Leveraging Information Flow-Based Fuzzy Cognitive Maps for Interpretable Fault Diagnosis in Industrial Robotics

Marios Tyrovolas[1,2(✉)], Chrysostomos Stylios[1,2], Khurshid Aliev[3], and Dario Antonelli[3]

[1] Department of Informatics and Telecommunications, University of Ioannina, 47 150 Arta, Greece
tirovolas@kic.uoi.gr, stylios@isi.gr
[2] Industrial Systems Institute (ISI), Athena RC, 265 04 Patras, Greece
[3] Department of Management and Production Engineering, Politecnico di Torino, 10129 Turin, Italy
{khurshid.aliev,dario.antonelli}@polito.it

Abstract. In Industry 4.0, Artificial Intelligence (AI) is revolutionizing manufacturing with innovations such as automated fault detection in robotics. However, many current AI models are opaque, obscuring decision-making processes and reducing worker trust. Additionally, these models rely on correlative learning, making them susceptible to adopting spurious correlations that affect their reliability and generalizability. This paper presents the use of Information Flow-Based Fuzzy Cognitive Maps (IF-FCMs) for fault detection and diagnosis in industrial robotics, aiming to overcome these challenges. IF-FCMs, building on FCMs known for their intuitive causal structure and interpretability, integrate Liang-Kleeman Information Flow analysis for rigorous data-driven causality analysis. This approach effectively distinguishes authentic causal links from spurious correlations, enhancing the predictive and explanatory power of FCMs. Moving beyond previous studies that used synthetic data, which often lack real-world complexity and variability, this study employs actual industrial robot data. Numerical simulations demonstrate that IF-FCMs outperform traditional FCMs in terms of both diagnostic accuracy and interpretability, underscoring their potential for tackling manufacturing challenges.

Keywords: Fuzzy Cognitive Maps · eXplainable AI · Information Flow · Industrial Robotics · Fault detection

1 Introduction

Industry 4.0 has revolutionized manufacturing by integrating Artificial Intelligence (AI) and the Industrial Internet of Things (IIoT), which enhances both operational efficiency and safety [1]. For example, AI can be crucial in monitoring industrial robots to promptly

Published by Springer Nature Switzerland AG 2024
L. M. Camarinha-Matos and F. Ferrada (Eds.): DoCEIS 2024, IFIP AICT 716, pp. 98–110, 2024.
https://doi.org/10.1007/978-3-031-63851-0_6

detect any faults, ensuring their continuous and efficient operation. These technological advancements allow industries to quickly identify and resolve issues, significantly reducing production downtime, financial losses, and safety risks [2]. However, the success of AI in such applications relies not only on the ability to accurately detect faults but also on workers' understanding of their root causes. Without a deep understanding of these causes, persistent inefficiencies and errors may occur.

To bridge this gap, explainable AI (XAI) systems are increasingly used to demystify AI decisions and provide transparent explanations of detected faults, helping workers take informed corrective actions [3]. XAI approaches include post-hoc methods that explain "black-box" models and intrinsically interpretable models, which are transparent by design [4]. However, post-hoc methods may inaccurately approximate the behavior of the underlying black-box model, rationalize biased models, or incorrectly assume feature independence, leading to potentially misleading explanations in case of interrelated features [5, 6]. This has prompted a shift towards intrinsically interpretable models and causal XAI, which learns the causal mechanisms of the analyzed system, thus providing unbiased explanations that align with human reasoning and reveal the actual root causes of the model's behavior rather than merely correlations [7].

In this context, Fuzzy Cognitive Maps (FCMs) have emerged as a promising solution for modeling and simulating complex dynamic systems and performing predictive tasks, such as pattern classification and time series forecasting [8, 9]. Described as interpretable recurrent neural networks, FCMs are directed graphs where nodes, referred to as concepts, represent system components, and weighted edges describe the causal relationships between these concepts [9]. FCMs are recognized for their intuitive causal structure and transparent, interpretable feature-based explanations, making them ideal for industrial applications such as predictive maintenance and fault detection [10]. Developed through expert knowledge and/or learning algorithms, FCMs offer unmatched flexibility, adapting to new knowledge or compensating for the lack of historical data using expert assessments [11].

