

13th CIRP Global Web Conference (CIRPe 2025)

Enabling Joint Benchmarking of Automated Root Cause Analysis and Causal Discovery in Manufacturing Using the *causRCA* Dataset

Carl Willy Mehling*, Sven Pieper, Tobias Lüke, Julius Döbelt, Steffen Ihlenfeldt

Fraunhofer Institute for Machine Tools and Forming Technology IWU, Nöthnitzer Str. 44, 01187 Dresden, Germany

* Corresponding author. Tel.: +49 351 4772-2633. E-mail address: carl.willy.mehling@iwu.fraunhofer.de

Abstract

As manufacturing systems become more automated and interconnected, diagnosing faults and identifying their root cause has become increasingly complex for human operators. Data-driven methods can help prevent costly downtime by leveraging Causal Discovery (CD) to map out how different machine components affect each other, while automated Root Cause Analysis (RCA) tracks down fault origins. However, progress in developing RCA and CD methods is hindered by the lack of real-world datasets that support their joint benchmarking in realistic manufacturing environments. We introduce the *causRCA* manufacturing dataset to fill this gap. The dataset comprises 170 data recordings from normal operation of a CNC vertical lathe and 100 simulated fault data recordings generated through a hardware-in-the-loop setup that combines a digital twin of the lathe with a physical controller. The dataset includes an expert-validated causal graph connecting the 92 included variables and alarms, serving as ground truth for evaluating both CD and causal RCA methods. We illustrate the versatility of *causRCA* through exemplary benchmarks that compare supervised RCA methods, unsupervised RCA methods, and CD algorithms on the dataset. Furthermore, we demonstrate its potential for answering research questions regarding causal RCA methods by analyzing how the quality of learned causal graphs affects RCA performance. All data, code, and documentation are publicly available to accelerate research in CD and automated RCA.

© 2025 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

Peer review under the responsibility of the scientific committee of the CIRPe 2025

Keywords: Root Cause Analysis, Causal Discovery, Fault Diagnosis, Benchmark, Manufacturing, Artificial Intelligence, Digital Twin, Hardware-in-the-Loop

1. Introduction

Modern manufacturing systems are increasingly automated and complex, making the identification of failure causes challenging for human operators and costly for manufacturers [1, 2]. According to an industrial report, unplanned downtime costs manufacturers an average of 11 percent of annual revenues, amounting to 1.4 trillion USD for the 500 biggest manufacturers [3]. Preventing repeated failures and minimizing downtime is therefore a top priority, and the automation of root cause analysis (RCA) plays a central role.

RCA refers to the process of systematically identifying the actual underlying cause of a problem in order to prevent its recurrence [1]. However, many current data-driven approaches

focus merely on associations, neglecting the need for underlying causal assumptions to draw a causal conclusion [4].

Causal graphs (CG) provide a human- and machine-readable way to make cause-and-effect assumptions visible and accessible to RCA and can be inferred from machine data using causal discovery (CD) algorithms. Recent surveys summarize advances in automated RCA and industrial causal discovery, covering methods that range from data-driven to hybrid approaches, and discussing practical challenges in applying them to manufacturing settings [1, 2, 5, 6].

Although the combination of CD and RCA is actively researched, progress is hindered by the lack of realistic datasets supporting their joint evaluation. In this work, we provide such a dataset and illustrate its use with example benchmarks,

focusing on reactive fault diagnosis that helps operators isolate causes and make decisions after faults have occurred [7].

1.1. Related Datasets

Available datasets often focus on either CD or RCA and have limited applicability to manufacturing scenarios (Table 1). For the evaluation of CD algorithms, a true causal graph is needed to compare the learned graph with the true causal structure. These validated true causal graphs are missing in common datasets used for evaluating fault diagnosis and RCA algorithms [8–10]. For real-world data, this true causal graph often has to be reconstructed from expert knowledge, making comprehensive CD benchmarks on real-world data scarce.

Recent contributions (causalAssembly [11], ACCP [12], CausalRivers [13]) provide validated real-world causal structures but are not usable for RCA evaluation. RCA algorithms require labeled root cause information to compare predicted causes with the true causes of a fault scenario. These true causes are often only available for synthetic datasets, which miss the real manufacturing complexity and noise characteristics [14]. Some datasets address real-world manufacturing fault diagnosis [15] but lack ground truth graphs. Others target non-manufacturing domains (IT systems [8, 16], power systems [17]) with different characteristics.

To our knowledge, few manufacturing time-series datasets are publicly available that combine real-world manufacturing data with a known causal structure and labeled causes, allowing for joint CD and RCA benchmarking (Table 1).

