A Novel Framework for Enhanced Interpretability in Fuzzy Cognitive Maps

SUBMISSION DATE / POSTED DATE

28-04-2023 / 02-05-2023

CITATION

Tyrovolas, Marios; Liang, X. San; Stylios, Chrysostomos (2023): A Novel Framework for Enhanced Interpretability in Fuzzy Cognitive Maps. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22718032.v1

DOI

10.36227/techrxiv.22718032.v1

A Novel Framework for Enhanced Interpretability in Fuzzy Cognitive Maps

Marios Tyrovolas, *Student Member, IEEE,* X. San Liang,
and Chrysostomos Stylios, *Senior Member, IEEE*

***Abstract*—A Fuzzy Cognitive Map (FCM) is a graph-based tool for knowledge representation that intends to model any complex system through an interactive structure of nodes interacting with each other through causal relationships. Owing to their flexibility and inherent interpretability, FCMs have been used in various modeling and prediction tasks, particularly in situations where humans make final decisions, such as industrial anomaly detection. However, FCMs can unintentionally absorb spurious correlations presented in collected data during development, leading to poor prediction accuracy and interpretability. To address this limitation, this article proposes a novel framework for constructing FCMs based on the Liang-Kleeman Information Flow (L-K IF) analysis, a causal inference tool. The actual causal relationships are identified from the data using an automatic causal search algorithm, and these are then imposed as constraints in the FCM learning procedure to rule out spurious correlations and improve the predictive and explanatory power of the model. Numerical simulations were conducted by comparing the proposed approach with state-of-the-art FCM-based models, thereby demonstrating the promising performance of the developed FCM.**

***Index Terms*—Fuzzy Congnitive Maps, explainable AI, quantitative causality, Information Flow, Industry 4.0**

I. 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 [1]. 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. Several AI solutions, such as Support Vector Machines and Artificial Neural Networks, have been presented, demonstrating great accuracy in predicting production line malfunctions [2]. 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 [3]. Moreover, the lack of transparency and interpretability prevents the identification of the weaknesses of the employed AI algorithms through their explanations and, eventually, their improvement. Therefore, it is necessary to create AI models with more interpretable behavior, whose explanations can contribute to finding the root cause of a decision, helping humans respond appropriately [4].

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 [5]. Specifically, depending on the application, different explanation methods can be employed from an XAI system, such as the detection of important pixels for image classification or *IF-THEN* rules that express input-output data relationships [6], [7]. Therefore, transitioning from AI to XAI is imperative to successfully integrate automated decision-making into production systems, where humans make supervision and final decisions.

A. State-of-the-Art & Motivation

Over the last few years, the research community has proposed two distinct categories of XAI methodologies based on how they are implemented, named:

- **Post hoc techniques**: Techniques that build a second interpretable surrogate model, i.e., *explainer*, to approximate the underlying model and explain its predictions.
- **Intrinsic interpretable models**: Models that can explain their predictions by themselves.

Because post hoc techniques are adapted to the underlying model, they may not accurately imitate it, leading to incorrect interpretations [8]. Furthermore, even if the approximation from the *explainer* is good, the interpretations will be erroneous if the underlying model misunderstands the relationships among the training data [9]. Finally, another disadvantage is that the explanations of these techniques can be easily controlled to be acceptable through specific frameworks, even if the base model is highly biased [10]. Thus, academics have turned to intrinsic interpretable models, whose decisions can be explained without additional techniques, and are able to represent assimilated knowledge in a manner consistent with human thought [11].

A widely used intrinsic interpretable tool for knowledge representation is Fuzzy Cognitive Maps (FCMs), a type of

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).

Marios Tyrovolas and Chrysostomos Stylios are with the Department of Informatics and Telecommunications, University of Ioannina, Arta 47150, Greece (e-mails: tirovolas@kic.uoi.gr, stylios@isi.gr).

X. San Liang is with the Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai 200438, China, and also with the AI Group, Division of Frontier Research, Southern Marine Laboratory, Zhuhai 519000, China (e-mail: xsliang@fudan.edu.cn).

recurrent neural network typically incorporating fuzzy logic features during its development, which can model complex systems, such as industrial systems [12]. Specifically, FCMs are directed graphs consisting of nodes called concepts that represent the components of the modelled system and weighted edges that describe the causal relations between them. The advantages of FCMs are threefold:

1) Their ability to use experts' assessments when the collected data are insufficient,
2) Their intrinsic explainability, as concepts and weights have well-defined meanings for the system under analysis and the transparent inference process.
3) The experts' capacity to modify the FCM's weights to encode rules that have not yet been observed in data (e.g., a new type of fault in the manufacturing system), a level of flexibility that cannot be achieved in other intrinsic interpretable models.

Considering the above, FCMs have piqued the interest of researchers and proved extremely useful in different domains. For instance, in the industry context, the authors of [13] proposed an FCM-based model for fault diagnosis in a tank-pipeline system that successfully identified various simulated faults, whereas the authors of [14] presented an FCM-based supervisor of manufacturing systems for failure detection and decision analysis. Lastly, [15] proposed FCMs as a health indicator prognostic method for engines' remaining useful life in the context of predictive maintenance. Nevertheless, it should be mentioned that even if the literature frequently mentions FCMs' interpretable nature, it mainly rests on the fact that their concepts and weights have a clear meaning without demonstrating their explanatory performance. Therefore, thorough numerical simulations should be performed to determine the capabilities of FCMs to explain their decisions.

