij} are the cofactors of the covariance matrix C. An algorithm (Algorithm 1) for multivariate time-series causality analysis is developed based on (7). As observed from the algorithm, a statistical significance test is conducted to draw safe conclusions about the actual causal relationships for each pair of variables, estimated by $\hat{T}_{i\to j}$.

Algorithm 1: Quantitative causal inference

Input: Dataset of d time series
Output: a causal graph $\mathcal{G} = (V, E)$, where V and E are the set of vertexes and edges, and IFs along edges
initialize $\mathcal{G}$ such that all vertexes are isolated;
set a significance level α
for *each* $(i, j) \in V \times V$ **do**
  compute $\hat{T}_{i\to j}$ by (7);
  if $\hat{T}_{i\to j}$ *is significant at level* α **then**
    add $i \to j$ to $\mathcal{G}$;
    record $\hat{T}_{i\to j}$;
return $\mathcal{G}$, together with the IFs $\hat{T}_{i\to j}$

Nevertheless, the importance of the causal relationship must be assessed more than by inspecting the presence of causality between variables. For this purpose, the normalization of the estimated significant IF rates has been proposed with the normalizer of $\hat{T}_{2\to1}$ being

$$\hat{Z} = \left|\left(\widehat{\frac{dH_1^*}{dt}}\right)\right| + \sum_{j=2}^{d} |\hat{T}_{j\to1}| + \left|\left(\widehat{\frac{dH_1^{noise}}{dt}}\right)\right|, \tag{8}$$

where

$$\left(\widehat{\frac{dH_1^*}{dt}}\right) = \frac{1}{detC} \cdot \sum_{j=1}^{d} \Delta_{1j} C_{j,d1}, \tag{9}$$

$$\left(\widehat{\frac{dH_1^{noise}}{dt}}\right) = \frac{1}{2}\frac{\hat{g}_{11}}{C_{11}}, \tag{10}$$

and $\hat{g}_{11} = \frac{Q_{N,1}\Delta t}{N}$. Finally, the normalized IF from X_2 to X_1 is:

$$\tau_{2\to1} = \frac{T_{2\to1}}{\hat{Z}} \tag{11}$$

which lies on $[-1, 1]$. When $|\tau_{2\to1}|$ is 1, X_2 has the greatest causal impact on X_1. Furthermore, simply swapping the indices in the above equations yields $\tau_{1\to2}$.

2.3 L-K IF analysis on binary time series

Previous studies utilizing L-K IF analysis to identify causal relations have not focused on discrete-valued signals that take a few values, such as a binary time series. However, real-world datasets, particularly in industrial settings, often comprise binary variables such as the state of a proximity sensor or button. Consequently, an experiment was conducted to verify the efficiency of the causal inference tool in effectively handling binary data.

Let be three series X_1, X_2, and X_3, generated from three autoregressive processes:

$$X_1(n+1) = 0.1 + 0.4X_1(n) - 0.8X_3(n) + e_1(n+1), \tag{12a}$$

$$X_2(n+1) = 0.7 + 0.7X_3(n) - 0.8X_2(n) + e_2(n+1), \tag{12b}$$

$$X_3(n+1) = 0.5 + 0.5X_3(n) + e_3(n+1), \tag{12c}$$

where X_3 is the confounder of the other two ($X_3 \to X_1$ and $X_3 \to X_2$) without any other causality, and the errors, $e_1 \sim N(0,1)$, $e_2 \sim N(0,1)$ and $e_3 \sim N(0,1)$ are independent. After initializing the variables with random values and generating 10,000 samples for each, L-K IF analysis was performed. Table 1a depicts the derived IF rates and their respective confidence intervals at the 99% confidence level. The results demonstrate that the only significant IF rates are $T_{3\to1}$ and $T_{3\to2}$ as they lie within the intervals [0.1975, 0.2091] and [0.0613, 0.0657], respectively, which is in agreement with the actual relations. The rest of Ts take both negative and positive values; thus, they cannot be distinguished from zero. It is noteworthy that creating pseudorandom values can lead to slightly different results for different series. Nevertheless, the mean is expected to converge to the same value when an ensemble of series is examined. Subsequently, the experiment was repeated using the binarized time series, that is, the series discretized into 0 or 1. After repeating the L-K IF analysis (Table 1b), it is concluded that the proposed technique reliably captures the causal relations in a qualitative sense, even if the time series have been binarized.

3 Proposed methodology

Figure 1 illustrates the proposed methodology, outlining the major phases of constructing an FCM-based model and interpreting its predictions.

Table 1 IF rates for the series generated with (12) and their respective confidence intervals (99% confidence level)

		To		
	variables	X_1	X_2	X_3
From	X_1	\	0.0018 ± 0.0027	-0.0023 ± 0.0085
	X_2	-0.0013 ± 0.0029	\	0.0008 ± 0.0034
	X_3	0.2033 ± 0.0058	0.0635 ± 0.0022	\
		$T_{3\to1}$	$T_{3\to2}$	

(a) Raw Time Series

		To		
	variables	X_1	X_2	X_3
From	X_1	\	0.0011 ± 0.0025	0.0039 ± 0.0054
	X_2	0.0019 ± 0.0023	\	0.0013 ± 0.0022
	X_3	0.0918 ± 0.0046	0.0187 ± 0.0020	\
		$T_{3\to1}$	$T_{3\to2}$	

(b) Binarized Time Series

3.1 Data pre-processing

Once data are collected from the target system, such as a manufacturing system, suitable data pre-processing techniques are employed. Initially, since FCM can only handle numeric data, categorical variables, including class attributes in a classification problem, need to be encoded. The numerical representative ($a_j \in [0,1]$) for each class label ($class_j$) is calculated using the following formula:

$$a_j = \frac{j-1}{m-1}, \tag{13}$$

where $j \in \{1,\ldots,m\}$ and $m \geq 2$ the number of class labels.

In the context of FCMs, an essential step is the assignment of fuzzy values to concepts, known as data fuzzification. Fuzzification is practically considered a data normalization procedure that computes the concepts' initial activation values for each data observation. Traditional normalization techniques include min-max and z-score normalization; however, they present some weak points, such as out-of-bounds error when a new value is outlying and susceptibility to outliers. Furthermore, min-max normalization yields different normalizations for different data separations, such as in cross-validation. To address these issues, the Generalized Logistic (GL) algorithm was utilized in this study for data normalization (Cao et al. 2016). This algorithm makes no assumptions about the distribution of variables but instead uses a generalized logistic function to approximate the cumulative distribution function (CDF) of each variable. The main advantage of this method is its robustness against outliers. The algorithm maps values from interval $(-\infty, \infty)$ to the interval [0, 1].

3.2 Information flow-based fuzzy cognitive map (IF-FCM)

After preparing the data, the next step is to define the FCM architecture that determines the type and number of concepts. In the context of classification, the literature presents two main FCM architectures, which differ in the number of output concepts (OCs), but also in how they assign a class label for each data instance. The first architecture, known as the *class-per-output architecture* (CpO), maps each class label to a separate OC with m total outputs. The predicted class is then indicated by the OC with the highest activation value in the last iteration of the reasoning process. In contrast, in the second architecture, referred to as the *single-output architecture* (SO), the class attribute is mapped to a single OC C_n. The estimated activation value of C_n is then assigned to one of the class labels by dividing the activation interval ([0, 1] or $[-1, 1]$) into partitions, each corresponding to a class label (Papakostas et al. 2008). The classification process in an FCM-SO can be summarized as follows:

Step 1: Consider the kth data observation in the dataset as the initial state vector

$$\mathbf{A}_k^{(0)} = [A_{1k}^{(0)}, A_{2k}^{(0)}, \ldots, A_{nk}^{(0)} = 0], \tag{14}$$

where $A_{ik}^{(0)} \in [0,1]$, $i \in \{1,2,\ldots,n-1\}$ are the initial activation values of the input concepts, and $A_{nk}^{(0)}$ the initial activation value of the OC