Table 1. Overview of Causal Discovery and RCA datasets

Dataset	Ref	True Graph	Fault Diag.	Real World	True Cause	Variables
LEMMA-RCA	[8]	X	✓/X	X	✓	51-216
HVAC systems	[9]	X	✓	✓/X	X	24-56
SWaT	[10]	X	✓/X	✓	✓	51
PyRCA	[14]	✓	X	X	✓	dynamic
causal-Assembly	[11]	✓/X	X	✓/X	X	98
ACCP ESS	[12]	✓	X	✓	X	233
CausalRivers	[13]	✓	X	✓	X	44-666
ADAPT	[17]	✓	✓	X	✓	128
CIPCaD	[15]	✓	✓	✓	X	17
RCEval	[16]	✓	✓	✓	✓	49-376
causRCA	ours	✓	✓	✓	✓	11-92

1.2. Contributions and Structure of this Work

This work introduces causRCA, a comprehensive manufacturing dataset addressing the lack of resources that jointly support CD and RCA benchmarking. It features real-world CNC vertical lathe operational data, fault data with known causes from a realistic hardware-in-the-loop (HIL) simulation, and an expert-curated causal graph. The dataset

enables the individual evaluation of CD and RCA, as well as their interaction, allowing research into how the quality of causal graphs influences diagnostic performance. We provide example benchmarks for CD, supervised RCA, and unsupervised RCA, and release all data and code under an open-source license to foster reuse and reproducibility.

The rest of the work is structured as follows: we introduce the causRCA dataset and its generation, present benchmark results for CD and RCA, evaluate the impact of causal graph quality on RCA performance, and conclude with a discussion.

2. causRCA - Real-World Dataset for Causal Discovery and Root Cause Analysis in Machinery

This section first describes causRCA's data generation methodology, followed by an overview of the dataset's hierarchical structure and key characteristics.

2.1. Dataset Generation using Hardware-in-the-loop setup

To generate training data with known fault causes, we developed a HIL simulation combining a digital twin of the machine's behavior with an industrial PLC/NC controller. Fig. 1 illustrates this dual-path data generation architecture.

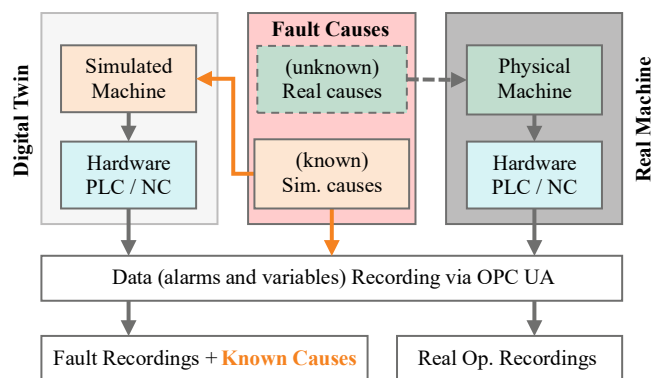


Fig. 1. The data creation setup with Digital Twin and Real Machine, both having a physical PLC/NC to create realistic data recordings.

The left path depicts the digital twin: a virtual machine model in ISG-virtuos controlled by physical Sinumerik 840Dsl PLC/NC hardware. The HIL setup preserves authentic control timing and communication protocols, allowing controlled fault injection via the known cause.

The right path captures machine data from the physical vertical lathe during production. In this case, the actual cause remains unknown, but operational patterns become visible.

In the HIL setup, the Virtuos simulation runs on a workstation with a dedicated fieldbus card connected to the PLC/NC bus. The controller receives all I/Os from Virtuos as simulated values, making the virtual plant indistinguishable from the real machine. This coupling ensures the PLC/NC operates with its native cycle time and processes simulated feedback with the same determinism as in production. [18]

Both the digital twin and real machine use the OPC UA server embedded in the hardware PLC/NC, which provides the identical variables in the identical OPC UA information model.

A historian subscribes to all OPC UA variables and stores them alongside the information model. In the HIL setup, additional writable fields are included to record which variables were manipulated and which fault cause was simulated. These entries are historized together with the measured process values and alarms, enabling direct comparison of manipulations and controller responses.

Injection of simulated faults is realized through the Virtuoso.dll interface, which is addressed by a Python-based orchestration tool. The procedure follows a repeatable five-step cycle: (i) the virtual machine is initialized to a defined state, (ii) the CNC program is executed on the controller, (iii) simulated faults are injected by forcing sensor readings, (iv) manipulated variables, associated causes, and controller responses are stored via the OPC UA server, and (v) the simulation is stopped, reset, and restarted for the next run. This cycle is repeated for multiple runs per scenario with slight alterations of the occurrence time of the manipulations.

By historizing manipulated variables, causes, measured values, and alarms, we obtain a dataset that directly links injected faults to the resulting outputs of the controller. The digital twin records include ground truth annotations specifying the manipulated variables and exact time of manipulation, as well as expert-derived cause labels, which provide essential information for training and evaluating RCA methods.