To evaluate the performance of FCMs in terms of interpretability, it is important to describe how they can be developed. Currently, two fundamental FCM construction methods have been proposed in the literature: a) expert-based and b) data-driven [16]. In the first method, FCM concepts and weights are determined only by the knowledge of domain experts, which is incorporated into the model using fuzzy logic theory [17]. However, as the developed model depends on their expertise level, its performance may not be satisfactory as they may overlook essential aspects of the problem and assign inappropriate weight values [16]. In contrast, in the data-driven approach, FCM parameters are defined using learning algorithms [18]. Specifically, during FCM learning, either the presence of all weights is assumed, leading to an over-parameterized model, or they are calculated using the correlation coefficients between the variables [19], [20]. Nevertheless, the dataset may contain *spurious correlations* that are unintentionally absorbed from the FCM and bias its learning, leading to poor prediction accuracy and interpretability [21], [22]. Incorrect explanations are an important issue for successfully implementing FCMs for industrial anomaly detection, as the model directs plant supervisors to the wrong parts of the manufacturing system where the root cause of the fault cannot be found. Thus, developing a new method that identifies the authentic causal relations between problem variables and rules out possible spurious correlations, is considered essential [23]. In this direction, the authors in [24] 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.

One solution to these limitations is to develop a method that identifies the real causal structure of an FCM from observational data, without requiring domain experts. In this way, a) the injection of spuriousness into these cognitive networks can be avoided, thus improving their prediction accuracy and interpretability, and b) large-scale problems can be efficiently encountered. To the best of the authors' knowledge, no data-driven causal discovery method has been proposed for constructing robust and interpretable FCMs.

B. Contribution

In this paper, a novel approach for FCM construction is introduced based on the causal inference tool Liang–Kleeman Information Flow (L-K IF) analysis. In more detail, in contrast to [24], the proposed technique does not require expert involvement because it identifies the actual causal relationships from the data using an automatic causal search algorithm. Finally, the derived causal links are imposed as constraints in the FCM learning procedure, aiming to rule out spurious correlations and thus 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 should be highlighted that, even if the examined case study concerns anomaly detection, the proposed method can be used effectively in other prediction problems.

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

II. 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.

A. Fuzzy Cognitive Maps

As mentioned in Section I-A, an FCM consists of n concepts $C_i\ ,i \in \{1,2,\ldots,n\}$, and weights $w_{ij} \in [-1,1]$ that indicate the causal relation 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 which is determined via a reasoning rule, where the most common is

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

where t is the iteration step, $A_m^{(p)}$ denotes the value of the m-th concept at p-th iteration step, w_{ji} denotes the causal weight from j-th concept to i-th concept, and $f(\cdot)$ denotes the activation function that normalizes the concepts' activation values within a specified interval [12]. The most known activation functions are bivalent, trivalent, hyperbolic tangent, and sigmoid, where depending on which is selected, $A_m^{(p)}$ receives values within the $[0,1]$ or $[-1,1]$ intervals [25]. 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:

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

The activation values of the concepts in each iteration step are calculated using (2) and an initial state vector $\mathbf{A}^{(0)}$ as an input that contains input data (e.g., sensor data), and triggers the FCM to start an iterative reasoning process. 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 [26].

B. Information Flow

As mentioned above, to develop efficient FCMs, it is important to identify authentic causal relations between the modelled system variables, which can be achieved through causality analysis. In the context of causality analysis, [27] proposed a framework for quantifying the causal relations between dynamic system variables, in which causality is expressed as a physical notion called Information Flow (IF). 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. 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 [28]. For the aforementioned system, the IF from x_2 to x_1 is

$$T_{2\to1} = -E\left(\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right) + \frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2 \mathrm{g}_{11}\rho_1}{\partial x_1^2}\right), \tag{4}$$

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, $\mathrm{g}_{11} = \sum_{k=1}^{2} b_{1k}^2$, 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 ($T_{2\to1} = 0$) if the evolution of the latter is independent of the former (neither F_1 nor g_{11} depends on x_2) [29].

As a further step, [27] 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):

$$T_{2\to1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^2 C_{1,d1}}{C_{11}^2 C_{22} - C_{11}C_{12}^2}, \tag{5}$$

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: $\dot{X}_{j,n} = (X_{j,n+k} - X_{j,n})\,/(k\Delta t)$, with $k \geq 1$ some integer. The IF in the opposite direction, i.e., $T_{1\to2}$, is obtained by swapping indices 1 and 2. Besides, writing (5) as a function of correlation and/or correlation-like quantities gives

$$T_{2\to1} = \frac{r}{1-r^2}(\acute{r}_{2,d1} - r\,\acute{r}_{1,d1}), \tag{6}$$

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 ($T_{2\to1} \neq 0$), 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 [30]. In detail, given a dataset of d time-series variables, the IF from X_2 to X_1 is

$$\hat{T}_{2\to1} = \frac{1}{detC}\cdot\sum_{j=1}^{d}\Delta_{2j}C_{j,d1}\cdot\frac{C_{12}}{\mathrm{C}_{11}}, \tag{7}$$