Step 2: Applying the employed reasoning rule recurrently, calculate the state vector

$$\mathbf{A}_k^{(l)} = [A_{1k}^{(l)}, A_{2k}^{(l)}, \ldots, A_{nk}^{(l)}], \tag{15}$$

in the steady state l, whereas $|A_{ik}^{(l)} - A_{ik}^{(l-1)}| < \varepsilon$, with ε being a small positive number (usually 10^{-5}), and

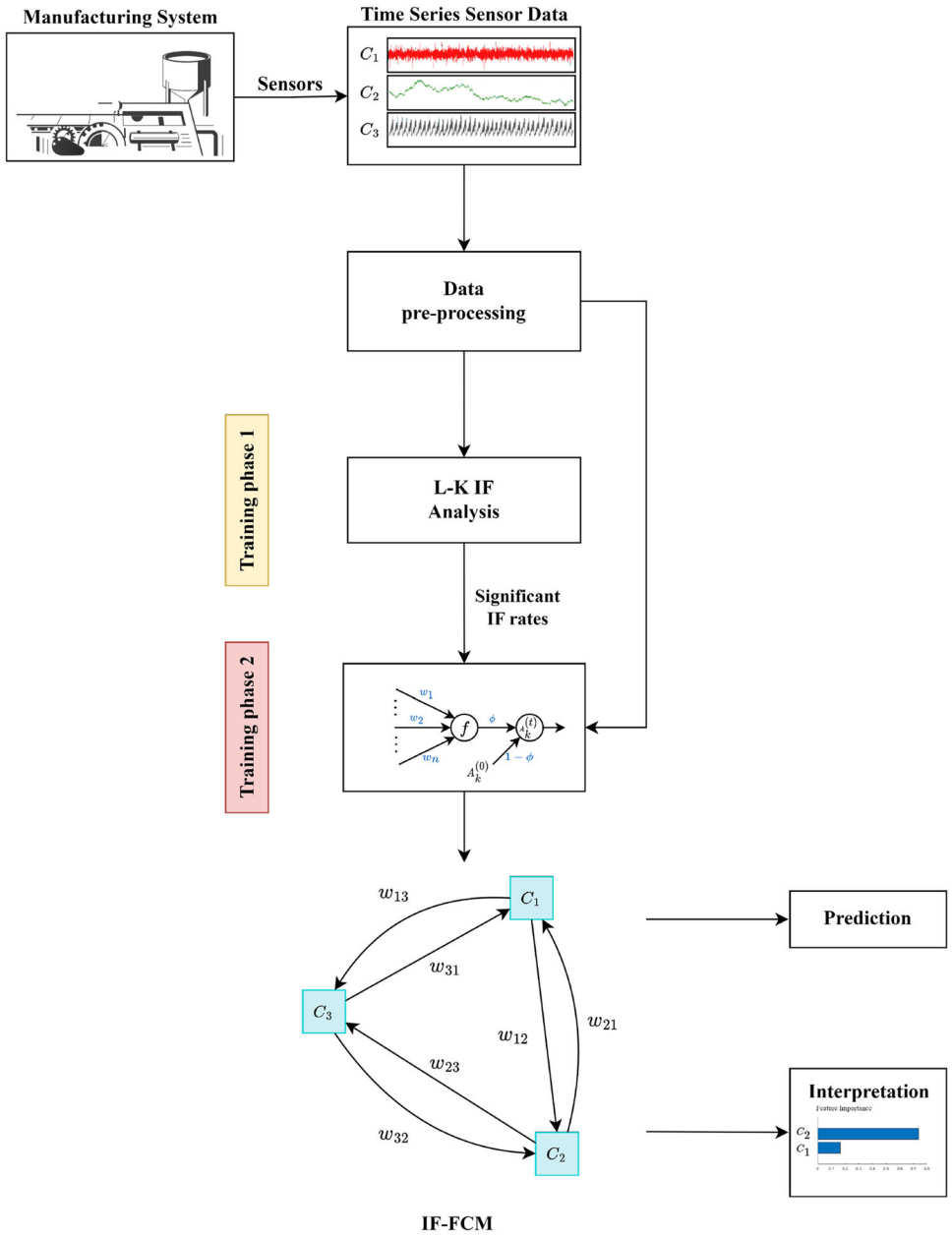

Fig. 1 Proposed methodology scheme

$i \in \{1, 2, \ldots, n\}$. The maximum number of iterations is denoted by T and defined by the user. $A_{nk}^{(l)}$ is the activation value of the OC in the last iteration.

Step 3: Once the reasoning process is complete, assign $A_{nk}^{(l)}$ to one of the numerical representatives of the class labels. This is accomplished using $m-1$ defined decision thresholds that divide the activation interval into m partitions. Therefore, depending on the range $A_{nk}^{(l)}$ falls into, the FCM predicts the corresponding class label. To determine the decision thresholds, a "threshold-moving" approach is employed, which identifies the optimal value based on a predefined evaluation metric. In this paper, we locate the decision threshold by considering the maximum value of the Geometric Mean (16), which describes the balance of classification performance on both majority and minority classes, allowing us to determine the ideal position of the classification hyperplane (Kubat et al. 1997).

$$G - mean = \sqrt{TPR * TNR} \tag{16}$$

In this study, the second architecture was selected because of its lower parameter count and computational requirements. Additionally, a comprehensive analysis of the architectures conducted in prior research, the study by Papakostas et al. (2012), concluded that the SO

architecture outperformed the CpO architecture on seven of the eight datasets analyzed.

3.3 IF-FCM learning

After determining the architecture, a learning procedure is performed to adapt the FCM behavior based on the collected data (Fig. 1). The proposed approach is divided into two phases. In the first phase (*Training phase 1*), Algorithm 1 is executed to determine the causal relationships between the dataset variables. The algorithm is computationally efficient, even when the scales of the original variables differ greatly; therefore, raw encoded data are used.

In the second phase (*Training phase 2*), the parameters that define the FCM response are tuned, including the weights and parameters associated with the activation function and reasoning rule. Consequently, a challenging question arises regarding the choice of the appropriate reasoning rule and activation function. Previous studies have shown that using (1) in conjunction with the activation functions mentioned in Sect. 2.1 often leads the FCM to converge to the same equilibrium point regardless of the initial state vector (Boutalis et al. 2009). However, this behavior is undesirable in forecasting tasks, such as anomaly detection, as the model predicts only one class label. Furthermore, the use of bounded activation functions can lead to saturation issues, where the activation values of concepts tend to approach the lower or upper boundary of the specified interval when they receive a strong negative or positive influence, respectively (Nápoles et al. 2022a). Finally, the sigmoid function deceives the simulation results by activating unexpected concepts based on their received influence, as it returns 0.5 when its argument is zero (Mpelogianni and Groumpos 2018).