2.2. Overview of the causRCA dataset

The causRCA dataset comprises 270 records from a CNC vertical lathe: 170 from real-world production and 100 from HIL fault simulations. The records contain value changes from up to 92 variables and alarms. For each variable that changes during the up to 10-minute length of a recording, all value changes, as well as the initial value at the start of the recording, are captured using OPC UA subscriptions onto the NC/PLC controller's OPC UA Server. These value changes are accompanied by alarms that are triggered by the controller during the recording and captured through OPC UA Events.

The 92 variables and alarms are also structured as nodes within the full expert causal graph. The expert-graph features a hierarchical structure of its subsystems, with three of its subsystems (coolant, hydraulics, probe) used for creating fault scenarios with the HIL simulation.

The three subsystems plus 49 additional variables form the complete causal graph, enabling benchmarking at multiple complexity levels: from focused 11-node subsystem analysis to comprehensive system-wide CD or RCA (Fig. 2, Table 2).

Full Causal Graph (92 Nodes, 104 Edges)

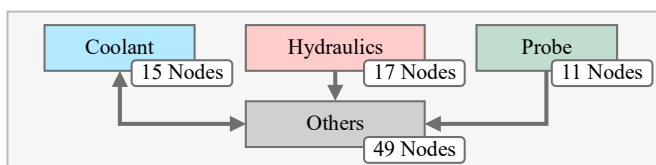


Fig. 2. Causal graph of the dataset (92 nodes, 104 edges) grouped by machine components: Coolant (15 nodes), Hydraulics (17 nodes), Probe (11 nodes), and Others (49 nodes), illustrating both dataset composition and interdependencies between machine subsystems.

The fault datasets include complete time series and binary alarm states, ground truth labels for manipulated variables and periods, expert-assigned diagnoses from 10 predefined types, and standardized timestamps. Normal operation datasets provide baseline behavior without known faults.

The dataset is publicly available at zenodo.org [19] under Apache License 2.0, with code and documentation at github.com [20].

Table 2. Dataset statistics for causRCA showing the hierarchical decomposition into sub-causal graphs and their corresponding fault scenarios. This dataset also contains 170 data points from normal operation.

(Sub-) Graph	Nodes	Edges	Scenarios	Datasets	Diagnoses	Faulty Vars.	Alarms
Full	92	104	19	100	10	14	18
Coolant	15	10	4	25	2	6	6
Hydraulics	17	18	9	41	5	6	7
Probe	11	15	6	34	3	2	2
Other	49	61	-	-	-	-	3

3. Benchmarking Causal Discovery on causRCA

To demonstrate causRCA's utility for causal discovery, we evaluated four established CD algorithms, providing baseline performance metrics for comparing methods on real-world manufacturing data. We selected *PC*, *FCI*, *FGES*, and *PCMCI* algorithms, following established benchmarking practices, specifically adapting methodologies from Pham et al. [21] and their accompanying source code.

We implemented dataset ensemble learning with majority voting – executing algorithms on multiple sub-datasets and determining edge presence by majority votes. Concatenation approaches showed significantly worse performance and runtime. We assessed performance using F1 score and Structural Hamming Distance (SHD) against the ground-truth expert graphs. Table 3 shows the results across the three subsystem graphs (Coolant, Hydraulics, Probe) and the complete system (Full).

Table 3. CD Algorithm Performance Comparison (dataset ensemble learning approach with majority voting)

Algo.	Coolant		Hydraulics		Probe		Full	
	F1	SHD	F1	SHD	F1	SHD	F1	SHD
<i>PC</i>	0.29	14	0.40	13	0.37	10	0.15	99
<i>FCI</i>	0.29	14	0.40	13	0.37	10	0.15	99
<i>FGES</i>	0.13	30	0.27	33	0.25	17	0.20	157
<i>PCMCI</i>	0.26	14	0.16	37	0.29	26	0.20	264

PC and *FCI* achieved identical performance: F1 scores 0.29–0.40, SHD values 10–14. *PC* and *FCI* produced identical results across all test cases, suggesting that with default settings, both algorithms converge to similar solutions.

FGES and *PCMCI* showed weaker performance on the subsystem graphs. *FGES* reached F1 scores of 0.13–0.27, and *PCMCI* achieved 0.16–0.29. The drop compared to *PC* and

FCI stems mainly from higher false discovery rates (FDR), indicating that *FGES* and *PCMCI* inserted more false positive edges (spurious causal connections) into the learned graphs.

For example, *PCMCI* on hydraulics reached an FDR above 0.8, indicating that the majority of discovered edges were incorrect. The additional, spurious edges are also the reason for the larger SHD values (17–37) of *FGES* and *PCMCI* in comparison with *PC* and *FCI* (10–14). These findings suggest that both *FGES* and *PCMCI* tend to overfit by adding spurious edges when applied to subsystem graphs, whereas *PC* and *FCI* maintain a more conservative structure.