While FCMs are valuable for their interpretability and causal reasoning capabilities, their effectiveness can be compromised by subjective expert opinions or the limitations of data-driven correlative learning algorithms that do not differentiate between causal and spurious correlations [12, 13]. To overcome these challenges, Information Flow-Based Fuzzy Cognitive Maps (IF-FCMs) have been introduced, which incorporate Liang-Kleeman Information Flow (L-K IF) analysis to identify authentic causal relationships within observational data, significantly enhancing model predictive and explanatory power by eliminating spurious correlations [14]. The effectiveness of IF-FCMs was first shown in an XAI model used for detecting and diagnosing faults in industrial systems, employing a synthetic dataset. However, because synthetic data may not fully represent the complexity and variability of real-world scenarios, it is crucial to further validate IF-FCMs with real-world data to confirm their practical utility and robustness.

This study investigates the application of IF-FCMs in detecting and diagnosing faults in industrial robots using data from an actual industrial robot manipulator. Preliminary results reveal the high diagnostic accuracy and interpretability of IF-FCMs, highlighting their potential in practical scenarios. This research enriches the study of FCMs and

contributes to developing understandable, reliable AI technologies, advancing safer and more efficient manufacturing practices in Industry 4.0.

The remainder of this paper is structured as follows: Sect. 2 delves into the connection between current research and human-centric systems. Section 3 provides an overview of our theoretical framework. Section 4 describes the experimental setup and IF-FCMs application in our case study. Section 5 discusses the experimental results and offers quantitative and qualitative insights into the prediction accuracy and interpretability. Finally, Sect. 6 concludes the study with the main contributions and future research directions.

2 Relationship to Human-Centric Systems

Industry 4.0 enhances efficiency and inspires innovative business models, services, and products. However, this digital transformation tends to prioritize manufacturing process optimization, often neglecting the human dimension [15]. Thus, there is a critical need for a human-centric production approach, where AI and automation foster an industrial environment that prioritizes worker well-being, safety, trust, and human-machine collaboration. Applying IF-FCMs for fault detection and diagnosis in industrial robots represents a significant step towards this goal, embodying the principles of transparency, interpretability, and user-friendliness, which are critical for realizing human-centric systems [16].

IF-FCMs significantly improve the accessibility and understanding of AI technologies for industrial workers by offering clear explanations for AI decisions, fostering trust, and facilitating intuitive human-automation interaction through causality. This approach not only advances FCMs and XAI but also promotes sustainable, resilient industrial environments centered on human needs. This research underscores the value of interpretable AI in linking technological advancements to their practical, user-oriented applications, aiming for a future in which technology supports human needs and ensures a collaborative, safe, and efficient workplace.

3 Theoretical Background

This section outlines the theoretical foundations of our research, including FCMs, Information Flow for rigorous causality analysis in multivariate time-series data, and IF-FCMs.

3.1 Fuzzy Cognitive Maps

Fuzzy Cognitive Maps (FCMs), introduced by Kosko, are a soft computing methodology represented as directed graphs. These graphs consist of n interconnected concepts C_i, $i \in 1,2,\ldots,n$, linked by signed weights $w_{ij} \in [-1,1]$. Positive weights indicate that two concepts change in the same direction, while a negative weight means they change in opposite directions.

Each concept C_i in an FCM has an activation value, $A_i^{(t)} \in [0,1]$ or $[-1,1]$, indicating the activation level of that concept in the t th iteration of the FCM's recurrent reasoning process, computed using the following generalized reasoning rule.

$$A_i^{(t+1)} = \phi f\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} A_j^{(t)} w_{ji}\right) + (1-\phi) A_i^{(0)} \tag{1}$$

where t denotes the iteration step, $f(\cdot)$ is the activation function that normalizes concept's activation values to the allowed interval, parameter $\varphi \in [0,1]$ controls the nonlinearity of the reasoning rule, and $A_i^{(0)}$ is the initial activation value of the i th concept provided by domain experts or extracted from the available data. Because commonly used activation functions, such as bivalent, trivalent, hyperbolic tangent, and sigmoid, can impact the reliability of FCMs' reasoning process, Nápoles *et al.* [17] mitigated these issues by introducing a re-scaled activation function.

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

where $\overline{\mathbf{A}}^{(\mathbf{t})} = \left[\overline{\mathrm{A}}_1^{(\mathrm{t})}, \overline{\mathrm{A}}_2^{(\mathrm{t})}, \ldots, \overline{\mathrm{A}}_{\mathrm{n}}^{(\mathrm{t})}\right]$ is the raw state vector given by $\overline{\mathbf{A}}^{(\mathbf{t})} = \mathbf{A}^{(\mathbf{t})}\mathbf{W}$, with $\mathbf{W} \in \mathbb{R}^{n \times n}$ being the weight matrix, and $\|\cdot\|_2$ denotes the Euclidean norm.