where $C_{j,d1}$ is the sample covariance between X_j and $\dot{X}_1$, and Δ_{ij} the are the cofactors of the covariance matrix C. An algorithm for multivariate time-series causality analysis is developed based on (7) (Algorithm 1). 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 $\alpha$
  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 $\alpha$ 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 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 $\hat{T}_{2 \to 1}$ being

$$\hat{Z} = \left| \left(\widehat{\frac{\mathrm{d}H_1^*}{\mathrm{d}t}} \right) \right| + \sum_{j=2}^{d} |\hat{T}_{j \to 1}| + \left| \left(\widehat{\frac{\mathrm{d}H_1^{noise}}{\mathrm{d}t}} \right) \right|, \tag{8}$$

where

$$\left(\widehat{\frac{\mathrm{d}H_1^*}{\mathrm{d}t}} \right) = \frac{1}{detC} \cdot \sum_{j=1}^{d} \Delta_{1j} C_{j,d1}, \tag{9}$$

$$\left(\widehat{\frac{\mathrm{d}H_1^{noise}}{\mathrm{d}t}} \right) = \frac{1}{2} \frac{\hat{\mathrm{g}}_{11}}{C_{11}}, \tag{10}$$

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

$$\tau_{2 \to 1} = \frac{T_{2 \to 1}}{\hat{Z}} \tag{11}$$

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

C. L-K IF Analysis on Binary Time Series

Until now, studies that employed L-K IF analysis to identify causal relations have not focused on discrete-valued signals that take a few values, such as binary time series. However, real datasets, especially in the industry, usually contain binary variables, such as the state of a proximity sensor or button. Consequently, an experiment was conducted to ensure that the causal inference tool successfully handled binary data types.

Let be a dataset of three time series X_1, X_2, and X_3, where X_3 is the confounder of the other two without any other causality, which are expressed mathematically as

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

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

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

where 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 Ia 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 $T_{3 \to 1}$ and $T_{3 \to 2}$ 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 Ib), it is concluded that the proposed technique reliably captures the causal relations in a qualitative sense, even if the time series have been binarized.

III. Proposed Methodology

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

A. Data pre-processing

After collecting data from the analyzed system, for example, a manufacturing system, appropriate data pre-processing techniques are applied. Initially, because FCM processes only numeric data, it is necessary to encode categorical variables, including class attributes, in a classification problem. 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 $j \in \{1, \dots, m\}$ and $m \geq 2$ the number of class labels.

In the context of FCMs, a mandatory process is the assignment of fuzzy values to concepts, which is called data fuzzification. Fuzzification is practically considered a data normalization procedure that computes the concepts' initial activation values for each data observation. Representative normalization techniques are the 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 susception to outliers. In addition, concerning the min-max normalization, different data separations, for example, in cross-validation, yield different normalizations. With this in mind,

		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	$\boxed{0.2033 \pm 0.0058}$	$\boxed{0.0635 \pm 0.0022}$	\
		$T_{3\to1}$	$T_{3\to2}$	

(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	$\boxed{0.0918 \pm 0.0046}$	$\boxed{0.0187 \pm 0.0020}$	\
		$T_{3\to1}$	$T_{3\to2}$	

(b) Binarized Time Series

TABLE I: IF rates for the series generated with (12) and their respective confidence intervals (99% confidence level).

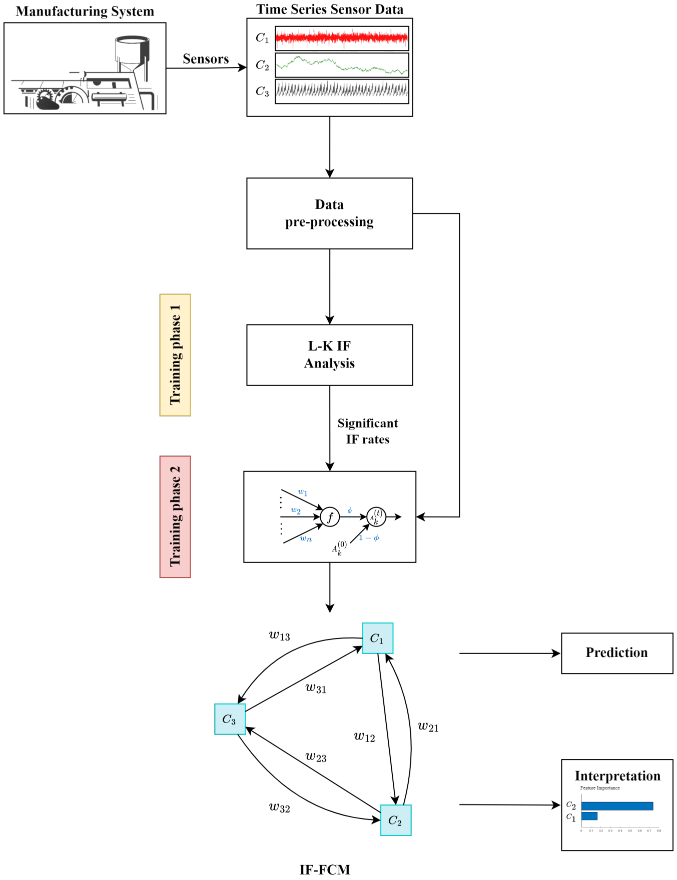

Fig. 1: Proposed Methodology Scheme

the Generalized Logistic (GL) algorithm was employed in this study to normalize the data [31]. This algorithm makes no assumptions about the distribution of variables but instead uses a generalized logistic function to approximate each variable's corresponding cumulative density function (CDF). The main advantage of this method is its inherent robustness against outliers. The algorithm maps values from interval $(-\infty, \infty)$ to interval $[0, 1]$.

B. Information Flow-Based Fuzzy Cognitive Map (IF-FCM)

Having prepared the data, the FCM architecture that specifies the type and number of concepts must be defined. In the context of classification, the literature presents two main FCM architectures, whose difference lies in the number of output concepts (OCs), but also in the way of dealing with the prediction of the class label for each data instance. In the first *class-per-output architecture* (CpO), each class label is mapped to a different OC with m total outputs, and the OC with the highest activation value in the last iteration of the reasoning process indicates the predicted class. In contrast, in the second architecture, called *single-output architecture* (SO), the class attribute is mapped to a single OC C_n whose predicted activation value should be assigned to one of the class labels [32]. To achieve this, the activation interval ($[0, 1]$ or $[-1, 1]$) is divided into partitions, each corresponding to a class label. More specifically, the prediction process in an FCM-SO is as follows:

Step 1: Consider the k-th 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 $i \in \{1, 2, \ldots, n\}$. 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 class labels. This is accomplished using $m-1$ defined decision thresholds that divide the activation interval into m partitions. Therefore, depending on the range $A_{nk}^{(l)}$ belongs, the FCM predicts the corresponding class label. To determine the decision thresholds, a "threshold-moving" approach is employed, which finds the best 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 and, therefore, determines the ideal position of the classification hyperplane [33].

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

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, such as the work presented in [34], concluded that the SO architecture outperformed the CpO architecture on seven of the eight datasets analyzed.

C. IF-FCM Learning

After the architecture is determined, 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 are very different; therefore, raw encoded data are used.

In the second phase (*Training phase 2*), the parameters defining the FCM response are tuned, which are the weights and parameters (if any) of the employed activation function and reasoning rule. Consequently, a challenging question arises regarding the choice of the appropriate reasoning rule and activation function. Various researchers have shown that using (1) in conjunction with the activation functions mentioned in Section II-A, FCM usually converges to the same equilibrium point regardless of the initial state vector [35]. This behavior is undesirable in forecasting tasks, such as anomaly detection, because the model predicts only one class label. Beyond that, through these bounded activation functions, the saturation problem appears, where the activation values of the concepts are placed during the iterative reasoning process in the lower or upper boundary of the specified interval when receiving a high negative or positive influence, respectively, [36]. 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 [35]

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 [37], 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

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

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 [36], the convergence features of the reasoning mechanism were analyzed. Through a mathematical proof by contradiction, they concluded that in an FCM that employs (18) and (19), when $\varphi \in [0,1)$, there is no pair of different initial stimuli leading to the same fixed-point attractor, or more loosely, there is no unique equilibrium point for all initial state vectors.

For $\varphi = 1$, the findings were 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, as $k \to \infty$, converges to such an eigenvector [38]. In particular, when λ_1 is real, the method converges to a fixed point. Nevertheless, because asymmetry is a distinguishing characteristic of causation, the convergence of FCM when $\varphi = 1$ should be analyzed without the above characteristics 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 a slower convergence [39], [40]. Therefore, the case of $\varphi = 1$ enables modeling scenarios in which FCM converges to a unique fixed-point attractor without the necessity 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 FCM's weights and the controllable parameter φ. However, the difference between the proposed method and 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, the weights of the edges whose estimated IF rate was significant were the only tunable parameters, while the rest were set to zero "a priori". Thus, in addition to improving the generalizability and interpretability of the developed FCM, the training time also decreases as the dimensions of the optimization problem are reduced. Therefore, a candidate solution is encoded as a $(\text{SIFs}+1)$-dimensional vector, where SIFs is the number of significant IF rates and parameter φ.

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

For FCM learning, that is, the optimal weight values and φ identification, we chose the Particle Swarm Optimization (PSO) meta-heuristic algorithm because of the effectiveness shown in the literature [41]. Specifically, PSO defines a population of candidate solutions called particles and compares them iteratively based on a cost function, which in this particular study, is

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

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 iteration, allowing greater flexibility at the beginning [42].

D. 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 on 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 specific decision. The ability to find the features that play a more critical role in classifying a specific sample as an anomaly enables root cause analysis [43].

D.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 [12]:

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

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.

D.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 activation values of all concepts across iterations [44]. 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 [45], [46]. The plot above can also help investigate how relative changes in the initial concept values impact the reasoning process, such as whether a change accelerates, stabilizes, or dies away.

IV. 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.

A. Dataset Description

We adopted Matzka's PMAI4I dataset to perform the experiments, which is a synthetic yet realistic dataset that represents industrial predictive maintenance data [47]. It contains 10,000 samples, each one containing one categorical variable (product quality $\in$ {"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, besides 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 minutes.
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 at least one of the aforementioned failure modes is true, the process fails and the machine failure value is set to one. However, during training, the FCM is fed only with the values of the input variables and system condition, without knowing the root cause of the fault. Consequently, the purpose of the FCM-based classifier is, on the one hand, to detect the anomaly's presence in the analyzed manufacturing system, and on the other hand, to indicate the most important input variable(s) of each true positive prediction that is likely to be the root cause of the fault by exploiting its inherent interpretability characteristics.

B. Simulations Execution

Following the methodology described in Section III, the data are first pre-processed, starting from encoding the categorical features, as in this case, the product quality, where each category value is assigned an integer starting from zero. Specifically, "low" corresponds to 0, "medium" to 1, and "high" to 2. Because the dataset is imbalanced, an SMOTE-based algorithm is used to address this issue, generating artificial instances of the minority class "1" [48]. In particular, this study applies the hybrid algorithm SMOTE-ENN, which

TABLE II: 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

TABLE III: LIST OF HYPER-PARAMETERS FOR MODEL TUNING

Model	Hyper-parameters
FCM-A	*g = 0 to 10*
FCMB	*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	*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	*method: inverse* *transfer function: sigmoid, tanh* *phi: 0.5 to 1.0* *T: 5, 10, 15* *alpha: 0, 0.01, 100*
FCM-SSF	*density: 10% to 100%* *slope parameter of the sigmoid: -1 to 10* *offset parameter of the sigmoid: -1 to 1*

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, leading to better overall performance than executing one of these techniques alone. Finally, data fuzzification was performed to prepare the data for training and decision-making.

B.1 Training phase 1

Applying Algorithm 1 to the dataset and then normalizing the significant IF rates yield Tab. II, whose results are consistent with the dataset description since:

1) The product quality determines the wear time added to the tool, which in turn causes the appearance of TWF and OSF.
2) The process temperature at each time step has been calculated using the air temperature samples, suggesting causality between them.
3) Air and process temperature are responsible for the appearance of HDF in the system. Therefore, information flows from these two 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 is causal to tool wear.
3) There is feedback from machine failure to tool wear.

B.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 the PSO execution. According to Tab. II, the final number of tunable weights is nine, plus the reasoning rule parameter φ. For the PSO parameter initialization, the population size is set to 100, and the cost function parameters α_1 and α_2 are set to 0.8 and 0.2, respectively. Finally, a hybrid function continues the optimization after the termination of the original solver to obtain a more accurate solution. The algorithm was implemented using MATLAB global optimization toolbox.

B.3 Performance Analysis of IF-FCM

After training the IF-FCM, we compared its predictive and explanatory power against state-of-the-art FCM-based models, including FCM-A [49], FCMBinaryClassifier (FCMB) [50], FCMMulticlassClassifier (FCMMC) [50], Long-Term Cognitive Network (LTCN) [37], and a Fuzzy Cognitive Map that uses the "*Stability based on Sigmoid Functions*" method (FCM-SSF) [51]. In addition, to highlight the significance of L-K IF analysis in improving FCM performance, two additional models were developed that employ (18) and (19) for their decision-making while being trained through PSO; however, their *Training phase 1* differs. Specifically, in the first model, called correlation coefficient-based FCM (CCFCM), the weights correspond to the correlation coefficients between variables whose p-value is less than 0.05, whereas in the second model (FCM-FC), all weights are included without performing primary data analysis to determine the relationships between concepts.

To identify possible issues such as overfitting and to check the generalizability of the models, stratified 10-fold cross-validation was used for the simulations. Simultaneously, hyper-parameter tuning was conducted to achieve the optimal performance of each model, considering the variables displayed in Tab. III. For FCM-SSF, we randomly produced 91 maps with network densities varying between 10% and 100%, and the model with the best performance was chosen.

Tab. IV demonstrates each model's mean prediction accuracy, "Area under the ROC Curve" (AUC) score and Cohen's kappa coefficient for all folds. According to the results, the LTCN, FCMB, and FCMMC are the three best-performing cognitive networks on these metrics, followed by FCM-FC and IF-FCM, whereas FCM-A performs the worst. The poor performance of FCM-A is because in the algorithm proposed in [49], the loop for calculating the classification error of each candidate threshold only considers false negatives minimiza-

TABLE IV: Mean Accuracy, AUC, and Kappa coefficient For each FCM-based model

Model	Accuracy	AUC	Kappa
LTCN	0.95168	0.95061	0.90295
FCMB	0.94159	0.94357	0.88412
FCMMC	0.93464	0.93432	0.86426
FCM-FC	0.85080	0.88231	0.70238
IF-FCM	0.82235	0.85465	0.64600
FCM-SSF	0.81956	-	0.63907
CCFCM	0.81029	0.81434	0.62009
FCM-A	0.68311	0.86089	0.35364

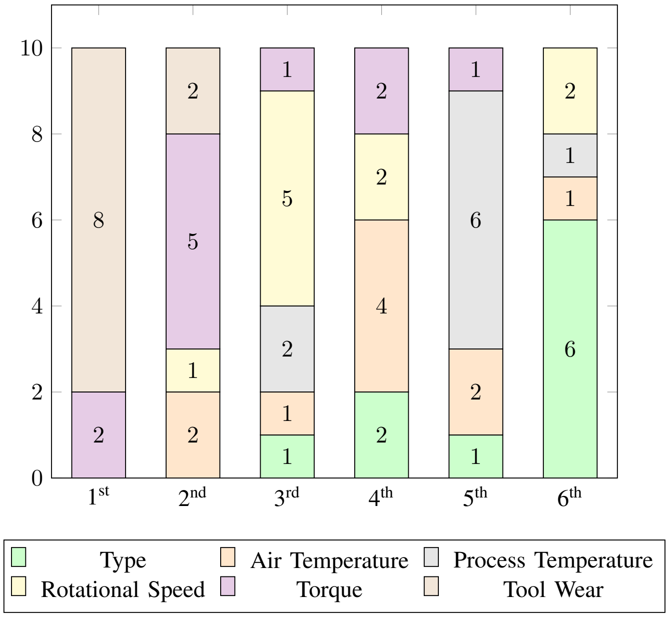

Fig. 2: The input features ranking based on their global importance for the examined models.

tion rather than the optimal balance between false negatives and false positives. This issue can also explain the discrepancy between the accuracy and AUC scores. Because the AUC is classification-threshold-invariant and provides an aggregate performance measure across all possible decision thresholds, its value combined with poor accuracy indicates that the chosen decision threshold is not optimal. Furthermore, the AUC score cannot be computed in FCM-SSF because it is based on the CpO architecture, which chooses the OC with the highest activation value in the last iteration, while the ROC Curve uses a decision threshold.

Regarding interpretability, the *global feature importance* in IF-FCM is computed using (24). As observed from its causal structure (Tab. II), the most important feature is tool wear with six incoming and outcoming edges, followed by air temperature with three edges, process temperature with 2, and the rest with one outcoming edge. The LOFO (Leave-One-Feature-Out) importance method, which is an XAI technique based on iterative variable removal for determining the mean value and standard deviation of the importance of each feature, is used to verify the results [52]. This method was chosen because, unlike linear models, which struggle to deliver meaningful information when dealing with correlated features, LOFO eliminates this concern while exhibiting solid generalization, as the feature importance is calculated across cross-validation splits. We, therefore, performed LOFO for various 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), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB). The results showed that all ML models recognized tool wear as the most important feature, except for XGB, which assessed process temperature as the first driver in decision-making and tool wear as the second (mean value 0.000754 and 0.000736, respectively). In addition, Logistic Regression (LR), an intrinsic interpretable model, was developed, whose feature coefficients indicate that torque is the most impactful feature, followed by air temperature and tool wear.

Regarding the examined FCM-based models, the LTCN exported its three most critical features through its mechanism: a) torque, b) tool wear, and c) rotational speed. In addition, the FCM-A based on an SO architecture, where input concepts are connected directly and only to the OC without feedback, does not represent the actual causal relationships, implying an inability to calculate the concept's degree of centrality [49, Fig. 1]. For example, product quality does not directly affect machine failure. The same problem appears in FCMB and FCMMC, where the fully connected map structure suggests the existence of spurious correlations and does not allow the calculation of each concept's centrality, because all concepts have the same number of edges. Summarizing the above outcomes in Fig. 2, it is observed that the largest percentage of models agreed on the importance of each feature. Identifying the *global feature importance* of the PMAI4I dataset has also concerned other researchers. Specifically, in [48], the authors also concluded that tool wear had the most considerable effect, followed by torque and rotational speed, thereby providing additional safety for the correctness of the outcomes.

Regarding local interpretability, Tab. V presents the average success rate of explaining the correctly predicted anomalous data instances, that is, determining the appropriate input features as the most important according to the type of fault that appeared in each observation. For example, in the case of TWF detection, the model should calculate tool wear as the most important variable, whereas in PWF, either the torque or rotational speed should be calculated. Furthermore, it shows the average success rate of each failure mode. As can be observed, the proposed model possesses the highest degree of interpretability (87.49% success), outperforming all the 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%. Then, the 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. Concerning FCMMC and FCMB, the reason is twofold: a) the class label is extracted after a predetermined number of iterations (i.e., hyper-parameter depth) without the

TABLE V: SUCCESS RATE OF LOCAL EXPLANATIONS FOR EACH FCM-BASED MODEL

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

TABLE VI: ACCURACY-INTERPRETABILITY TRADE-OFF AND AGGREGATE POWER FOR EACH FCM-BASED MODEL

Model	Trade-off	Aggregate Power
IF-FCM	0.05253	1.69723
LTCN	0.20895	1.69441
FCM-SSF	0.01590	1.65502
FCM-FC	0.08399	1.61761
FCMMC	0.35344	1.51584
FCMB	0.55420	1.32899
CCFCM	0.30541	1.31517
FCM-A	-	-

models being stabilized, and b) the fully connected structure results in the unintentional absorption of spurious correlations. 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. Moreover, the fully connected structure problem plagues the FCM-FC, leading to poor interpretability. Finally, FCM-A cannot interpret individual predictions because its topology, which contains spurious correlations, combined with the 1-step reasoning, the employed reasoning rule, and the sigmoid activation function, deceives the simulation results, with the values of all input concepts being 0.5, in the last iteration.