The complete system proved more challenging than the subsystems. In our benchmark, F1 scores dropped markedly from subsystem graphs (11 – 17 nodes, $F1 \leq 0.40$) to the full system graph (92 nodes, $F1 \leq 0.20$). The decrease in causal discovery performance with increasing node count aligns with observations from other causal discovery benchmarks [21, 22].

The reason is structural: the number of possible causal graphs grows super-exponentially with the number of nodes, making discovery an NP-hard problem [23]. This not only degrades accuracy but also increases computational time [24], which we also witnessed in our benchmark experiments.

Our evaluation used default hyperparameters without extensive tuning, which likely contributed to the modest performance. Better parameter selection can improve causal discovery accuracy and partly mitigate the observed effects.

4. Benchmarking Root-Cause-Analysis on causRCA

To demonstrate the utility of causRCA for benchmarking RCA models, we provide an example benchmark comprising the evaluation of three supervised and two unsupervised RCA models on all fault datasets. The 100 fault datasets used for evaluation were generated with the HIL simulation and are accompanied by known manipulated variables (causes) and manually labeled diagnoses for each fault scenario.

Supervised RCA models treat the task as a multi-class classification problem with ranked outputs. In a training phase, the models are presented with samples, each consisting of a dataset and the true diagnoses as labels. During inference, the supervised RCA predicts a ranked list of diagnoses, with the most probable diagnosis on top.

Unsupervised RCA models try to identify anomalous variables that are likely causes of the fault scenario at hand. They take as input a dataset and rank the contained variables based on their probability of explaining the fault scenario. Table 4 summarizes our categorization of RCA methods.

Table 4. Supervised RCA and Unsupervised RCA categorization.

RCA Method	Training	Inference	Output	
	Input	Input	List of	Example
Super-vised	record + diag. labels	dataset record	ranked diagnoses	[diag2, diag1, diag3, ...]
Unsuper-vised	/	dataset record	ranked variables	[varZ, varX, varY, ...]

4.1. Example Benchmark Supervised RCA

We evaluate three supervised RCA algorithms on causRCA: two non-causal and one causal. For training and testing of each model, we use a three-fold stratified cross-validation as implemented in *scikit-learn* [25]. K-fold cross-validation splits the data into k equally sized subsets (folds) and iteratively uses one-fold for testing while training on the remaining folds. Stratification maintains the overall distribution of classification labels across folds and ensures that each train–test split contains all diagnosis types.

Evaluation proceeds under two difficulty settings. *Full* presents all available variables and alarms to the model, leaving it to the model to learn which inputs are relevant. *Sub* presents only variables and alarms from the designated subgraph (i.e., *Coolant*, *Hydraulics*, or *Probe*) to the model, resulting in fewer misclassifications. The limitation to relevant variables, as in *Sub*, mimics situations where expert knowledge is present to limit the scope of the RCA effectively.

We evaluate model performance using Mean Average Precision at k ($MAP@k$), a metric normally used for assessing ranked outputs in recommender systems, which is also highly descriptive for measuring RCA performance. $MAP@k$ measures whether, and at what rank, the correct results appear within the top k positions of the model’s predicted ranking. It emphasizes placing true positives near the top of the ranking, rewarding models that prioritize correct diagnoses. [26]

For supervised RCA, we chose $k=2$, equaling the maximum number of correct diagnoses that are active at the same time in the fault datasets used for evaluation. Table 5 summarizes the resulting $MAP@2$ for the included algorithms:

- BaselineSupervisedRCA (*Base*) identifies the oldest active alarm at diagnosis time and ranks associated diagnoses based on their number of occurrences in the training set.
- LogisticRegressionRCA (*LogReg*) utilizes the full system state as input features in a logistic regression classifier, ranking diagnoses based on their predicted probability.
- CausalPrioLogisticRegressionRCA (*CausReg*) applies the provided causal graph for feature selection and then ranks diagnoses using the *LogReg* classifier.

All three supervised models achieve high mean $MAP@2$ (M) scores on causRCA. *Base* performs well on *Coolant* and *Hydraulics* but is limited by low accuracy on *Probe* (0.54), resulting in a modest overall mean ($M=0.82$). Its strong results on two subsystems, achieved by simply ranking diagnoses based on occurrence frequency, suggest that the benchmark is relatively straightforward, likely due to the limited number of diagnosis choices.

LogReg achieves strong results under *Sub* ($M_{Sub}=0.98$) but drops in *Full* ($M_{Full}=0.93$). The decline is most evident on *Probe*, where performance decreases from 0.97 (*Sub*) to 0.80 (*Full*), indicating sensitivity to irrelevant inputs when all variables are included.