The FCM iterative reasoning process starts with an initial state vector $\mathbf{A}^{(0)} \in \mathbb{R}^{\mathrm{n}}$, provided by domain experts or derived from datasets. The chosen reasoning rule is then applied recurrently to update the concepts' activation values in each iteration. This continues until the FCM converges to a fixed-point attractor or reaches a predefined maximum number of iterations, T, where the FCM exhibits cyclic or chaotic behavior [18].

3.2 Information Flow

To address the challenges and further capabilities of data-driven FCMs, causality inference from historical data has become crucial, particularly with recent advancements in AI. Notably, Liang argued that causality, in the Newtonian sense, is a real physical notion called Information Flow (IF), which can be established from physics' first principles [19]. This insight led to the Liang-Kleeman Information Flow (L-K IF) analysis, a rigorous framework for causality analysis in time-series data, marking a significant shift from traditional qualitative or empirically-based formalisms, such as transfer entropy [20]. The L-K IF framework introduces a method to quantify the causality between state variables in a d-dimensional continuous-time stochastic system, with the IF rate, $T_{j \to i}$, , which measures how the entropy of one variable (X_j) contributes to another's (X_i) marginal entropy per unit time. A nonzero $T_{j \to i}$ signifies causality and its magnitude reflects causality's strength [21]. The analytical expression for $\mathrm{T}_{\mathrm{j} \to \mathrm{i}}$ is given in detail

in [19]. In detail, for d time series $X_1, X_2, \ldots, X_d$, , under the assumption of a linear model with additive, independent noises, the maximum likelihood estimator (MLE) for IF from X_2 to X_1 is [22]:

$$\hat{T}_{2\to 1} = \frac{1}{detC} \cdot \sum_{j=1}^{n} \Delta_{2j} C_{j,d1} \cdot \frac{C_{12}}{C_{11}} \tag{3}$$

where C_{ij} denotes the sample covariance between X_i and X_j, Δ_{ij} are the cofactors of the covariance matrix $C = \left(C_{ij}\right)$, and $C_{i,dj}$ is the covariance between X_i and $\frac{dX_j}{dt}$ calculated using the Euler forward scheme. Given that this formula is an MLE of the analytical expression, evaluating its statistical significance using the Fisher information matrix is crucial for validating the causality.

3.3 Information Flow-Based Fuzzy Cognitive Maps

Building on FCMs and IF's mathematical rigor, physical interpretability, and computational efficiency, IF-FCMs have recently emerged as a crucial advancement, integrating L-K IF analysis with FCMs to rule out spurious correlations and thereby enhance both predictive and explanatory power. This subsection details the development, structure, and interpretability of IF-FCMs.

The construction of an IF-FCM starts by mapping problem variables to distinct concepts, establishing a single-output model architecture. Subsequently, the model undergoes a two-phase learning process. In *Training Phase 1*, the L-K IF analysis identifies statistically significant IF rates from the data, uncovering causal relationships among variables. These IF rates are essential for guiding the next phase. *Training Phase 2* focuses on learning the IF-FCM by optimizing weights and parameters associated with the activation function and reasoning rule to fit the model to the observed data. Unlike traditional methods, the causal relationships identified are used as constraints in the learning algorithm. This algorithm adjusts only the weights with significant IF rates and sets others to zero a priori, ensuring the FCM accurately represents actual system interactions.

In *Training Phase 2* of IF-FCM learning, population-based metaheuristic algorithms, such as Particle Swarm Optimization (PSO), are commonly used [23]. The foundational research on IF-FCM initially employed PSO, but due to its limitations with high-dimensional problems and multiple local optima, this study uses Social Learning PSO (SL-PSO) [25]. SL-PSO enhances PSO by incorporating a social learning mechanism that allows particles (solutions) to learn from top-performing peers, called demonstrators. This method encourages particles to update their position based on multiple peers, thus enhancing solution space exploration and aiding in escaping local optima. Additionally, SL-PSO uses a dimension-dependent parameter control strategy, which adjusts demonstrators' influence based on problem dimensionality, allowing particles to maintain diversity in higher dimensions and balance exploration with exploitation. Comparative experiments have also shown SL-PSO's lower time complexity compared to other PSO variants across various dimensional problems. In the IF-FCM context, SL-PSO optimizes solutions encoded as $x = \left[\varphi, w^{(1)}, w^{(2)}, \ldots, w^{(\mathrm{SIFs})}\right]$, where SIFs represents