Finally, the performance analysis results are summarized in Tab. VI, which shows the aggregate predictive and explanatory power of the examined models and their accuracy-interpretability trade-off. According to this, IF-FCM has the maximum overall power, surpassing all other FCM-based models, whereas it has the second-best trade-off following FCM-SSF. The relevant list follows the LTCN, whose aggregate power is quite close; however, there is an imbalance between prediction and interpretation abilities. Among the considered models, excluding FCM-A because of its lack of interpretability, CCFCM exhibited the poorest overall performance.

Based on the experiments and comparisons conducted, it can be concluded that the IF-FCM is a reliable predictor of machine failures. The model's improved explanatory power is attributed to its ability to capture authentic causal relationships between problem variables. IF-FCM's results align with most of the examined models, and research works at the global interpretability level; however, it is essential to note that different models focus on different features. Regarding local interpretability, IF-FCM outperformed the other models, providing more correct explanations. Overall, the method's ability to rule out spurious correlations improves FCM's aggregate power, making it a powerful interpretable model.

V. CONCLUSIONS

This study presents a novel approach to constructing Fuzzy Cognitive Maps (FCMs) using Liang-Kleeman Information Flow (L-K IF) analysis, a causal inference tool. While other FCM-based implementations suffer from spurious correlations between problem variables, this proposal employs an automatic causal search algorithm to identify authentic causal relations from the data, and then it imposes them as constraints in the FCM learning procedure to rule out misleading relations. The effectiveness of the technique was evaluated using a realistic synthetic dataset to design an XAI model to detect anomalies and identify the root causes of an industrial system. The results confirm that the developed FCM has improved interpretability and predictive power compared with other FCM-based models. In the future, we plan to continue our work in this direction and investigate the issues of metaheuristic learning algorithms.

REFERENCES

[1] A. Gilchrist, *Industry 4.0*, 1st ed. Berlin, Germany: APress, Jun. 2016.

[2] Z. Li, Y. Wang, and K.-S. Wang, "Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario," *Adv. Manuf.*, vol. 5, no. 4, pp. 377–387, Dec. 2017.

[3] J.-R. Rehse, N. Mehdiyev, and P. Fettke, "Towards explainable process predictions for industry 4.0 in the DFKI-smart-Lego-factory," *KI - Künstl. Intell.*, vol. 33, no. 2, pp. 181–187, Jun. 2019.

[4] M. Carletti, C. Masiero, A. Beghi, and G. A. Susto, "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, Oct. 2019.

[5] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[6] R. Ibrahim, "Effective explainable artificial intelligence using visual explanations in images," Ph.D. dissertation, 2022.

[7] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2018.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, Aug. 2016.

[10] 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: ACM, Feb. 2020.

[11] J. M. Alonso, C. Castiello, and C. Mencar, "Interpretability of fuzzy systems: Current research trends and prospects," in *Springer Handbook of Computational Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 219–237.

[12] B. Kosko, "Fuzzy cognitive maps," *Int. J. Man. Mach. Stud.*, vol. 24, no. 1, pp. 65–75, Jan. 1986.

[13] K.-S. Lee, S.-H. Kim, M. Sakawa, M. Inuiguchi, and K. Kato, "Process fault diagnosis by using fuzzy cognitive map," *Trans. Soc. Instrum. Control Eng.*, vol. 33, no. 12, pp. 1155–1163, 1997.

[14] C. D. Stylios and P. P. Groumpos, "Fuzzy cognitive map model for supervisory manufacture systems," in *Intelligent Systems for Manufacturing*, ser. IFIP advances in information and communication technology. Boston, MA: Springer US, 1998, pp. 137–146.

[15] M. Tirovolas and C. Stylios, "Introducing fuzzy cognitive map for predicting engine's health status," *IFAC-PapersOnLine*, vol. 55, no. 2, pp. 246–251, 2022.

[16] E. I. Papageorgiou and C. D. Stylios, "Fuzzy cognitive maps," in *Handbook of Granular Computing*. Chichester, UK: John Wiley & Sons, Ltd, 2008, pp. 755–774.

[17] C. D. Stylios and P. P. Groumpos, "Modeling complex systems using fuzzy cognitive maps," *IEEE Trans. Syst. Man Cybern. A Syst. Hum.*, vol. 34, no. 1, pp. 155–162, Jan. 2004.

[18] E. I. Papageorgiou, "Learning algorithms for fuzzy cognitive maps—a review study," *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, vol. 42, no. 2, pp. 150–163, Mar. 2012.

[19] D. Czerwinski, M. Czerwinska, P. Karczmarek, and A. Kiersztyn, "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, Jul. 2021.

[20] G. Nápoles, A. Jastrzębska, C. Mosquera, K. Vanhoof, and W. Homenda, "Deterministic learning of hybrid fuzzy cognitive maps and network reduction approaches," *Neural Netw.*, vol. 124, pp. 258–268, Apr. 2020.

[21] "Causality for Machine Learning," https://ff13.fastforwardlabs.com.

[22] Z. Wang and A. Culotta, "Robustness to spurious correlations in text classification via automatically generated counterfactuals," *Proc. Conf. AAAI Artif. Intell.*, vol. 35, no. 16, pp. 14 024–14 031, May 2021.

[23] G. Nápoles, J. L. Salmeron, W. Froelich, R. Falcon, M. Leon Espinosa, F. Vanhoenshoven, R. Bello, and K. Vanhoof, "Fuzzy cognitive modeling: Theoretical and practical considerations," in *Intelligent Decision Technologies 2019*, ser. Smart innovation, systems and technologies. Singapore: Springer Singapore, 2020, pp. 77–87.

[24] A. Yosef, E. Shnaider, M. Schneider, and A. Rothstein, "Relative influences and the reliability of weights in fuzzy cognitive maps," *Fuzzy Sets And Systems*, vol. 449, pp. 100–119, Nov. 2022.

[25] O. Orang, P. C. d. L. e. Silva, and F. G. Guimarães, "Time series forecasting using fuzzy cognitive maps: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2201.02297

[26] B. Kosko, "Hidden patterns in combined and adaptive knowledge networks," *Int. J. Approx. Reason.*, vol. 2, no. 4, pp. 377–393, Oct. 1988.

[27] X. S. Liang, "Unraveling the cause-effect relation between time series," *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 90, no. 5-1, p. 052150, Nov. 2014.

[28] ——, "Information flow within stochastic dynamical systems," *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 78, no. 3 Pt 1, p. 031113, Sep. 2008.

[29] ——, "Information flow and causality as rigorous notions ab initio," *Phys. Rev. E.*, vol. 94, no. 5, Nov. 2016.

[30] ——, "Normalized multivariate time series causality analysis and causal graph reconstruction," *Entropy (Basel)*, vol. 23, no. 6, p. 679, May 2021.

[31] 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, p. 359, Sep. 2016.

[32] G. A. Papakostas, Y. S. Boutalis, D. E. Koulouriotis, and B. G. Mertzios, "Fuzzy cognitive maps for pattern recognition applications," *Intern. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 08, pp. 1461–1486, Dec. 2008.

[33] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97, no. 1. Citeseer, 1997, p. 179.

[34] G. A. Papakostas, D. E. Koulouriotis, A. S. Polydoros, and V. D. Tourassis, "Towards hebbian learning of fuzzy cognitive maps in pattern classification problems," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10 620–10 629, Sep. 2012.

[35] V. Mpelogianni and P. P. Groumpos, "Re-approaching fuzzy cognitive maps to increase the knowledge of a system," *AI Soc.*, vol. 33, no. 2, pp. 175–188, May 2018.

[36] 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.

[37] G. Napoles, Y. Salgueiro, I. Grau, and M. L. Espinosa, "Recurrence-aware long-term cognitive network for explainable pattern classification," *IEEE Trans. Cybern.*, vol. PP, Apr. 2022.

[38] R. V. Mises and H. Pollaczek-Geiringer, "Praktische verfahren der gleichungsauflösung ." *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, no. 2, pp. 152–164, 1929. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/zamm.19290090206

[39] J. J. Leader, "Limit orbits of a power iteration for dominant eigenvalue problems," *Applied mathematics letters*, vol. 4, no. 4, pp. 41–44, 1991.

[40] ——, "Power iterations and the dominant eigenvalue problem," NAVAL POSTGRADUATE SCHOOL MONTEREY CA, Tech. Rep., 1992.

[41] E. I. Papageorgiou, K. E. Parsopoulos, C. S. Stylios, P. P. Groumpos, and M. N. Vrahatis, "Fuzzy cognitive maps learning using particle swarm optimization," *J. Intell. Inf. Syst.*, vol. 25, no. 1, pp. 95–121, Jul. 2005.

[42] G. Nápoles, E. Papageorgiou, R. Bello, and K. Vanhoof, "On the convergence of sigmoid fuzzy cognitive maps," *Inf. Sci. (Ny)*, vol. 349-350, pp. 154–171, Jul. 2016.

[43] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mech. Syst. Signal Process.*, vol. 163, no. 108105, p. 108105, Jan. 2022.

[44] P. Barbrook-Johnson and A. S. Penn, "Fuzzy cognitive mapping," in *Systems Mapping*. Cham: Springer International Publishing, 2022, pp. 79–95.

[45] L. S. Soler, K. Kok, G. Camara, and A. Veldkamp, "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.*, vol. 7, no. 2, pp. 149–175, Jun. 2012.

[46] F. Liu, Y. Peng, Z. Chen, and Y. Shi, "Modeling of characteristics on artificial intelligence IQ test: A fuzzy cognitive map-based dynamic scenario analysis," *Int. J. Comput. Commun. Control*, vol. 14, no. 6, p. 653, Feb. 2020.

[47] S. Matzka, "Explainable artificial intelligence for predictive maintenance applications," in *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*. IEEE, Sep. 2020.

[48] S. Sridhar and S. Sanagavarapu, "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, Sep. 2021.

[49] W. Froelich, "Towards improving the efficiency of the fuzzy cognitive map classifier," *Neurocomputing*, vol. 232, pp. 83–93, Apr. 2017.

[50] P. Szwed, "Classification and feature transformation with fuzzy cognitive maps," *Appl. Soft Comput.*, vol. 105, no. 107271, p. 107271, Jul. 2021.

[51] G. Napoles, M. Leon, I. Grau, and K. Vanhoof, "Fuzzy cognitive maps tool for scenario analysis and pattern classification," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, Nov. 2017.

[52] A. Erdem, "Aerdem4/lofo-importance," Jan. 2023.