CausReg, in contrast, is both accurate and stable: it reaches the highest means ($M=0.98$) and avoids the degradation

observed for *LogReg* in the Probe subgraph, demonstrating that causal-graph-guided feature selection effectively identifies relevant inputs. Its performance indicates that embedding causal knowledge into supervised RCA helps maintain accuracy and improves robustness under high-dimensional settings where many irrelevant variables may be present.

Table 5. MAP@2 for prediction of diagnosis labels by supervised RCA models on causRCA subgraphs and mean over all subgraph evaluations.

Supervised RCA Algo	Coolant		Hydraulics		Probe		Mean (M)	
	Full	Sub	Full	Sub	Full	Sub	Full	Sub
Base	1.0	1.0	0.93	0.93	0.54	0.54	0.82	0.82
LogReg	1.0	1.0	0.99	0.98	0.80	0.97	0.93	0.98
CausReg	1.0	1.0	0.98	0.98	0.97	0.97	0.98	0.98

4.2. Example Benchmark Unsupervised RCA

We evaluate five unsupervised RCA algorithms on causRCA using Mean Average Precision at k (MAP@k), the same metric employed for supervised methods. Unlike supervised models, which produce a ranked list of diagnoses after a training phase, unsupervised models directly process each fault dataset and return an ordered list of variables ranked by their likelihood of causing the observed fault scenario.

We set $k = 3$ for evaluating the unsupervised methods, as no benchmark fault scenario involves more than three simultaneous anomalous variables. Table 6 summarizes MAP@3 scores for the two non-causal (*TimeRank*, *Baro*) and three causal unsupervised models (*CausTR*, *RandWalk*, *PageRank*) under evaluation:

- *TimeRankUnsupervisedRCA* (*TimeRank*) serves as a simple baseline model and ranks variables by their proximity to alarm activations, prioritizing those with the most recent value changes before the alarm.
- *Baro* [27] is a representative non-causal benchmark that uses robust scaling to differentiate normal (before alarm) and abnormal periods (after alarm) and ranks variables by deviation magnitude. Our implementation follows [21].
- *CausalPrioTimeRankRCA* (*CausTR*) uses the provided causal graph to get all variables that can potentially cause the active alarms and then ranks these using *TimeRank*.
- *RandomWalk* [28] (*RandWalk*) performs random walks on the causal graph and ranks variables by visitation frequency, reflecting their likelihood of contributing to the fault. Our implementation follows [21].
- *PageRank* [29] evaluates node importance in the causal graph based on the number and quality of incoming edges, ranking variables with stronger and more informative connections as more likely root causes. Our implementation follows [21].

All unsupervised RCA algorithms achieve lower MAP@3 mean scores (M) than supervised models, confirming the difficulty of diagnosis without labeled examples. Among the non-causal methods, *TimeRank* performs reasonably well when the candidate set is restricted ($M_{Sub}=0.60$), but its performance

drops sharply when all variables are present ($M_{Full}=0.24$). *Baro* shows an even stronger dependence, performing well on subgraphs ($M_{Sub}=0.75$) but failing almost completely on full graphs ($M_{Full}=0.09$). These results indicate that non-causal models benefit heavily from preselection and accumulate false positives when many unrelated variables are included.

Causal methods, in contrast, exhibit varying robustness across the evaluated RCA methods and graph sizes. While *CausTR* achieves near identical scores in *Full* and *Sub* ($M=0.77-0.78$), *PageRank* and *RandWalk* show a significant performance drop in *Full* ($M_{Full}=0.04$ and $M_{Full}=0.00$).

CausTR achieves the highest mean score and outperforms its non-causal predecessor *TimeRank* by combining situation-specific information (active alarms) with causal information (causes of alarms). The combination yielded stable performance, regardless of variable preselection and graph size.

In contrast, *PageRank* and *RandWalk* rely solely on structural information. As implemented, they are situation-agnostic, ranking nodes purely by structural importance, independent of the specific situation and current alarms.

On small graphs, *PageRank* performs strongly, notably on Coolant and Hydraulic, where faulty nodes are structurally central. In the *Full* graph, the structurally central nodes do not coincide with the actual causes, leading to poor performance.

RandWalk, ranks lowest, showing that undirected random exploration of the graph is insufficient.

The overall scaling issues of methods that rely solely on data or solely on causal structure indicate that neither is sufficient for effective RCA in complex systems. The good scalability of *CausTR* indicates that combining live data with causal information can improve RCA performance and robustness.

Table 6. MAP@3 for prediction of faulty variables by unsupervised RCA models on causRCA subgraphs and mean over all subgraph evaluations. Methods using the expert causal graph as input are marked with [*].