Recently, to solve the issues mentioned above, a new rule called *quasi nonlinear reasoning rule* was proposed, which involves a re-scaled activation function acting as a normalizer (Nápoles et al. 2022b), and is mathematically expressed as

$$A_i^{(t+1)} = \underbrace{\varphi f\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j^{(t)} w_{ji}\right)}_{\text{nonlinear component}} + \underbrace{(1-\varphi)A_i^{(0)}}_{\text{linear component}}, \tag{17}$$

where the parameter $\varphi \in [0,1]$ controls the nonlinearity of the reasoning rule, and $f(\cdot): \mathbb{R}^n \to \mathbb{R}^n$ is the activation function defined as

$$f(\mathbf{X}) = \begin{cases} \dfrac{\mathbf{X}}{\|\mathbf{X}\|_2}, & \mathbf{X} \neq \overrightarrow{0} \\ 0, & otherwise \end{cases} \tag{18}$$

such that $\|\cdot\|_2$ denotes the Euclidean norm. Using a matrix-like notation, (17) is rewritten as

$$\mathbf{A}^{(t)} = \varphi f\left(\mathbf{A}^{(t-1)}\mathbf{W}\right) + (1-\varphi)\mathbf{A}^{(0)}. \tag{19}$$

In the study conducted by Nápoles et al. (2022a), the convergence properties of the above reasoning mechanism were thoroughly examined. Through a mathematical proof by contradiction, it was concluded that an FCM employing (18) and (19) does not have a unique equilibrium point for all initial state vectors when $\varphi \in [0,1)$. They also explored the case of $\varphi = 1$ based on the symmetry and diagonalizability of the derived $\mathbf{W}$. In more detail, using the appropriate matrix properties, the authors equated the reasoning rule with the *power iteration method* formula and concluded that for a diagonalizable weight matrix $\mathbf{W}$ with eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, if an initial stimulus u_0 has a nonzero projection along an eigenvector associated with λ_1, then u_k converges to such an eigenvector as $k \to \infty$ (Mises and Pollaczek-Geiringer 1929). In particular, when λ_1 is real, the method converges to a unique fixed point. Nevertheless, because asymmetry is a distinguishing characteristic of causation, the convergence of the FCM for $\varphi = 1$ should be analyzed without relying on the diagonalizability of $\mathbf{W}$. In the context of the power iteration method, studies have demonstrated that even if $\mathbf{W}$ is not diagonalizable, the same outcomes are achieved, albeit with slower convergence (Leader 1991). Therefore, the case of $\varphi = 1$ enables modelling scenarios in which the FCM converges to a unique fixed-point attractor without the need for symmetry in $\mathbf{W}$.

In this paper, we utilize the reasoning rule presented in (19) and the activation function of (18). The learning algorithm eventually adjusts the weights of the FCM and the controllable parameter φ. However, the difference between the proposed method and the existing methods is that the normalized significant IF rates computed in *Training phase 1* are imposed as constraints in *Training phase 2* to avoid capturing spurious correlations. In detail, only the weights of edges with significant estimated IF rates are tunable parameters, while the remaining weights are set to zero "a priori". This approach improves the generalizability and interpretability of the developed FCM while reducing the training time by reducing the dimensions of the optimization problem. Therefore, a candidate solution is encoded as a (SIFs + 1)-dimensional vector, where SIFs is the number of significant IF rates and parameter φ.

$$x = [\varphi, w^{(1)}, w^{(2)}, \ldots, w^{(\mathrm{SIFs})}]. \tag{20}$$

For FCM learning, which involves determining the optimal weight values and φ, we have chosen the Particle Swarm Optimization (PSO) metaheuristic algorithm due to its effectiveness in the literature (Papageorgiou et al. 2005; Bas et al. 2022). PSO starts with a random population of candidate solutions called particles. Through iterations, the particles are evaluated using a defined cost function and updated accordingly. The process continues until a satisfactory solution is found or a stopping criterion is met, such as the maximum number of function evaluations. The cost function used in this study is defined as follows:

$$\mathcal{E}(x) = \alpha_1 G(x) + \alpha_2 H(x), \tag{21}$$

where x represents a candidate solution, $0 \leq G(\cdot) \leq 1$ denotes the FCM's mean absolute prediction error (22), and $0 \leq H(\cdot) \leq 1$ denotes the accumulated dissimilarity between two consecutive FCM state vectors (23). The parameters $\alpha_1, \alpha_2 \in [0, 1]$ indicate the relevance of the FCM's prediction accuracy versus stability, for which $\alpha_1 + \alpha_2 = 1$, ensuring that the cost function is always bounded in the interval [0, 1].

$$G(x) = \frac{1}{K}\sum_{k=1}^{K}\left|Y_k - A_{n,k}^{(l)}\right| \tag{22}$$

$$H(x) = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{l}\frac{2\,\omega_t(A_{ik}^{(t)} - A_{ik}^{(t-1)})^2}{K\,n\,(T-1)} \tag{23}$$

In (22) and (23), K represents the number of training observations, n is the number of FCM concepts, Y_k is the expected value of the output concept in k-th data observation, and $\omega_t = \frac{t}{T}$ is the importance of the t-th iteration in the reasoning process, which increases linearly with the number of iterations. The rationale behind ω_t is that the learning algorithm should focus primarily on stabilizing the last iterations, allowing greater flexibility at the beginning (Nápoles et al. 2016).

3.4 Interpretation of FCM's predictions

The proposed FCM can explain its predictions, supporting two levels of interpretability: (a) *global* and (b) *local*. At the global level, IF-FCM provides a holistic view of the influence of each input variable in the decision-making process, whereas, at the local level, it provides numeric explanations for individual predictions by calculating the importance of each input feature to this particular decision. The ability to find the features that play a critical role in classifying a sample as an anomaly enables root cause analysis (Brito et al. 2022).

3.4.1 Global interpretability

The relevant literature demonstrates several methods for examining the overall contribution of each feature to the decision-making process of an FCM. The most widespread method is based on graph theory and states that the concept's importance can be measured via its degree of centrality (Kosko 1986):

$$CEN(C_i) = in(C_i) + out(C_i), \tag{24}$$

where $in(C_i)$ and $out(C_i)$ refer to the number of incoming and outcoming edges of each concept C_i, respectively. The most significant feature of the FCM is the one whose sum of the concepts acting on it and those affected by it is the largest.

3.4.2 Local interpretability

To explain the decision for a given data instance, FCMs provide a dynamic, semi-quantitative method that analyzes the propagation of effects from one concept to another using a plot of the activation values of all concepts across iterations (Barbrook-Johnson and Penn 2022). The final activation values of the input concepts after FCM stabilization reflect their contribution to the prediction, with concepts with larger absolute values interpreted as more important or influenced/influential (Soler et al. 2012; Liu et al. 2020). Furthermore, this plot enables the investigation of how relative changes in the initial concept values impact the reasoning process, providing insights into whether changes accelerate, stabilize, or diminish over time.

4 Experimental results

In this section, we present the results of numerical simulations designed to assess the efficacy of the proposed methodology. First, we provide a detailed description of the dataset used in the simulation. Next, we outline the application of the proposed methodology to the dataset. Finally, we compared our model with state-of-the-art FCM-based models in terms of their prediction accuracy, interpretability, and aggregate power.

4.1 Dataset description

One of the standard datasets used in industrial process anomaly detection is Matzka's PMAI4I dataset, which we adopted in our experimental evaluation (Matzka 2020). This synthetic yet realistic dataset represents industrial predictive maintenance data, and has been widely recognized and accepted as a reliable benchmark for evaluating

various XAI methods (Ghasemkhani et al. 2023; Mylonas et al. 2023). It contains 10,000 samples covering a diverse range of variables to provide a holistic view of the industrial data. These variables include one categorical variable (product quality ∈ {"low", "medium", "high"}), five numerical variables (air temperature, process temperature, rotational speed, torque and tool wear) and a binary target variable indicating the machine failure ("0" = Healthy, "1" =Faulty. For each sample, apart from the fault, its type is known to be one of the following:

1. **Tool wear failure (TWF)**: the tool fails at a random tool wear time between 200 and 240 min.
2. **Heat dissipation failure (HDF)**: if the difference between the air and the process temperature is less than 8.6 K while the tool's rotational speed is less than 1380 rpm, a failure is caused.
3. **Power failure (PWF)**: if the required power (i.e., the product of torque and rotational speed in rad/s) is less than 3500 W or greater than 9000 W, the system fails.
4. **Overstrain failure (OSF)**: the process fails by overstrain when the product of tool wear and torque exceeds 11.000 minNm for low quality (L) products, 12.000 for medium quality (M), and 13.000 for high quality (H), respectively.
5. **Random failures (RNF)**: regardless of process parameter values, there is a 0.1% probability of failure.