the number of statistically significant IF rates, aiming to minimize the following objective function

$$\varepsilon(x) = \alpha_1 \underbrace{\frac{1}{K}\sum_{k=1}^{K}\left|Y_k - A_{nk}^{(l)}\right|}_{G(x)} + (1-\alpha_1)\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{l}\frac{2\omega_t(A_{ik}^{(t)} - A_{ik}^{(t-1)})^2}{K\,n\,(T-1)}}_{H(x)}, \quad (4)$$

where $G(\cdot)$ reflects the mean absolute prediction error across all instances and $H(\cdot)$ measures the aggregate dissimilarity between successive state vectors, indicating the model's convergence. K is the sample size, n the number of FCM concepts, Y_k the expected output for each instance, and $\omega_t = \frac{t}{T}$ weights the importance of iterations, with later iterations weighted more to ensure convergence and allowing initial flexibility in the model's dynamics. The coefficient $\alpha_1 \in [0,1]$ balances accuracy against stability in the model's performance.

Once developed, an IF-FCM can be used for specific predictive tasks, such as binary classification to detect faults in industrial robots. In classification tasks, each k^{th} data instance starts as an initial state vector $A_k^{(0)} = \left[A_{1k}^{(0)}, A_{2k}^{(0)}, \ldots, A_{nk}^{(0)} = 0\right]$ in the IF-FCM. This vector is then recurrently updated through the reasoning rule until the reasoning process concludes. The predicted class label is determined by the final activation value, $A_{nk}^{(l)}$, , of the output concept. The predefined activation interval ([0,1] or [−1,1]) is divided into m partitions corresponding to the possible class labels. The final activation value $A_{nk}^{(l)}$ then assigns a class label based on which partition it falls into. In binary classification, the model uses a single threshold to distinguish between classes, optimizing the balance between true-positive and true-negative rates.

Along with the generated predictions, the IF-FCM model can provide global or local explanations through its inherent interpretability. On the one hand, global interpretability reveals the significant concepts across all data, providing a holistic view of the model's logic. In contrast, local interpretability, which is the focus of our study, explores the rationale behind predictions for specific data instances. This characteristic is particularly valuable in fault detection, as it identifies which input features (i.e., concepts) are responsible for the model's predictions, aiding in diagnosing faults. To locally explain a given data instance, IF-FCM used the final activation values of concepts after the reasoning process. Concepts with higher absolute final activation values are interpreted as more influential in the decision-making process for that specific instance.

4 IF-FCMs in Robotic Fault Detection and Diagnosis

In this section, we outline the experimental framework and dataset employed for the numerical simulations, followed by a detailed discussion of the application of IF-FCMs in our case study. This setup was designed to rigorously test the efficacy of the IF-FCM model in a real-world environment, providing insights into its predictive accuracy and interpretability.

4.1 Industrial Robot Manipulator Test Bench Setup

In our study, we investigated the performance of IF-FCMs in detecting and diagnosing operational issues in collaborative robots (cobots), which are increasingly utilized in manufacturing because of their ability to work safely alongside human operators. Cobots provide advantages, including lightweight and flexible design, ease of programming, improved productivity, and reduced injury risks, all in compliance with the ISO TS 15066 standards. However, cobots can encounter errors such as grip loss or unintended protective stops, often triggered by their force sensors during unexpected contact or owing to limitations in power and force leading to object slippage.

In this context, we utilized a UR3 cobot, which features six degrees of freedom, a 500mm reach radius, and a repeatability of $\pm$ 0.1 mm, and includes safety mechanisms like immediate stopping upon unexpected contact and limitations on power and force for safety. Our experimental setup involved a cobot performing a pick-and-place task between two points (A and B) under varying conditions of three static process hyper-parameters: workload (1, 2, and 3 kg), movement speed (60%, 80%, and 100% of the maximum speed of 210 rad/s), and gripping force (80, 100, and 120N, based on the OnRobot RG6 gripper's capacity). Each test scenario was replicated 25 times to examine the impact of these parameters on the occurrence of 'Grip Loss' and 'Protective Stop' failures. The experimental setup, including predefined waypoints for consistent gripper movement and controlled arm speed, is depicted in Fig. 1, focusing on these two primary failure modes for an in-depth analysis.