Unsuperv. RCA Algo	Coolant		Hydraulics		Probe		Mean (M)	
	Full	Sub	Full	Sub	Full	Sub	Full	Sub
<i>TimeRank</i>	0.35	0.85	0.18	0.47	0.19	0.47	0.24	0.60
<i>Baro</i>	0.00	0.88	0.26	0.75	0.00	0.63	0.09	0.75
<i>CausTR</i> [*]	0.99	1.00	0.74	0.74	0.57	0.58	0.77	0.78
<i>RandWalk</i> [*]	0.00	0.14	0.00	0.14	0.00	0.32	0.00	0.20
<i>PageRank</i> [*]	0.00	1.00	0.11	0.98	0.00	0.26	0.04	0.75

5. Causal Discovery Performance Impact on RCA

To investigate the relationship between causal graph quality and RCA performance, we evaluated the *CausTR* unsupervised RCA algorithm using the causal graphs learned by the four CD algorithms presented in Section 3 and compare it with the RCA performance of *CausTR* with the expert graph. Table 7 presents MAP@3 scores achieved by *CausTR* using different causal graphs as input. RCA performance varies substantially depending on the underlying causal graph. Using the expert graph ($F1=1.00$), *CausTR* achieves MAP@3 scores of 1.00 for Coolant, 0.74 for Hydraulics, and 0.58 for Probe subsystems. Graphs generated by CD algorithms yield lower MAP@3

scores: PC and FCI (both achieving $F1=0.29-0.40$) result in $MAP@3$ scores ranging from 0.37 to 0.87, while FGES ($F1=0.13-0.27$) produces scores between 0.19 and 0.55, and PCMCI ($F1=0.16-0.29$) achieves 0.21 to 0.56.

Table 7. $MAP@3$ performance scores for unsupervised RCA *causTR* across subsets using causal graphs obtained from four CD algorithms and the expert-graph. The F1 score compares the learned graph with the expert graph.

Graph origin for CausTR	Coolant (Sub)		Hydraulics (Sub)		Probe (Sub)	
	F1	$MAP@3$	F1	$MAP@3$	F1	$MAP@3$
PC	0.29	0.87	0.40	0.48	0.40	0.37
FCI	0.29	0.87	0.40	0.48	0.40	0.37
FGES	0.13	0.55	0.27	0.37	0.25	0.19
PCMCI	0.26	0.56	0.16	0.37	0.29	0.21
Expert Graph	1.00	1.00	1.00	0.74	1.00	0.58

To systematically examine this relationship, we generated modified versions of the ground truth graphs with controlled degradation levels. By randomly adding and removing edges, we created graph variants with F1 scores ranging from 0.3 to 1.0 in increments of 0.1.

Fig. 3 displays $MAP@3$ distributions when CausTR uses these degraded graphs across all subsystems. The results reveal an increasing trend: median $MAP@3$ rises from ~ 0.45 at $F1=0.3$ to ~ 0.75 at $F1=1.0$. Notably, interquartile ranges decrease with higher F1 scores. The widest variation occurs at lower F1 scores (0.3-0.5), while performance becomes more consistent at higher scores (0.8-1.0).

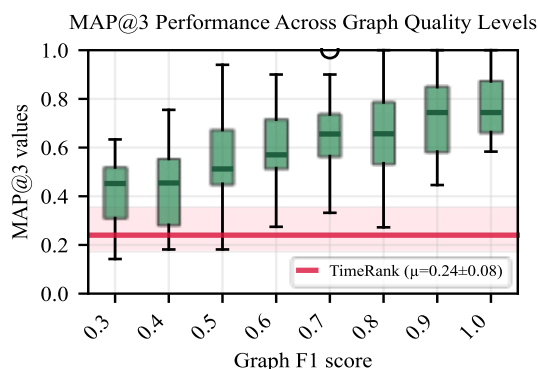


Fig. 3. Distribution of $MAP@3$ scores for CausTR across different causal graph quality levels (F1 scores from 0.3 to 1.0). Box plots show median, quartiles, and range for all three subsystems combined. The red line indicates TimeRank baseline performance ($\mu=0.24\pm 0.08$).

Our analysis reveals practical deployment insights. The non-linear F1- $MAP@3$ relationship shows diminishing returns at higher F1 scores: modest improvements (F1 0.3 \rightarrow 0.5) yield substantial RCA gains, suggesting perfect causal knowledge is not necessary to achieve good RCA performance. However, poor graphs also increase performance variance, making the RCA performance on these graphs less predictable and reliable.

Subsystem-specific performance ceilings ($MAP@3$: 1.0 for Coolant vs. 0.58 for Probe with perfect graphs) indicate that causal structure is beneficial but not sufficient in every scenario. As CausTR is a relatively simple algorithm, future research will likely lift this ceiling and produce better

unsupervised RCA performance by capitalizing stronger on the causal graph and machine learning.