If any of the mentioned failure modes is present, the process fails, and the machine failure value is set to one. However, during training, the FCM receives only the input variable values and system condition information without knowing the root cause of the fault. Hence, the FCM-based classifier aims to achieve two objectives: first, detecting the presence of anomalies in the analyzed manufacturing system, and second, identifying the most significant input variable(s) for each true positive prediction, which are likely to be responsible for the fault. This is accomplished by leveraging the inherent interpretability characteristics of the FCM.

Table 2 Significant IF Rates In The PMAI4I

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0	0	0	0	0	**−5.59e−4**	0
X_2	0	0	**0.3512**	0	0	**2.1e−3**	**1.2e−2**
X_3	0	0	0	0	0	0	**−3.8e−3**
X_4	0	0	0	0	0	**7.55e−5**	0
X_5	0	0	0	0	0	**−1.5e−3**	0
X_6	0	0	0	0	0	0	**0.7e−2**
X_7	0	0	0	0	0	**−8.13e−2**	0

4.2 Simulations execution

Following the methodology described in Sect. 3, the data are first pre-processed. This includes encoding categorical features, such as product quality, where each category value is assigned an integer (e.g., "low" is represented as 0, "medium" as 1, and "high" as 2). In addition, because the dataset is imbalanced, an SMOTE-based algorithm is used to address this issue, generating artificial instances of the minority class "1" (Sridhar and Sanagavarapu 2021). In particular, this study applies the hybrid algorithm SMOTE-ENN, which merges undersampling and oversampling using Edited Nearest Neighbors and SMOTE, respectively. This combination strengthens the bias towards the minority class while weakening it towards the majority class, resulting in improved overall performance compared to using these techniques individually. Finally, data fuzzification is performed to prepare the data for training and decision-making.

4.2.1 Training phase 1

By applying Algorithm 1 to the dataset and subsequently normalizing the significant IF rates, we obtain the results presented in Table 2. These findings align with the dataset description since:

1. The product quality influences the wear time of the tool, leading to the occurrence of TWF and OSF.
2. The process temperature at each time step is derived from the air temperature samples, indicating a causal relationship between these variables.
3. Both air temperature and process temperature contribute to the emergence of HDF in the system, establishing an information flow from these variables to the target variable.

Nonetheless, beyond the obvious links, the algorithm discovered that:

1. Rotational speed and torque do not directly affect machine failure but indirectly through tool wear.
2. Air temperature has a causal influence on tool wear.
3. There exists a feedback loop from machine failure to tool wear.

4.2.2 Training phase 2

As mentioned previously, during *Training phase 2*, the weights of all FCM edges with insignificant IF rates were set to zero before starting PSO execution. According to Table 2, the final number of tunable weights is nine along with the reasoning rule parameter φ. For the PSO

parameter initialization, a population size of 100 was chosen, and the cost function parameters α_1 and α_2 were set to 0.8 and 0.2, respectively. Additionally, to achieve a more accurate solution, a hybrid function was employed to continue the optimization after the termination of the original solver. The algorithm was implemented using MATLAB global optimization toolbox.

4.2.3 Experimental setup

After training the IF-FCM, we compared its predictive and explanatory power with several state-of-the-art FCM-based models. These include FCM-A (Froelich 2017), FCMBinaryClassifier (FCMB) (Szwed 2021), FCMMultiClassClassifier (FCMMC) (Szwed 2021), Long-Term Cognitive Network (LTCN) (Nápoles et al. 2022b), and a Fuzzy Cognitive Map using the "*Stability based on Sigmoid Functions*" method (FCM-SSF) (Nápoles et al. 2017). Furthermore, to emphasize the significance of L-K IF analysis in enhancing FCM performance, we developed two additional models that utilize (18) and (19) for decision-making while being trained through PSO. However, their *Training phase 1* varies. In the first model, called correlation coefficient-based FCM (CCFCM), the weights correspond to the correlation coefficients between variables with a p value less than 0.05. In the second model (FCM-FC), all weights are included without performing initial data analysis to determine the relationships between concepts.

To avoid possible issues such as overfitting and ensure the generalizability of the models, stratified 10-fold cross-validation was used for the simulations. Simultaneously, hyper-parameter tuning was conducted to optimize the performance of each model, considering the variables displayed in Table 3. As for FCM-SSF, 91 maps were randomly generated, varying in network densities from 10% to 100%, and the model delivering the best performance was selected.

Table 3 List of Hyper-parameters for model tuning

Model	Hyper-parameters
FCM-A[d]	g = 0 to 10
FCMB[b]	Activation: sigmoid
	Activation_m: 1
	Depth: 2, 3, 5
	Epochs: 50, 100
	Batch size: 16, 256, 4096, –1
	Buffer_size: 1000
	Training_loss:logloss
	Optimizer: rmsprop
	Learning_rate: 0.001, 0.01, 0.05, 0.1, 0.5
FCMMC[b]	Activation: sigmoid
	Activation_m: 1
	Depth: 2, 3, 5
	Epochs: 50, 100
	Batch_size: 16, 256, 4096, −1
	Buffer_size: 1000
	Training_loss: softmax
	Optimizer: rmsprop
	Learning_rate: 0.001, 0.01, 0.05, 0.1, 0.5
LTCN[a]	Method: inverse
	Transfer function: sigmoid, tanh
	Phi: 0.5 to 1.0
	T: 5, 10, 15
	Alpha: 0, 0.01, 100
FCM-SSF[c]	Density: 10% to 100%
	Slope parameter of the sigmoid: –1 to 10
	Offset parameter of the sigmoid: –1 to 1

[a] Nápoles et al. (2022b)
[b] Szwed (2021)
[c] Nápoles et al. (2017)
[d] Froelich (2017)

Table 4 Average Accuracy, AUC, and Kappa coefficient For each FCM-based model

Model	Accuracy	AUC	Kappa
LTCN[a]	**0.95168**	**0.95061**	**0.90295**
FCMB[b]	0.94159	0.94357	0.88412
FCMMC[b]	0.93464	0.93432	0.86426
FCM-FC	0.85080	0.88231	0.70238
IF-FCM[e]	0.82235	0.85465	0.64600
FCM-SSF[c]	0.81956	N/A	0.63907
CCFCM	0.81029	0.81434	0.62009
FCM-A[d]	0.68311	0.86089	0.35364

[a] Nápoles et al. (2022b)
[b] Szwed (2021)
[c] Nápoles et al. (2017)
[d] Froelich (2017)
[e] Proposed model

4.2.4 IF-FCM's predictive power

Table 4 displays the average prediction accuracy, the "Area under the ROC Curve" (AUC) score, and the Cohen's kappa coefficient for each model across all folds. According to the results, the LTCN, FCMB, and FCMMC exhibit the highest performance among the cognitive networks, followed by FCM-FC and IF-FCM. In contrast, FCM-A demonstrates the lowest performance. The poor

performance of FCM-A can be attributed to the algorithm proposed by Froelich (2017), where the loop for computing the classification error of each candidate threshold focuses only on minimizing false negatives, rather than achieving an optimal balance between false negatives and false positives. This issue also explains the discrepancy between its accuracy and AUC score. Since AUC is threshold-invariant and provides an aggregate performance measure across all possible decision thresholds, a combination of low accuracy and high AUC suggests that the selected decision threshold is not optimal.

The FCM-SSF model is based on the CpO architecture. As mentioned in Sect. 3.2, this architecture selects the OC with the highest activation value in the final iteration to make its decision. This mechanism inherently lacks a decision threshold, which is a pivotal component in computing the AUC score. Specifically, the AUC score is a performance metric that evaluates the ability of a model to discriminate between the positive and negative classes. The ROC Curve is constructed by plotting the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at various decision threshold levels, typically ranging from 0 to 1. Different points on the ROC Curve are obtained by varying this threshold, and the AUC score is the area under this curve. The essence of the AUC score lies in its ability to assess the model's performance across all possible thresholds. Therefore, given that the CpO architecture of the FCM-SSF does not utilize a decision threshold, it becomes inherently incompatible with the ROC Curve. Without the ability to vary the decision threshold, generating the ROC Curve and computing the AUC score by extension is impossible. Therefore, in Table 4, we employed the symbol "N/A" (not applicable, not available) to indicate that the AUC score is not applicable or computable for the FCM-SSF model.