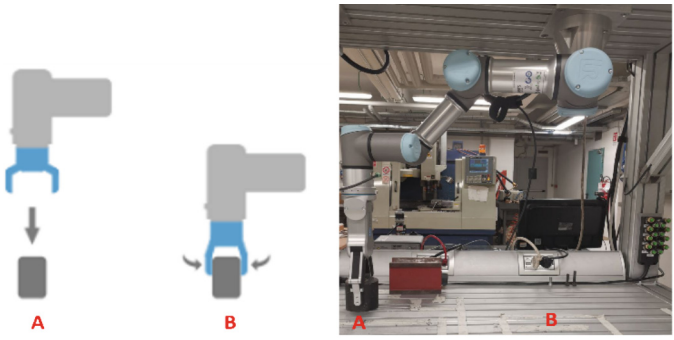

Fig. 1. The cobot test bench setup with a workpiece, highlighting Points A (pickup) and B (placement).

4.2 Dataset Description

In this study, we developed the "UR3 CobotOps" dataset, a comprehensive, multidimensional time-series dataset compiled from real-time operational data of the UR3 cobot using both MODBUS and RTDE protocols [25]. The RTDE interface is critical in this setup, providing seamless synchronization between external applications and the cobot's

controller via TCP/IP without affecting real-time performance. Operating at 125 Hz on port 30004, the RTDE supports configurable synchronization of the robot and joint status, tool, safety status, analog, and digital I/Os. This dataset captures several operational parameters of the UR3 cobot: electrical currents at its joints (Current $J_0 - J_5$), joint temperatures (Temperature $J_0 - J_5$), joint speeds (Speed $J_0 - J_5$), gripper current (Tool Current), number of operation cycles (Cycle), and recorded faults (Protective Stop and Grip Lost), with faults labeled as "False" (no fault) or "True" (fault present) at each time step. These parameters are listed in Table 1.

Table 1. UR3 CobotOps dataset variables.

Variable Name	Description	Units	Type	Range of values (Mix-Max)
Current $J_0 - J_5$	Joint current	A	Float	[–6.24, 6.80]
Temperature $J_0 - J_5$	Joint temperature	°C	Float	[27.81, 45.37]
Speed $J_0 - J_5$	Joint speed	°/s	Float	[–1.62, 2.67]
Tool Current	Gripper current	A	Float	[0.02, 0.60]
Cycle	Operation cycles count	-	Integer	{1, 2, 3,…, 264}
Protective Stop	Emergency halt status	-	Boolean	{"False", "True"}
Grip Lost	Actual grip loss status	-	Boolean	{"False", "True"}

4.3 IF-FCM Deployment in Case Study

In this subsection, we discuss the application of the IF-FCM model in our case study, focusing on data preprocessing and model development. Initially, we refined the raw data by removing non-informative predictors such as "Timestamp" and "Cycle," eliminating missing values, and converting Boolean variables to binary integers. Subsequently, we labeled the dataset to indicate system health, creating a binary target variable "System Failure", where "0" indicates healthy and "1" indicates faulty conditions for samples exhibiting one of the two examined faults. This process produced a binary classification dataset to train the IF-FCM for fault detection. Most importantly, owing to its inherent interpretability, IF-FCM also provides local feature-based explanations of the most influential input features for its decisions, enabling fault diagnosis.

In fault detection and diagnosis applications, data are often unbalanced, with fewer samples for fault classes than for normal ones. To address this, we employed the SMOTE-ENN technique [26] and Stratified K-Fold Cross-Validation [27] with ten splits, enhancing model robustness and reducing overfitting risks. However, caution is required when applying these techniques to time-series data. SMOTE-ENN might create synthetic samples that overlook dependencies, and K-fold cross-validation could disrupt the temporal sequence. Nonetheless, these methods proved effective in our study, which focused on classifying sensor data patterns rather than their temporal sequences.

In the model development (Sect. 3.3), the L-K IF analysis revealed 210 causal relationships among the dataset variables, which were subsequently used as constraints in the SL-PSO algorithm. Regarding the implementation of the SL-PSO, we used the MATLAB version from GitHub[1], enhancing it with MATLAB's "parfor" for parallel particle evaluation. Furthermore, we adopted the default parameter values for SL-PSO, which have been shown to perform robustly across various problem scales [24]. Finally, based on previous studies, we set the α_1 parameter of the objective function to 0.8 [28].

5 Results and Discussion

This section evaluates the efficacy of the proposed IF-FCM model for fault detection and diagnosis in the UR3 cobot, leveraging its inherent interpretability. We compare the proposed model's performance against a traditional FCM, which differs from the IF-FCM in that it assumes all weights are present during training—a common practice in existing research for constructing FCMs automatically in the absence of detailed knowledge about the true causal model structure. Consequently, this model is prone to assimilate spurious correlations from the data. This comparative analysis aims to demonstrate the benefits of leveraging L-K IF analysis to enhance the accuracy and interpretability of FCMs.

We assessed the predictive capabilities of both models using various metrics, including accuracy, area under the curve (AUC), kappa statistic, precision, recall, and F1-Score (Table 2). Our analysis shows that the IF-FCM model marginally outperforms the traditional FCM in this case study. Notably, the IF-FCM achieves a slightly higher accuracy, enhancing its ability to correctly identify faults in industrial robots. It also achieves a superior AUC, indicating more effective discrimination between fault and no-fault conditions. Moreover, the IF-FCM exhibits increased precision, resulting in fewer false alarms, which is essential for avoiding unnecessary, costly interventions in manufacturing processes. Finally, although both models show similar kappa statistics and recall rates, the overall metrics suggest that the IF-FCM is a more robust option for this application.

Table 2. Prediction Performance Metrics of IF-FCM vs. FCM Across Ten Folds

Model	Accuracy	AUC	Kappa	Precision	Recall	F1-Score
IF-FCM	**88.84%**	**94.61%**	**0.7767**	**88.95%**	88.93%	**88.84%**
FCM	88.43%	92.26%	0.7686	86.63%	**91.04%**	88.74%

To assess the interpretability of the two models, we thoroughly evaluated the consistency between ground truth facts and the local feature-based explanations each model provided for their true-positive predictions of 'Protective Stop' and 'Grip Lost' faults. The process began by identifying the data instances in which the models accurately predicted each fault type. Subsequently, we computed each feature's average local importance score across these instances and cross-validation folds. As shown in Table 3, for the "Protective Stop" fault, the IF-FCM model identified Speed J_4, Current.

[1] https://github.com/ranchengcn/SL-PSO_Matlab.

J_2, and Current J_1 as key predictors, while the traditional FCM model emphasized Current J_4, Temperature J_2, and Speed J_2. For "Grip Lost" faults, the IF-FCM prioritized Current J_2, Current J_1, and Speed J_4, contrasting with the traditional FCM, highlighting Speed J_2, Speed J_5, and Speed J_0.

To validate the relevance of these features, we employed *Analysis of Variance* (ANOVA), a model-independent statistical technique, to initially assess the impact of experimental hyperparameters (speed, workload, and gripping force) on the faults [29]. Specifically, ANOVA indicated that 'workload' was the most significant factor for 'Protective Stop' faults (F-value = 7.78, p-value = 0.0032), supporting the IF-FCM findings and field experience, which suggests that a high workload causes out-of-scale values of joint current or wrist force and sudden speed changes, potentially triggering protective stops (Fig. 2a). However, the traditional FCM's focus on Temperature J_2 did not find support in the ANOVA results or field observations, as temperature is not typically a contributing factor to such faults.

Regarding 'Grip Lost', ANOVA highlighted 'gripping force' as the critical hyperparameter (F-value = 5.38, p-value = 0.0135), which aligns with field observations that attribute grip loss to insufficient gripping force, excessive weight, or poor contact (Fig. 2b). These findings also support IF-FCM's focus on Current J_2, Current J_1 and Speed J_4, explaining that high weight may lead to exceeded current in the joints, while sudden rotational movements by the wrist might destabilize the gripped object (i.e., change the center of gravity of the piece concerning the tool center point of the gripper), thus causing detachment when the gripping force is insufficient.

Further statistical analyses were conducted using ANOVA to determine the features significantly associated with each fault type. In this regard, the dataset was divided into subsets tailored to each fault scenario. For "Protective Stop" faults, we analyzed data under normal conditions and those exhibiting protective stops, while for "Grip Lost" faults, we isolated corresponding instances. These focused subsets helped us compute the average statistical significance of each feature across the cross-validation folds.

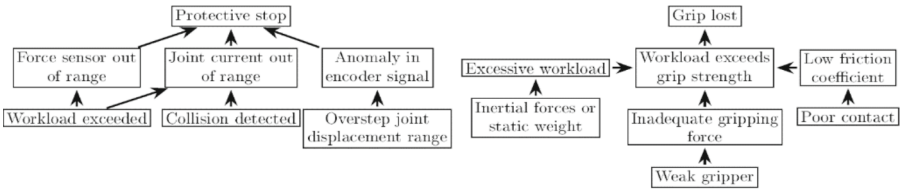

Fig. 2. Cognitive Maps of possible causes for the two examined types of faults.

The results detailed in Table 3 from this ANOVA analysis confirm the relevance of Current J_2 and Current J_1 in triggering protective stops, both identified by IF-FCM as among the three most critical features. In contrast, the traditional FCM model only recognized the Current J_4 from the top five features. For 'Grip Lost' faults, ANOVA identified Current J_2, Speed J_5, and Speed J_4 as essential factors. The IF-FCM model corroborates many of these findings by identifying two of the three critical features

Table 3. Comparison of Feature-Based Explanations from ANOVA, IF-FCM, and FCM

Fault Type	ANOVA (Ground Truth Features)	IF-FCM	FCM
Protective Stop	– Current J_2 (0.361) – Current J_3 (0.311) – Current J_1 (0.244) – Tool Current (0.069) – Current J_4 (0.064)	– Speed J_4 – **Current $\mathbf{J}_2$** – **Current $\mathbf{J}_1$** – Speed J_5 – Speed J_2	– Speed J_2 – **Current $\mathbf{J}_4$** – Temperature J_2 – Current J_0 – Speed J_1
Grip Lost	– Current J_2 (0.995) – Speed J_5 (0.646) – Speed J_4 (0.569) – Current J_1 (0.543) – Speed J_3 (0.470)	– **Current $\mathbf{J}_2$** – **Current $\mathbf{J}_1$** – **Speed $\mathbf{J}_4$** – **Speed $\mathbf{J}_5$** – **Speed $\mathbf{J}_3$**	– Speed J_2 – **Speed $\mathbf{J}_5$** – Speed J_0 – **Current $\mathbf{J}_2$** – Current J_0
	#Common Features	**2/5**	1/2

specified by the ANOVA: Current J_2 and Speed J_4. However, the traditional FCM model only recognized Speed J_5. Moreover, the IF-FCM identified all the top five features marked by ANOVA as significant, whereas the traditional FCM identified only two features. Overall, this analysis highlights the enhanced explanatory power of the IF-FCM and its potential to facilitate manufacturing system enhancements by offering actionable insights, such as the adjustment of Speed J_4 to mitigate grip loss.

6 Conclusions

This paper presents a novel method for detecting and diagnosing faults in industrial robots using IF-FCMs, a recent advancement in causal XAI. Our tests on the UR3 collaborative robot show that IF-FCMs have superior predictive capabilities compared with traditional FCMs, resulting in more accurate fault identification. The method also offers enhanced interpretability, particularly when identifying critical features affecting robot faults. Our study reveals that IF-FCMs offer more consistent and reliable explanations for specific faults like 'Protective Stop' and 'Grip Lost,' closely aligning with field data and ANOVA analyses. This accuracy is attributed to the IF-FCMs' capacity to discern authentic causal relationships within the system. Future work will focus on developing IF-FCMs to not only analyze faults, but also recommend specific corrective actions. Additionally, we intend to expand the application of IF-FCMs to other industrial sectors, including energy management, predictive maintenance, and quality assurance. Overall, this study highlights the potential of IF-FCMs to enhance fault detection and diagnosis in industrial robotics, laying the groundwork for further research and applications in diverse manufacturing environments.

Acknowledgments. This research is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU under the call "Flagship actions in interdisciplinary scientific fields with a special focus on the productive fabric", project name "Greece4.0 - Network of Excellence for developing, disseminating and implementing digital transformation technologies in Greek Industry" (project code: TAEDR-0535864).

References

1. Jagatheesaperumal, S.K., Rahouti, M., Ahmad, K., Al-Fuqaha, A., Guizani, M.: The duo of artificial intelligence and big data for industry 4.0: applications, techniques, challenges, and future research directions. IEEE Internet Things J. **9**(15), 12861–12885 (2022)
2. Hsu, H.K., Ting, H.Y., Huang, M.B., Huang, H.P.: Intelligent fault detection, diagnosis and health evaluation for industrial robots. Mechanics **27**(1), 70–79 (2021)
3. Das, D., Banerjee, S., Chernova, S.: Explainable AI for robot failures. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. ACM, New York (2021)
4. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (2019)
5. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
6. Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: more accurate approximations to shapley values. Artif. Intell. **298**(103502), 103502 (2021)
7. Liang, X.S., Chen, D., Zhang, R.: Quantitative causality, causality-aided discovery, and causal machine learning. Ocean Land Atmos. Res., October 2023
8. Stylios, C.D., Groumpos, P.P.: Modeling complex systems using fuzzy cognitive maps. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **34**(1), 155–162 (2004)
9. Kosko, B.: Fuzzy cognitive maps. Int. J. Man Mach. Stud. **24**(1), 65–75 (1986)
10. Tirovolas, M., Stylios, C.: Introducing fuzzy cognitive map for predicting engine's health status. IFAC-PapersOnLine **55**(2), 246–251 (2022)
11. Napoles, G., Salgueiro, Y., Grau, I., Espinosa, M.L.: Recurrence-aware long-term cognitive network for explainable pattern classification. IEEE Trans. Cybern. **53**(10), 6083–6094 (2023)
12. Papageorgiou, E.I., Stylios, C.D.: Fuzzy cognitive maps. In: Handbook of Granular Computing, pp. 755–774. John Wiley & Sons, Ltd, Chichester (2008)
13. Wang, Z., Culotta, A.: Robustness to spurious correlations in text classification via automatically generated counterfactuals. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14024–14031 (2021)
14. Tyrovolas, M., Liang, X.S., Stylios, C.: Information flow-based fuzzy cognitive maps with enhanced interpretability. Granul. Comput. **8**(6), 2021–2038 (2023)
15. Mourtzis, D., Angelopoulos, J., Panopoulos, N.: A literature review of the challenges and opportunities of the transition from industry 4.0 to society 5.0. Energies **15**(17), 6276 (2022)
16. Rožanec, J.M., et al.: Human-centric artificial intelligence architecture for industry 5.0 applications. Int. J. Prod. Res. **61**(20), 6847–6872 (2023)
17. Nápoles, G., Grau, I., Concepción, L., Koutsoviti Koumeri, L., Papa, J.P.: Modeling implicit bias with fuzzy cognitive maps. Neurocomputing **481**, 33–45 (2022)
18. Nápoles, G., Rankovi¢, N., Salgueiro, Y.: On the interpretability of fuzzy cognitive maps. Knowl. Based Syst. **281**(111078), 111078 (2023)
19. Liang, X.S.: Information flow and causality as rigorous notions ab initio. Phys. Rev. E. **94**(5), November 2016

20. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**(2), 461–464 (2000)
21. Liang, X.S.: Unraveling the cause-effect relation between time series. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **90**(5-1), 052150 (2014)
22. Liang, X.S.: Normalized multivariate time series causality analysis and causal graph reconstruction. Entropy (Basel) **23**(6), 679 (2021)
23. Papageorgiou, E.I.: Learning algorithms for fuzzy cognitive maps—a review study. IEEE Trans. Syst. Man Cybern. C Appl. Rev. **42**(2), 150–163 (2012)
24. Cheng, R., Jin, Y.: A social learning particle swarm optimization algorithm for scalable optimization. Inf. Sci. (Ny) **291**, 43–60 (2015)
25. Tyrovolas, M., Aliev, K., Antonelli, D., Stylios, C.: UR3 CobotOps. UCI Machine Learning Repository (2024). https://doi.org/10.24432/C5J891
26. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. **6**(1), 20–29 (2004)
27. Purushotham, S., Tripathy, B.K.: Evaluation of classifier models using stratified tenfold cross validation techniques. In: Krishna, P.V., Babu, M.R., Ariwa, E. (eds.) ObCom 2011. CCIS, vol. 270, pp. 680–690. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29216-3_74
28. Nápoles, G., Papageorgiou, E., Bello, R., Vanhoof, K.: Learning and convergence of fuzzy cognitive maps used in pattern recognition. Neural. Process. Lett. **45**(2), 431–444 (2017)
29. Lindman, H.R.: Analysis of Variance in Experimental Design. Springer Texts in Statistics, Springer, New York (1991). 1992 edn.