6. Discussion

Our work addresses a practical need in fault diagnosis by providing a dataset that supports both CD and RCA benchmarking under realistic manufacturing conditions. Through *causRCA*, we show that combining real operational data, expert-validated causal structure, and labeled fault scenarios yields a valuable resource for fault diagnosis research. Our benchmarks demonstrate the usability of *causRCA* for evaluating causal discovery, supervised RCA, and unsupervised RCA separately and in combination, for example, to explore questions such as "How does the graph-learning performance of CD algorithms affect the performance of RCA methods using learned causal graphs?".

Because the dataset and graph structure are modular, researchers can evaluate subsystems individually or as a whole, allowing them to test methods across different levels of complexity. By working with these distinct subsystems and granularities, nuanced insights into how the algorithms perform under varying conditions are obtainable.

The RCA performance on different subsets indicates that the probe subsystem presents the most challenging diagnosis due to the significant variability in diagnoses from the same number of observed variables. Although the supervised RCA methods significantly outperform the unsupervised RCA methods, the effectiveness of the unsupervised approach is notable. Unsupervised RCA methods do not require labeled training data and can therefore diagnose faults on their first occurrence, which is not possible for supervised approaches. The ability of unsupervised RCA methods to operate without labeled data and diagnose faults on first occurrence opens the possibility of creating robust hybrid diagnostic systems that leverage unsupervised learning for novel faults and incorporate supervised models as labels become available.

A central finding in the evaluation of causal discovery methods is the significant challenge of learning effective causal representations for systems with many observed variables. However, our results also establish a clear, positive relationship between the quality of a discovered causal graph and RCA performance. Our analysis of the influence of the F1 score on the causal discovery performance of CausTR implies that even imperfect causal insights are valuable. This finding requires further validation using other causal methods, but it indicates that progress in improving CD will directly enhance automated RCA, making it a promising area for future research.

At the same time, *causRCA* represents only a single machine type, a CNC vertical lathe. This setup enables precise subsystem-level benchmarking; however, its generalizability to other machine classes or multi-machine remains to be evaluated. Many algorithms already perform strongly on *causRCA*, underlining the importance of extending benchmarking to multiple datasets to avoid overfitting to dataset-specific properties. For instance, our baseline methods used alarm detection times, which may not be available in other

industrial datasets. Our benchmarks also did not consider computational efficiency, runtime aspects, or real-time capabilities, which are critical for deployment.

Future work should therefore assess whether methods retain performance across different settings and datasets and systematically evaluate the runtime of CD and RCA methods.

7. Conclusion

Our work addresses the lack of datasets that allow the joint evaluation of CD and RCA methods by introducing causRCA. This publicly available dataset combines real-world machine data with realistic fault records, labeled causes, and a ground truth causal graph. In the work, we provided example benchmarks and code to demonstrate how causRCA enables rigorous evaluation of CD, supervised RCA, and unsupervised RCA methods. We evaluate the combination of CD with RCA and show that even imperfect learned causal graphs significantly enhance RCA performance, underscoring the practical value of including causal graphs in RCA. By releasing data, code, and benchmarks under an open-source license, we aim to foster progress toward more reliable, explainable, and applicable diagnostic systems in manufacturing.

Acknowledgements

The dataset originated from the research project KausaLAssist, funded by the German Federal Ministry of Education and Research (BMBF) under grant 02P20A150. The authors thank the project partners for their valuable contributions in making the collection and release of the dataset possible: KAMAX Holding GmbH & Co. KG, Schuster Maschinenbau GmbH, ISG Industrielle Steuerungstechnik GmbH, and SEITEC GmbH.

Declaration of generative AI usage in the writing process

During the preparation of this work, the authors used ChatGPT, Claude, and Grammarly to improve the language and readability of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