4.2.5 IF-FCM's explanatory power

Regarding interpretability, IF-FCM calculates the *global feature importance* using (24). Based on the causal structure presented in Table 2, tool wear is the most important feature with six incoming and outgoing edges, air and process temperature follow with three and two edges, respectively, while the rest have only one outgoing edge. To validate this finding, the LOFO (Leave-One-Feature-Out) method was employed, which is an XAI technique that iteratively removes each feature, retrains the model, and compares the resulting model error to a baseline model consisting of all features. This analysis assesses the mean feature importance value and standard deviation (Erdem 2023). LOFO was chosen due to its ability to handle correlated features, unlike linear models, and its robust generalization as it calculates feature importance across cross-validation splits. LOFO analysis was conducted for various black box machine learning (ML) models, such as Light Gradient-Boosting Machine (LightGBM), K-Nearest Neighbour (KNN), Decision Tree (DT), Multilayer Perceptron (MLP) classifier, Gaussian Naïve Bayes (NB), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB). In addition, an intrinsic interpretable Logistic Regression (LR) model was developed, where the feature coefficients provided insights into feature importance. The results presented in Fig. 2 indicate that tool wear was identified as the most impactful feature by seven out of the nine ML models, six models ranked torque as the second most influential feature, and so on. However, variations were observed among the models. For instance, XGB considered process temperature as the most significant feature and tool wear as the second, while LR highlighted torque as the most impactful, followed by air temperature and tool wear. These discrepancies underscore the potential differences in global interpretability across models, aligning with findings from previous studies (Li et al. 2022).

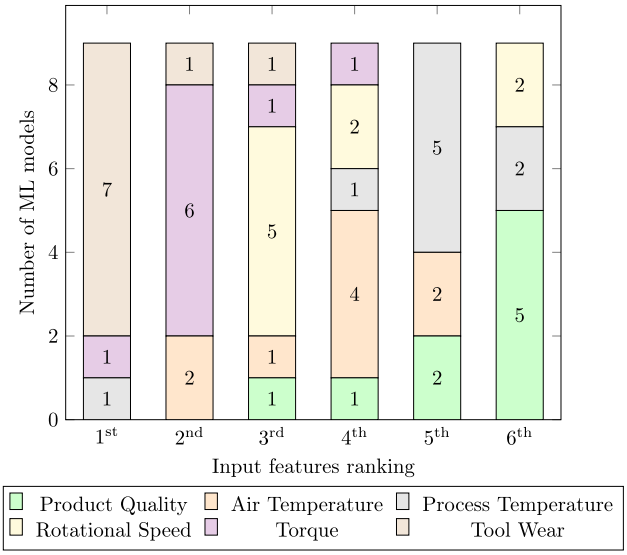

Fig. 2 The input features ranking based on their global importance for the examined ML models

Regarding the examined FCM-based models, the rankings of the input features' importance varied. In the LTCN model, the importance rankings of the input features were as follows: (1) torque, (2) tool wear, (3) rotational speed, (4) product quality, (5) process temperature, and (6) air temperature. Similarly, in FCM-SSF, product quality and tool wear were considered the most important features, each with three edges. Torque followed with two edges, while all other variables had only one edge. In contrast, the CCFCM assigned the highest importance to air temperature and rotational speed, with six edges each. Torque and

process temperature had four edges each, while product quality and tool wear had only two edges. However, the FCM-A model, which is based on SO architecture, does not represent the actual causal relationships and cannot calculate the degree of centrality for each concept. In this model, input concepts are connected only to the OC without feedback (Froelich 2017, Fig. 1). A similar issue arises in FCMB, FCMMC, and FCM-FC models, where the fully connected map structure suggests the presence of spurious correlations and does not allow for the calculation of each concept's centrality, as all concepts have the same number of edges. Determining the *global feature importance* of the PMAI4I dataset has also been a concern for other researchers. In particular, in the study conducted by Sridhar and Sanagavarapu (2021), the authors reached the conclusion that tool wear had the greatest effect, followed by torque and rotational speed, thereby providing additional assurance for the accuracy of the results.

The assessment of local interpretability was conducted by computing the success rate of the local explanations provided by IF-FCM for each failure mode. This rate indicates the percentage of correctly predicted anomalous data instances specific to each failure mode, in which the model effectively highlights the appropriate input features as the most important. This success rate serves as a measure of the model's accuracy, consistency, and coherence in attributing the correct input features to the detected fault. A higher success rate suggests an increased number of data observations, where the model precisely identifies the actual causal parameters of the failure, thus providing a measure of the correctness of the generated explanations. For instance, as shown in Fig. 3, the model should identify tool wear as the most significant input variable for a detected TWF. Conversely, in the event of a PWF, the model should underscore either the torque or the rotational speed.

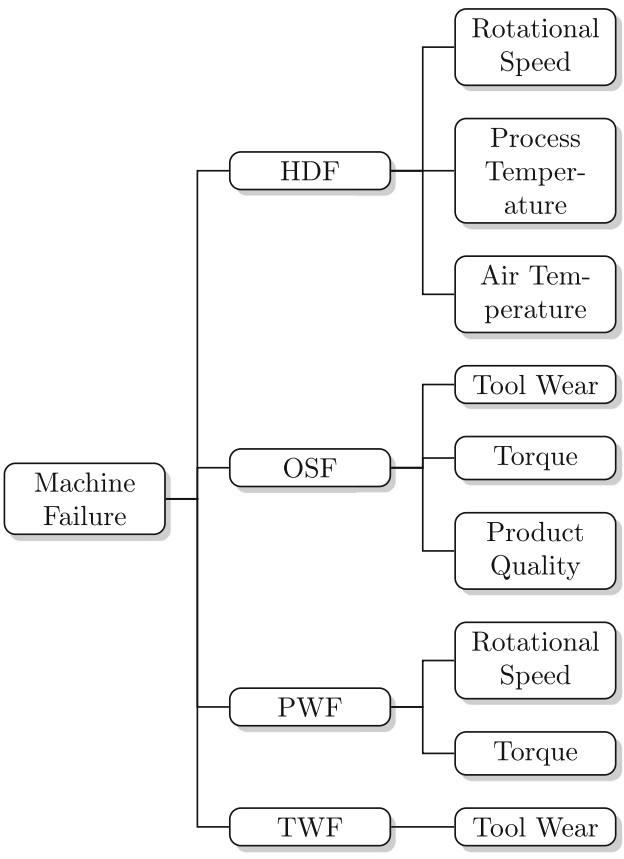

Fig. 3 Important input features for each failure mode

Table 5 presents both the average success rate for each failure mode and the overall average success rate across all four types of faults. As can be observed, the proposed model possesses the highest degree of interpretability with a success rate of 87.49%, outperforming all other FCM-based models. FCM-SSF is the second most interpretable FCM-based model, with an 83.55% success rate, whereas FCM-FC has a success rate of 76.68%. Among the other models, LTCN, FCMMC, CCFCM, and FCMB follow, the results of which (74.27%, 58.12%, 50.49%, and 38.74%, respectively) suggest that they provide confusing explanations for the modelled system. The issues with FCMMC and FCMB are twofold: (a) the class label is extracted after a predetermined number of iterations (i.e., hyper-parameter depth) without the models being stabilized, and (b) the fully connected structure results in the unintentional absorption of spurious correlations. Moreover, the fully connected structure problem plagues FCM-FC, leading to poor interpretability. Regarding the CCFCM, the correlation coefficient is unreliable because its value is significant in the case of spurious correlations, even if the two variables are not causally related. Finally, due to the capture of spurious correlations, 1-step reasoning, the employed reasoning rule, and the sigmoid activation function, FCM-A cannot interpret individual predictions, making it an inappropriate model. Notably, the simulation results revealed that all input concepts had an activation value of 0.5 in the final iteration. This uniformity implies that the influence of individual input concepts on the model's prediction for a specific data instance remains indeterminate. Consequently, neither the average success rate for each failure mode nor the overall success rate across all four fault types could be calculated. To represent this lack of local interpretability in Table 5, the symbol "N/A" (not applicable, not available) was used for the FCM-A metrics.

Figure 4 summarizes the performance analysis results, visually representing the interpretability and accuracy of each model. The primary goal is to evaluate the trade-off

Table 5 Success rate of local explanations for each FCM-based model

Model	TWF	HDF	PWF	OSF	Average Success
IF-FCM[e]	**0.98333**	**0.97841**	**0.5835**	**0.95422**	**0.87488**
FCM-SSF[c]	0.59853	0.79313	1	0.95027	0.83546
FCM-FC	0.38382	0.92430	0.86459	0.96661	0.76681
LTCN[a]	0.31353	1	0.83241	0.82306	0.74273
FCMMC[b]	0.24391	0.41835	0.68756	0.86153	0.58120
CCFCM	0	1	0.56574	0.45380	0.50488
FCMB[b]	0.21682	0.79908	0.15910	0.39627	0.38740
FCM-A[d]	N/A	N/A	N/A	N/A	N/A

[a] Nápoles et al. (2022b)

[b] Szwed (2021)

[c] Nápoles et al. (2017)

[d] Froelich (2017)

[e] Proposed model

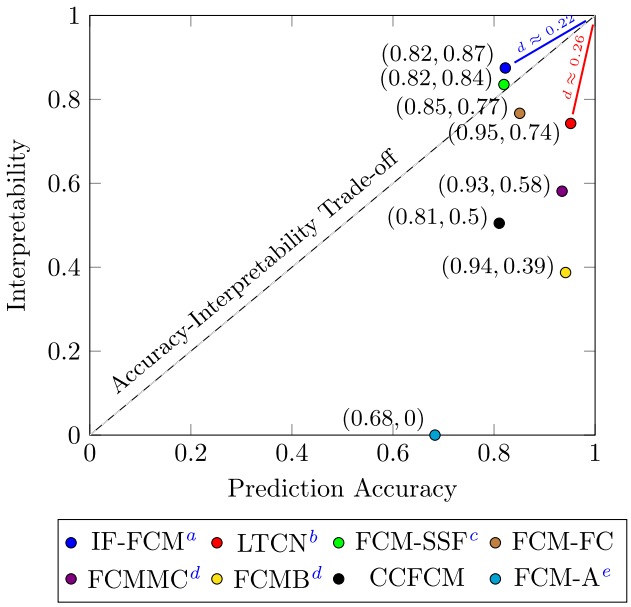

Fig. 4 Accuracy-interpretability trade-off and aggregate power for each FCM-based model. [a]Proposed Model, [b]Nápoles et al. (2022b), [c]Nápoles et al. (2017), [d]Szwed (2021), [e]Froelich (2017)

between accuracy and interpretability, along with the aggregate predictive and explanatory power of the models. The trade-off is measured as the absolute difference between the average accuracy and the average success rate score, while the aggregate power is quantified as their combined sum. In the scatter plot, the diagonal line $x = y$ represents the optimal trade-off, and the model positioned closer to the upper-right corner demonstrates the highest aggregate power. According to Fig. 4, IF-FCM has the maximum overall power (1.69723), surpassing all the other FCM-based models, whereas it has the second-best trade-off (0.05253), following FCM-SSF (0.01590). LTCN has the second-best aggregate power (1.69441); however, there is an imbalance between its prediction and interpretation scores (0.20895). Among the considered models, excluding FCM-A because of its lack of interpretability, CCFCM exhibited the poorest overall performance (1.31517).

Based on the conducted experiments and comparisons, it can be concluded that IF-FCM is a reliable predictor of machine failures. The model's enhanced explanatory power can be attributed to its ability to capture authentic causal relationships among problem variables. The global interpretability results of IF-FCM align with those of the examined models and previous research works. However, it is important to note that different models emphasize different features. In terms of local interpretability, IF-FCM outperformed the other models, providing more coherent explanations. Overall, the method's capability to rule out spurious correlations enhances the overall power of FCM, establishing it as a robust and interpretable model.

5 Conclusions

In this paper, a novel approach is presented for constructing FCMs using Liang-Kleeman Information Flow (L-K IF) analysis, an effective tool for causal inference. The motivation for this study stems from the pressing challenge of spurious correlations present in previous expert-based and data-driven FCM construction approaches. Our primary contribution is the formulation of a strategy that effectively mitigates spurious correlations, thereby enhancing the aggregate predictive and explanatory capabilities of FCM. By integrating L-K IF analysis into FCMs, we introduced an automated causal search algorithm that reliably

identifies authentic causal relationships from the data. These identified relationships subsequently served as constraints in the FCM learning process. To validate our approach, we applied it to a synthetic dataset tailored for industrial anomaly detection and root cause analysis as a proof-of-concept, resulting in improved performance of the developed FCM compared to other FCM-based models. While we acknowledge the potential value of incorporating additional datasets, our focus on a single dataset effectively highlights the unique contributions, innovations, and advantages of our method within a specific context, thus paving the way for future studies to explore its generalizability across multiple datasets. Moving forward, we plan to extend this study to real-world industrial data experiments, while investigating the challenges associated with metaheuristic learning algorithms.

Author Contributions Conceptualization, MT.; methodology, MT; validation, MT, XSL and CS; formal analysis, MT; investigation, MT, XSL.; writing-original draft preparation, MT; writing-review and editing, MT, XSL and CS; visualization, MT; supervision, CS; project administration, CS; funding acquisition, CS. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by HEAL-Link Greece. We acknowledge the support of this work by the project "Dioni: Computing Infrastructure for Big-Data Processing and Analysis." (MIS No. 5047222) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and cofinanced by Greece and the European Union (European Regional Development Fund). XSL is partially funded by the National Science Foundation of China under Grant #42230105, and by Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) through the startup foundation and scientific research program.

Data availability The data that support the findings of this study are available at https://doi.org/10.24432/C5HS5C.

Material and/or code availability To ensure reproducible science, all the necessary code and data required to replicate the results presented in this paper are available at the following link: https://doi.org/10.5281/zenodo.8124066. The code was implemented using both MATLAB and Python. Any updates to the code will be published in Zenodo, and the final DOI will be cited in the manuscript to ensure proper attribution and accessibility.

Declarations

Conflict of interest The authors declare they have no financial interests.

References

Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell 298(103502):103502. https://doi.org/10.1016/j.artint.2021.103502

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/access.2018.2870052

Alfeo AL, Cimino MGCA, Gagliardi G (2023) Concept-wise granular computing for explainable artificial intelligence. Granul Comput 8(4):827–838. https://doi.org/10.1007/s41066-022-00357-8

Alonso JM, Castiello C, Mencar C (2015) Interpretability of fuzzy systems: current research trends and prospects. Springer handbook of computational intelligence. Springer, Berlin, Heidelberg, pp 219–237. https://doi.org/10.1007/978-3-662-43505-2_14

Barbrook-Johnson P, Penn AS (2022) Fuzzy cognitive mapping. Systems mapping. Springer, Cham, pp 79–95. https://doi.org/10.1007/978-3-031-01919-7_6

Bas E, Egrioglu E, Kolemen E (2022) Training simple recurrent deep artificial neural network for forecasting using particle swarm optimization. Granul Comput 7(2):411–420. https://doi.org/10.1007/s41066-021-00274-2

Boutalis Y, Kottas TL, Christodoulou M (2009) Adaptive estimation of fuzzy cognitive maps with proven stability and parameter convergence. IEEE Trans Fuzzy Syst 17(4):874–889. https://doi.org/10.1109/tfuzz.2009.2017519

Brito LC, Susto GA, Brito JN et al (2022) An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mech Syst Signal Process 163(108105):108105. https://doi.org/10.1016/j.ymssp.2021.108105

Cao XH, Stojkovic I, Obradovic Z (2016) A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics 17(1):359. https://doi.org/10.1186/s12859-016-1236-x

Carletti M, Masiero C, Beghi A, et al (2019) Explainable machine learning in industry 4.0: evaluating feature importance in anomaly detection to enable root cause analysis. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, pp 21–26. https://doi.org/10.1109/smc.2019.8913901

Chen SJ, Chen SM (2002) A new method to measure the similarity between fuzzy numbers. In: 10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297), vol 3. IEEE, pp 1123–1126 vol.2, https://doi.org/10.1109/FUZZ.2001.1008852

Chen SM, Fang Yd (2005) A new method to deal with fuzzy classification problems by tuning membership functions for fuzzy classification systems. J Chin Inst Eng 28(1):169–173. https://doi.org/10.1080/02533839.2005.9670983

Chen SM, Jian WS (2017) Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups, similarity measures and PSO techniques. Inf Sci (NY) 391–392:65–79. https://doi.org/10.1016/j.ins.2016.11.004

Chen SM, Niou SJ (2011) Fuzzy multiple attributes group decision-making based on fuzzy preference relations. Expert Syst Appl 38(4):3865–3872. https://doi.org/10.1016/j.eswa.2010.09.047

Chen SM, Wang NY (2010) Fuzzy forecasting based on fuzzy-trend logical relationship groups. IEEE Trans Syst Man Cybern B Cybern 40(5):1343–1358. https://doi.org/10.1109/TSMCB.2009.2038358

Chen YC, Wang LH, Chen SM (2006) Generating weighted fuzzy rules from training data for dealing with the iris data classification problem. Int J Appl Sci Eng 4(1):41–52. https://doi.org/10.6703/IJASE.2006.4(1).41

Chen SM, Ko YK, Chang YC et al (2009) Weighted fuzzy interpolative reasoning based on weighted increment transformation and weighted ratio transformation techniques. IEEE Trans Fuzzy Syst 17(6):1412–1427. https://doi.org/10.1109/TFUZZ.2009.2032651

Czerwinski D, Czerwinska M, Karczmarek P, et al. (2021) Influence of the fuzzy robust gamma rank correlation, fuzzy c-means, and fuzzy cognitive maps to predict the Z generation's acceptance attitudes towards internet health information. In: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, pp 1–6. https://doi.org/10.1109/fuzz45933.2021.9494596

Egrioglu E, Bas E, Cansu T et al (2022) A new nonlinear causality test based on single multiplicative neuron model artificial neural network: a case study for turkey's macroeconomic indicators. Granul Comput 8(2):391–396. https://doi.org/10.1007/s41066-022-00336-z

Eichler M (2013) Causal inference with multiple time series: principles and problems. Philos Trans A Math Phys Eng Sci 371(1997):20110613. https://doi.org/10.1098/rsta.2011.0613

Erdem A (2023) Aerdem4/lofo-importance. https://github.com/aerdem4/lofo-importance/tree/master

Falcon R, Nápoles G, Bello R et al (2019) Granular cognitive maps: a review. Granul Comput 4(3):451–467. https://doi.org/10.1007/s41066-018-0104-7

Forward CF (2022) Causality for machine learning. Cloudera Fast Forward Labs Research, Santa Clara, CA. https://ff13.fastforwardlabs.com/FF13-Causality_for_Machine_Learning-Cloudera_Fast_Forward.pdf

Froelich W (2017) Towards improving the efficiency of the fuzzy cognitive map classifier. Neurocomputing 232:83–93. https://doi.org/10.1016/j.neucom.2016.11.059

Ghasemkhani B, Aktas O, Birant D (2023) Balanced K-Star: an explainable machine learning method for Internet-of-Things-enabled predictive maintenance in manufacturing. Machines 11(3):322. https://doi.org/10.3390/machines11030322

Gilchrist A (2016) Industry 4.0, 1st edn. APress, Berlin, Germany. https://doi.org/10.1007/978-1-4842-2047-4

Hlavackovaschindler K, Palus M, Vejmelka M et al (2007) Causality detection based on information-theoretic approaches in time series analysis. Phys Rep 441(1):1–46. https://doi.org/10.1016/j.physrep.2006.12.004

Kök İ, Okay FY, Muyanlı Ö et al (2023) Explainable artificial intelligence (XAI) for internet of things: a survey. IEEE Internet Things J 10(16):14764–14779. https://doi.org/10.1109/jiot.2023.3287678

Kosko B (1986) Fuzzy cognitive maps. Int J Man Mach Stud 24(1):65–75. https://doi.org/10.1016/s0020-7373(86)80040-2

Kosko B (1988) Hidden patterns in combined and adaptive knowledge networks. Int J Approx Reason 2(4):377–393. https://doi.org/10.1016/0888-613x(88)90111-9

Kubat M, Matwin S, et al. (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, Citeseer, p 179

Leader JJ (1991) Limit orbits of a power iteration for dominant eigenvalue problems. Appl Math Lett 4(4):41–44. https://doi.org/10.1016/0893-9659(91)90051-v

Lee KS, Kim SH, Sakawa M et al (1997) Process fault diagnosis by using fuzzy cognitive map. Trans Soc Instrum Control Eng 33(12):1155–1163. https://doi.org/10.9746/sicetr1965.33.1155

Li Z, Wang Y, Wang KS (2017) Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: industry 4.0 scenario. Adv Manuf 5(4):377–387. https://doi.org/10.1007/s40436-017-0203-8

Li X, Xiong H, Li X et al (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64(12):3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Liang XS (2008) Information flow within stochastic dynamical systems. Phys Rev E Stat Nonlin Soft Matter Phys 78(3 Pt 1):031113. https://doi.org/10.1103/physreve.78.031113

Liang XS (2014) Unraveling the cause-effect relation between time series. Phys Rev E Stat Nonlin Soft Matter Phys 90(5–1):052150. https://doi.org/10.1103/physreve.90.052150

Liang XS (2016) Information flow and causality as rigorous notions ab initio. Phys Rev E 94(5):052201. https://doi.org/10.1103/physreve.94.052201

Liang XS (2021) Normalized multivariate time series causality analysis and causal graph reconstruction. Entropy (Basel) 23(6):679. https://doi.org/10.3390/e23060679

Liu F, Peng Y, Chen Z et al (2020) Modeling of characteristics on artificial intelligence IQ test: a fuzzy cognitive map-based dynamic scenario analysis. Int J Comput Commun Control 14(6):653. https://doi.org/10.15837/ijccc.2019.6.3692

Loia V, D'Aniello G, Gaeta A et al (2016) Enforcing situation awareness with granular computing: a systematic overview and new perspectives. Granul Comput 1(2):127–143. https://doi.org/10.1007/s41066-015-0005-y

Matzka S (2020) Explainable artificial intelligence for predictive maintenance applications. In: 2020 Third International Conference on Artificial Intelligence for Industries (AI4I). IEEE, pp 69–74. https://doi.org/10.1109/ai4i49448.2020.00023

Mises RV, Pollaczek-Geiringer H (1929) Praktische verfahren der gleichungsauflösung. ZAMM - J Appl Math Mech/Z Angew Math Mech 9(2):152–164. https://doi.org/10.1002/zamm.19290090206

Mpelogianni V, Groumpos PP (2018) Re-approaching fuzzy cognitive maps to increase the knowledge of a system. AI Soc 33(2):175–188. https://doi.org/10.1007/s00146-018-0813-0

Mylonas N, Mollas I, Bassiliades N et al (2023) Local multi-label explanations for random forest. Communications in computer and information science. Springer, Cham, pp 369–384. https://doi.org/10.1007/978-3-031-23618-1_25

Nápoles G, Papageorgiou E, Bello R et al (2016) On the convergence of sigmoid fuzzy cognitive maps. Inf Sci (NY) 349–350:154–171. https://doi.org/10.1016/j.ins.2016.02.040

Nápoles G, Leon M, Grau I et al (2017) Fuzzy cognitive maps tool for scenario analysis and pattern classification. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp 644–651. https://doi.org/10.1109/ictai.2017.00103

Nápoles G, Jastrzębska A, Mosquera C et al (2020a) Deterministic learning of hybrid fuzzy cognitive maps and network reduction approaches. Neural Netw 124:258–268. https://doi.org/10.1016/j.neunet.2020.01.019

Nápoles G, Salmeron JL, Froelich W et al (2020b) Fuzzy cognitive modeling: theoretical and practical considerations. Intelligent decision technologies 2019. Smart innovation, systems and technologies. Springer, Singapore, pp 77–87. https://doi.org/10.1007/978-981-13-8311-3_7

Nápoles G, Grau I, Concepción L et al (2022a) Modeling implicit bias with fuzzy cognitive maps. Neurocomputing 481:33–45. https://doi.org/10.1016/j.neucom.2022.01.070

Nápoles G, Salgueiro Y, Grau I et al (2022b) Recurrence-aware long-term cognitive network for explainable pattern classification. IEEE Trans Cybern, pp. 1–12. https://doi.org/10.1109/tcyb.2022.3165104

Orang O, de Lima e Silva PC, Guimarães FG (2022) Time series forecasting using fuzzy cognitive maps: a survey. Artif Intell Rev 56(8):7733–7794. https://doi.org/10.1007/s10462-022-10319-w

Pant M, Kumar S (2022) Particle swarm optimization and intuitionistic fuzzy set-based novel method for fuzzy time series forecasting. Granul Comput 7(2):285–303. https://doi.org/10.1007/s41066-021-00265-3

Papageorgiou EI (2012) Learning algorithms for fuzzy cognitive maps–a review study. IEEE Trans Syst Man Cybern C Appl Rev 42(2):150–163. https://doi.org/10.1109/tsmcc.2011.2138694

Papageorgiou EI, Salmeron JL (2013) A review of fuzzy cognitive maps research during the last decade. IEEE Trans Fuzzy Syst 21(1):66–79. https://doi.org/10.1109/tfuzz.2012.2201727

Papageorgiou EI, Stylios CD (2008) Fuzzy cognitive maps. Handbook of granular computing. Wiley, Chichester, UK, pp 755–774. https://doi.org/10.1002/9780470724163.ch34

Papageorgiou EI, Parsopoulos KE, Stylios CS et al (2005) Fuzzy cognitive maps learning using particle swarm optimization. J Intell Inf Syst 25(1):95–121. https://doi.org/10.1007/s10844-005-0864-9

Papakostas GA, Boutalis YS, Koulouriotis DE et al (2008) Fuzzy cognitive maps for pattern recognition applications. Intern J Pattern Recognit Artif Intell 22(08):1461–1486. https://doi.org/10.1142/s0218001408006910

Papakostas GA, Koulouriotis DE, Polydoros AS et al (2012) Towards Hebbian learning of fuzzy cognitive maps in pattern classification problems. Expert Syst Appl 39(12):10620–10629. https://doi.org/10.1016/j.eswa.2012.02.148

Rehse JR, Mehdiyev N, Fettke P (2019) Towards explainable process predictions for industry 4.0 in the DFKI-smart-Lego-factory. KI - Künstl Intell 33(2):181–187. https://doi.org/10.1007/s13218-019-00586-1

Rohrer JM (2018) Thinking clearly about correlations and causation: graphical causal models for observational data. Adv Methods Pract Psychol Sci 1(1):27–42. https://doi.org/10.1177/2515245917745629

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Runge J, Heitzig J, Marwan N et al (2012) Quantifying causal coupling strength: a lag-specific measure for multivariate time series related to transfer entropy. Phys Rev E Stat Nonlin Soft Matter Phys 86(6 Pt 1):061121. https://doi.org/10.1103/physreve.86.061121

Shen VRL, Chung YF, Chen SM et al (2013) A novel reduction approach for petri net systems based on matching theory. Expert Syst Appl 40(11):4562–4576. https://doi.org/10.1016/j.eswa.2013.01.057

Slack D, Hilgard S, Jia E, et al. (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. ACM, New York, NY, USA, pp 180-186. https://doi.org/10.1145/3375627.3375830

Soler LS, Kok K, Camara G et al (2012) Using fuzzy cognitive maps to describe current system dynamics and develop land cover scenarios: a case study in the Brazilian amazon. J Land Use Sci 7(2):149–175. https://doi.org/10.1080/1747423x.2010.542495

Song H, Mioa C, Roel W et al (2009) Implementation of fuzzy cognitive maps using fuzzy neural network and application in prediction of time series. IEEE Trans Fuzzy Syst 18(2):233–250. https://doi.org/10.1109/tfuzz.2009.2038371

Song HJ, Miao CY, Wuyts R et al (2011) An extension to fuzzy cognitive maps for classification and prediction. IEEE Trans Fuzzy Syst 19(1):116–135. https://doi.org/10.1109/tfuzz.2010.2087383

Sridhar S, Sanagavarapu S (2021) Handling data imbalance in predictive maintenance for machines using SMOTE-based oversampling. In: 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, pp 44–49. https://doi.org/10.1109/cicn51697.2021.9574668

Stylios CD, Groumpos PP (1998) Fuzzy cognitive map model for supervisory manufacture systems. Intelligent systems for manufacturing. IFIP advances in information and communication technology. Springer, US, Boston, MA, pp 137–146. https://doi.org/10.1007/978-0-387-35390-6_12

Stylios CD, Groumpos PP (2004) Modeling complex systems using fuzzy cognitive maps. IEEE Trans Syst Man Cybern A Syst Hum 34(1):155–162. https://doi.org/10.1109/tsmca.2003.818878

Szwed P (2021) Classification and feature transformation with fuzzy cognitive maps. Appl Soft Comput 105(107271):107271. https://doi.org/10.1016/j.asoc.2021.107271

Tirovolas M, Stylios C (2022) Introducing fuzzy cognitive map for predicting engine's health status. IFAC-PapersOnLine 55(2):246–251. https://doi.org/10.1016/j.ifacol.2022.04.201

Tyrovolas M, Liang XS, Stylios C (2023) Information flow-based fuzzy cognitive maps with enhanced interpretability - Source Code. https://doi.org/10.5281/zenodo.8124066

Wang Z, Culotta A (2021) Robustness to spurious correlations in text classification via automatically generated counterfactuals. Proc Conf AAAI Artif Intell 35(16):14024–14031. https://doi.org/10.48550/arXiv.2012.10040

Wang J, Peng Z, Wang X et al (2021) Deep fuzzy cognitive maps for interpretable multivariate time series prediction. IEEE Trans Fuzzy Syst 29(9):2647–2660. https://doi.org/10.1109/tfuzz.2020.3005293

Wang X, Yang J, Lu W (2022) Bearing fault diagnosis algorithm based on granular computing. Granul Comput 8(2):333–344. https://doi.org/10.1007/s41066-022-00328-z

Yosef A, Shnaider E, Schneider M et al (2022) Relative influences and the reliability of weights in fuzzy cognitive maps. Fuzzy Sets Syst 449:100–119. https://doi.org/10.1016/j.fss.2022.01.011

Zadeh LA (1965) Fuzzy sets. Inf Contr 8(3):338–353. https://doi.org/10.1016/S0019-9958(65)90241-X