- [1] e Oliveira, E., Miguéis, V.L., Borges, J.L., 2023. Automatic root cause analysis in manufacturing: an overview & conceptualization
- [2] Pietsch, D., Matthes, M., Wieland, U., Ihlenfeldt, S. et al., 2024. Root Cause Analysis in Industrial Manufacturing: A Scoping Review of Current Research, Challenges and the Promises of AI-Driven Approaches 8, p. 277.
- [3] Siemens AG, Digital Industries, 2024. *The True Cost of Downtime 2024*, Erlangen, Germany.
- [4] Pearl, J., 2009. Causal inference in statistics: An overview 3.
- [5] Vuković, M., Thalmann, S., 2022. Causal Discovery in Manufacturing: A Structured Literature Review 6, p. 10.
- [6] Gong, C., Zhang, C., Di Yao, Bi, J. et al., 2024. Causal Discovery from Temporal Data: An Overview and New Perspectives 57, p. 1.
- [7] Webert, H., Simons, S., McGibney, A., 2025. A practical investigation of ML and Industry 4.0 for reactive fault detection in manufacturing systems 253, p. 1800.
- [8] Zheng, L., Chen, Z., Wang, D., Deng, C. et al., 2024. LEMMA-RCA: A Large Multi-modal Multi-domain Dataset for Root Cause Analysis *abs/2406.05375*.
- [9] Granderson, J., Lin, G., Chen, Y., Casillas, A., Im, P., Jung, S., Benne, K., Ling, J., Gorthala, R., Wen, J., Chen, Z., Huang, S., Vrabie, D., 2022. *LBNL Fault Detection and Diagnostics Datasets*. DOE Open Energy Data Initiative (OEDI); Lawrence Berkeley National Laboratory.
- [10] Goh, J., Adepu, S., Junejo, K.N., Mathur, A., 2017. A Dataset to Support Research in the Design of Secure Water Treatment Systems, in *Critical Information Infrastructures Security*, Springer International Publishing, Cham, p. 88.
- [11] Göbler, K., Windisch, T., Pychynski, T., Sonntag, S., Roth, M., Drton, M., 2023. *causalAssembly: Generating Realistic Production Data for Benchmarking Causal Discovery*. arXiv.
- [12] Mogensen, S.W., Rathsmann, K., Nilsson, P., 2024. Causal discovery in a complex industrial system: A time series benchmark, in *Proceedings of the Third Conference on Causal Learning and Reasoning*, PMLR.
- [13] Stein, G., Shadaydeh, M., Blunk, J., Penzel, N., Denzler, J., 2025. *CausalRivers – Scaling up benchmarking of causal discovery for real-world time-series*.
- [14] Liu, C., Yang, W., Mittal, H., Singh, M. et al., 2023. PyRCA: A Library for Metric-based Root Cause Analysis *abs/2306.11417*.
- [15] Giovanni Menegozzo, D. Dall’Alba, P. Fiorini, 2022. CIPCAd-Bench: Continuous Industrial Process datasets for benchmarking Causal Discovery methods, p. 2124.
- [16] Pham, L., Zhang, H., Ha, H., Salim, F. et al., 2025. RCAEval: A Benchmark for Root Cause Analysis of Microservice Systems with Telemetry Data, in *Companion Proceedings of the ACM on Web Conference 2025*, ACM, New York, NY, USA, p. 777.
- [17] Mengshoel, O.J., Darwiche, A., Cascio, K., Chavira, M. et al., 2008. Diagnosing Faults in Electrical Power Systems of Spacecraft and Aircraft, in *AAAI Conference on Artificial Intelligence*.
- [18] Scheifele, C., Verl, A., 2018. Hardware-in-the-Loop Simulation for Machines based on a Multi-Rate Approach, in *Proceedings of The 9th EUROSIM Congress on Modelling and Simulation, EUROSIM 2016*, Linköping University Electronic Press, p. 715.
- [19] Mehling, C.W., Pieper, S., Lüke, T. causRCA: Real-World Dataset for Causal Discovery and Root Cause Analysis in Machinery. Zenodo, 2025. <https://doi.org/10.5281/zenodo.15876410>.
- [20] Mehling, C.W., Pieper, S., Lüke, T. Benchmarking of Automated Root Cause Analysis and Causal Discovery in Manufacturing Using the causRCA Dataset (CIRPe 2025), 2025. <https://github.com/causalgraph/causRCA>.
- [21] Pham, L., Ha, H., Zhang, H., 2024. Root Cause Analysis for Microservice System based on Causal Inference: How Far Are We?, in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ACM, New York, USA, p. 706.
- [22] Ke, N.R., Didolkar, A., Mittal, S., Goyal, A., Lajoie, G., Bauer, S., Rezende, D., Bengio, Y., Mozer, M., Pal, C., 2021. *Systematic Evaluation of Causal Discovery in Visual Model Based Reinforcement Learning*.
- [23] Chickering, D.M., 1996. Learning Bayesian Networks is NP-Complete, in *Learning from Data*, Springer New York, p. 121.
- [24] Arjun Sondhi, Ali Shojaie, 2019. The Reduced PC-Algorithm: Improved Causal Structure Learning in Large Random Networks
- [25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. et al., 2011. Scikit-learn: Machine Learning in Python 12, p. 2825.
- [26] Schröder, G., Thiele, M., Lehner, W., 2011. Setting Goals and Choosing Metrics for Recommender System Evaluations, in *Proceedings of the UICERSTI2 Workshop at the 5th ACM Conference on Recommender Systems*, Chicago, USA, p. 53.
- [27] Pham, L., Ha, H., Zhang, H., 2024. BARO: Robust Root Cause Analysis for Microservices via Multivariate Bayesian Online Change Point Detection 1, p. 2214.
- [28] Spitzer, F., 2001. *Principles of random walk*, 2nd edn. Springer, New York, NY, Berlin, Heidelberg.
- [29] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab.