

Article

Model-Based Posterior Correction for Isolating Structurally Indistinguishable Faults

Young Eui Min  and Man-Je Kim * 

Convergence of AI, Chonnam National University, Gwangju 61186, Republic of Korea; alsduddml77@jnu.ac.kr

* Correspondence: jaykim0104@jnu.ac.kr

Abstract

Model-based fault diagnosis narrows down fault candidates by checking the consistency between a mathematical model of the plant and the measurements. Depending on the sensor configuration and the system structure, however, different faults may share an identical residual signature. In such cases, the diagnosis keeps the true fault inside the candidate set but cannot single it out, and this structural indistinguishability is not resolved by tuning thresholds or post-processing the residuals. This paper proposes a method that reuses the diagnostic model on the candidates left by the existing diagnosis in order to rank them. Assuming each candidate to be the true fault and propagating the model forward yields predictions whose comparison with the actual measurements exposes temporal differences that the residual signature does not carry. A limited computational budget is distributed over several candidates so that the comparison is well defined, and the weight given to the two sources of information is adjusted according to how confident the existing diagnosis is. On a three-tank fault diagnosis benchmark, the probability assigned to the true fault at instants where the existing diagnosis selected a different candidate increased from 0.2547 to 0.6735; the isolation accuracy over the three faults sharing an identical residual signature improved from 31.5% to 52.7%; and the accuracy over all seven faults improved from 59.4% to 76.7%. On that benchmark, learning-based methods taking the same residuals as input did not exceed 0.30 on the indistinguishable set even with 512 samples per fault, whereas the proposed method uses no fault data at all. On a second benchmark, an electric motor, the same procedure raised that probability from 0.3523 to 0.7110. Across twelve combinations of fault magnitude and process noise, the method was never below the existing diagnosis, although the gain exceeded 0.34 only in the six where the fault magnitude was at least 12.5 times the process noise standard deviation.

Keywords: model-based fault diagnosis; fault isolation; structural indistinguishability; residuals; posterior correction

MSC: 68T07



Academic Editor: Jonathan Blackledge

Received: 20 September 2026

Revised: 17 October 2026

Accepted: 20 October 2026

Published: 22 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Machinery and process plants are subject to a variety of faults during operation, and detecting a fault early while correctly identifying its cause is essential for stable operation. Fault diagnosis is the process of judging whether a plant is behaving abnormally and what has caused it, and it is commonly divided into fault detection, which determines whether a fault is present, and fault isolation, which identifies the actual fault among several possible causes [1]. This work addresses fault isolation.

Model-based fault diagnosis uses a mathematical model of the plant dynamics to diagnose faults. The discrepancy between the values predicted by the model and the actual measurements is expressed as a residual: under nominal conditions, a residual stays within a fixed range, and when a fault occurs, certain residuals leave that range. Because such residuals rely on a physical model, faults can be diagnosed without collecting a large amount of fault data in advance, which distinguishes this line of work from data-driven approaches that learn patterns from past operating records [2].

A single residual reacting to a particular fault is not by itself sufficient to determine the cause. Several residuals must be used together so that the set of residuals affected by each fault can be described. This description is called the fault signature, and the combination of residuals that react under each fault is used to narrow down the possible candidates. Structural analysis makes it possible to construct these residuals and fault signatures automatically from the plant equations and the sensor configuration [3,4].

This approach is interpretable because it relies on a physical model, and it remains applicable when fault data are scarce. Fault isolation, however, is not always achievable. When different faults produce the same combination of residuals, the two faults cannot be distinguished under the current measurement structure.

If the combination of reacting residuals differs across faults, the observed residual pattern can be used to narrow the candidate set. Conversely, if two faults exhibit the same reaction over every residual, they share an identical fault signature.

This is not an accident of a particular dataset. Which faults a residual reacts to are determined by the equations that constitute the residual, and which equations can be used to form a residual are determined by the sensor configuration and the structure of the system. Two faults that stand in the same structural relation to the measurements can therefore share an identical fault signature.

Such faults are the ones a structural isolability analysis marks as isolable only group-wise; they are referred to here as structurally indistinguishable faults. The three-tank system in Figure 1 exhibits this situation. A leak at the second valve f_{V_2} , a leak at the third valve f_{V_3} , and a leak in the third tank f_{T_3} share an identical residual signature. All three influence the measurements only through the unmeasured pressure p_3 and are therefore observed in the same way under the current sensor configuration. In the fault signature matrix obtained by structural analysis, the columns of the three faults coincide exactly.

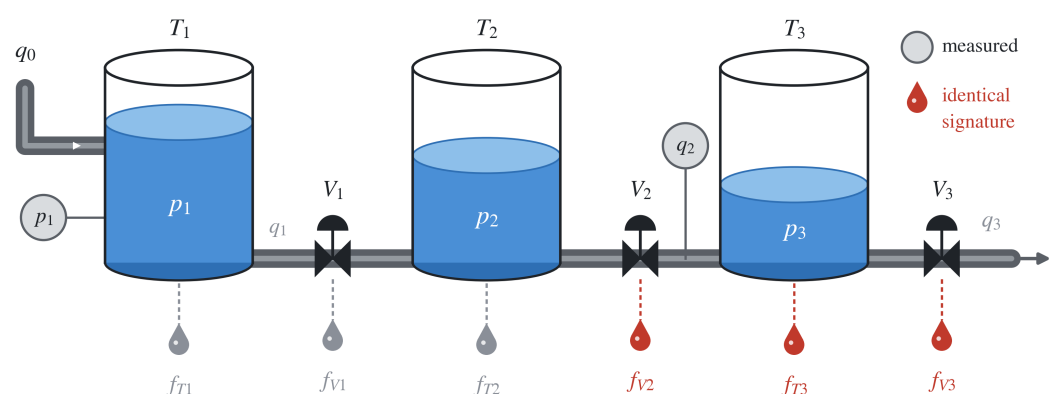


Figure 1. Configuration and measurement points of the three-tank system. Here p_1 , p_2 and p_3 are the liquid pressures in the three tanks and q_2 the flow through the second valve. Only the pressure p_1 and the flow q_2 are measured, while p_2 and p_3 are unobserved. The three faults marked in red share an identical residual signature.

Consequently, when one of these three faults occurs, consistency-based diagnosis does not remove the true fault from the candidate set, yet the pattern of threshold crossings

alone cannot determine which of the candidates is the actual fault. As shown in Figure 2, the true fault was contained in the candidate set 99.11% of the time in the experimental data, while the average candidate set size was 3.27, and the three faults, whenever they remained candidates, did so together. The issue is therefore not the failure to detect the fault but an inability to narrow the remaining candidates after detection.

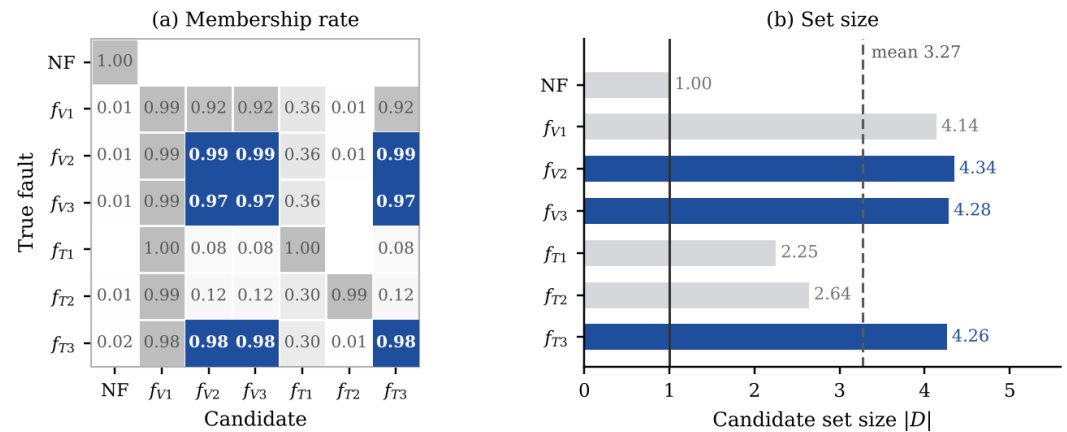


Figure 2. Characteristics of the output of the existing diagnosis. (a) Candidate membership rate for each true fault; the blue cells indicate the cases in which the three faults of the indistinguishable set contain one another as candidates. (b) Candidate set size, which for the three faults substantially exceeds the overall average.

Resolving this structural limitation requires information other than the residual pattern itself. This work turns to the model that is already being used for diagnosis. Even for faults sharing an identical residual signature, the way the system evolves under the assumption that each of them is the true fault may differ. Computing the future behaviour of each candidate and comparing it with the actual measurements therefore yield comparative information that the residual signature cannot provide.

Prior work has approached this difficulty in three ways, none of which applies without cost here. Refining the residuals produces no new discriminative information once the fault signatures coincide, while learning from fault data and injecting test inputs require, respectively, records of past failures and intervention in a plant that is in operation. Section 2 examines these directions in detail.

This work proposes a method that makes the already established diagnostic assets usable while at the same time generating discriminative evidence that the residual signature cannot provide. Model-based diagnosis already possesses a model of the plant, and this model can compute how the system will evolve under an assumed fault. In common practice, however, the model is used only to generate residuals and is no longer exploited once the candidate set has been computed. The proposed method mobilises this unused resource again to rank the candidates, evaluating the hypotheses of time t at time $t + L$ against the measurements observed in between. It operates as a layer that takes the output of the existing diagnosis as its input, so the diagnostic system need not be replaced or redesigned, and neither new fault data nor an input applied to the running plant is required. Obtaining isolation performance without paying the price demanded by each of the three directions above is what distinguishes this work.

The objective of this work is to isolate the candidates by making use of the existing diagnostic result in situations where candidates sharing a structurally identical fault signature prevent model-based diagnosis alone from achieving isolation. Three difficulties follow: (1) where to obtain discriminative evidence that is not contained in the residual signature; (2) how to allocate the computation needed to obtain that evidence within a limited budget; and (3) how to combine the new evidence with the existing diagnosis

without damaging the instants that were already judged correctly. The following method is proposed to solve these problems.

1. Assuming each fault candidate left by consistency-based diagnosis to be true, the future behaviour is computed, and discriminative evidence not contained in the residual signature is obtained from the agreement with the actual measurements.
2. A limited computational budget is distributed among the remaining candidates in proportion to the initial diagnostic probabilities.
3. When the new evidence is combined with the existing diagnostic result, the weight is adjusted according to the uncertainty of the diagnosis so that the outcome at instants already judged correctly is not damaged.
4. The method is compared with existing model-based diagnosis methods and learning-based methods on a three-tank benchmark, and both the isolation performance on structurally indistinguishable fault candidates and the overall diagnostic performance are evaluated.

2. Related Work

Attempts to improve fault isolation performance under a given sensor configuration are divided according to where the evidence for the decision is obtained. The following section reviews how model-based diagnosis has developed, examines the three directions in turn, and finally states the position of this work.

Model-based fault diagnosis originated in the 1980s along two branches with different academic backgrounds. In artificial intelligence, De Kleer and Williams [5] formalised consistency-based diagnosis, which infers possible fault combinations from the logical contradiction between a description of the nominal behaviour of the system components and the observations. The approach finds the minimal sets of healthiness assumptions that cannot all hold (called conflicts) and takes the minimal hitting sets of those conflicts as diagnostic candidates. In the same period, control theory pursued fault detection and isolation, in which residual signals were designed from the system model to detect deviations caused by faults. This branch focused on constructing residuals that are sensitive to particular faults and insensitive to the rest, using observers or parity equations.

The two streams developed separately for a long time, addressing discrete events and continuous dynamics, respectively, but began to merge into a single framework in the 2000s. Nyberg and Krysander [6] presented a diagnostic framework integrating diagnostic reasoning from Artificial Intelligence, residual design from control theory, and statistical hypothesis testing. Structural analysis served as an important intermediary in this merger. Krysander et al. [3] presented an algorithm that efficiently finds minimally structurally overdetermined sets, the smallest units from which residuals can be constructed, from a bipartite graph that encodes only the incidence between equations and variables. Because the method applies without knowledge of the specific form of the equations or the parameter values, it can be used to analyse diagnosability at an early stage of plant design.

Frisk et al. [4] implemented this structural analysis procedure as a toolbox, automating the design of residual generators and the analysis of fault isolability even for large models. The toolbox enumerates the minimally structurally overdetermined sets from the model equations, determines a computation order for each set, and emits residual generators with integral causality as code. It also computes which fault pairs are mutually isolable from the fault signature matrix, allowing the upper bound on diagnostic performance to be confirmed at the design stage. This approach has been applied to the air intake path and the exhaust aftertreatment system of automobiles [7,8], to multi-engine clusters of launch vehicles [9], and to chemical processes and hydraulic systems. What these cases

share is that a physical model of the plant exists while fault case data are difficult to obtain in sufficient quantity.

The isolability analysis computed from the fault signature matrix reveals at the design stage which fault pairs cannot be distinguished. This information is normally used to decide where to add sensors, but in a plant already in operation, the measurement configuration often cannot be changed. Under that constraint, the research divides into three directions.

2.1. Improving the Quality of the Residuals

The first direction attempts to extract more information from the measurements already available. In large plants, tens of thousands of equation combinations can form residuals, so selecting which residuals to use is itself a design problem. Murata et al. [9] addressed multi-engine clusters of launch vehicles and defined a residual sensitivity index from the signal-to-noise ratio of each residual with respect to an individual fault. Using this index, they selected—among the minimum-cardinality subsets that give each fault a unique signature—the combination with the highest sensitivity, and when the observed signature disagreed with the theoretical one, they estimated the active fault by penalising mismatches on highly sensitive residuals more heavily.

Suppression of false alarms has been studied in the same context. A false alarm is a threshold crossing that raises an alert when in fact no fault is present or when the residual in question should not react. In consistency-based diagnosis, a false alarm causes more than mere noise. Because Equation (3) removes from the candidate set any fault that does not react to a residual that has raised an alarm, a single spurious alarm can erase the true fault from the candidates. In this respect, a false alarm can be more harmful than a missed detection. Mohammadi et al. [8] addressed the urea injection system of heavy trucks: they estimated the predictive distribution of the residuals with an ensemble of probabilistic neural networks trained on nominal data only; treated intervals with high epistemic uncertainty as out-of-distribution data and excluded them from the decision; and then detected faults with an adaptive threshold. Within the same line of research, it has also been proposed that a neural network replace the residual generator itself [10], letting a regression model trained on nominal data account for the part where the physical model is incomplete, thereby transferring the computation graph of structural analysis into a neural network architecture.

Work in this field is effective in improving the accuracy and stability of diagnosis. For faults that share an identical fault signature, the binary structural sensitivity pattern is identical, although their residual magnitudes or predictive distributions need not be. Such quantitative differences may provide additional discrimination in some cases, but they do not in general remove the structural non-isolability indicated by coincident signature columns. Mohammadi et al. [8] report from their structural analysis that, even after introducing both neural network-based residuals and uncertainty quantification, the pump filter fault in their target plant remains inseparable from another filter fault and from a return pipe blockage.

2.2. Learning from Fault Data

The second direction obtains the evidence for the decision from fault data collected in the past. Even when the fault signatures are identical, the temporal behaviour produced by each fault differs, so a classifier can learn the difference given sufficient data containing that behaviour.

Jung et al. [7] presented a representative form of this combination. Model-based diagnosis first narrowed the candidates that were not contradicted by the observations, and a data-driven classifier then evaluated the relative likelihood of each candidate within that

set. In using the discriminative power provided by data while retaining the interpretability provided by the model, this work is the closest to the present one in its problem setting. In that study, however, the trained classifier created the ranking within the candidate set, and training data for each fault class were therefore required.

Lundgren and Jung [11] proposed an open-set classification scheme that divided the residual time series into fixed-length windows, modeled each window as a multivariate normal distribution, and decided whether to reject each fault class by comparing the Kullback–Leibler divergence with the class distributions obtained from training data. This work explicitly addresses situations in which different faults produce similar residual outputs and classification becomes ambiguous, and it notes that such ambiguity is particularly pronounced when the fault magnitude is small, which is a concern shared with the present paper.

Dash et al. [12] proposed a hybrid architecture that fed residuals computed from a bond graph model into a convolutional neural network, and they measured isolation performance while varying the number of training samples per fault from 16 to 512. Because this experiment treated the required amount of labelling quantitatively, it has direct relevance to the present work, and the comparative experiments in this paper follow its configuration.

This direction of study can in principle create the required distinction, but at the cost of obtaining data in which each fault actually occurred. Repeatedly inducing faults to collect data is not realistic in safety-critical plants, and for new plants or early stages of development, such records do not exist [13].

2.3. Stimulating the Plant to Obtain New Observations

The third direction applies a test input to the system so that the predictions of the fault hypotheses diverge as much as possible, and it then rules out the hypotheses that do not match the newly obtained measurements. The idea appeared first outside the diagnosis field. In chemical process research, the problem of model discrimination, which decides which of several candidate reaction models best explains the experimental data, was addressed early on, and because each experiment is costly, the choice of experimental conditions became the central decision.

In control theory, the same idea developed into active fault diagnosis. Nikoukhah [14] presented design conditions for input signals that guarantee fault detection in linear dynamic systems, and Šimandl and Punčochář [15] formulated active detection and control as a single optimisation problem. Scott et al. [16] introduced set-based methods that compute the reachable output range under each assumed fault as a zonotope, and they designed inputs that conclusively exclude faults whose range does not contain the measurement. This line of work has been applied to actuator and sensor fault discrimination in aircraft and robots.

This direction shares the goal of the present work in that it confronts structural indistinguishability directly. Because it requires deliberately injecting signals into a plant in operation, however, its applicability is restricted in situations where intervention is not permitted, such as a process in production or an aircraft in flight.

This work uses a fourth source. The evidence is obtained by computing the future behaviour of each candidate with the model already at hand, so neither new data nor intervention in the plant is required, and the basis for the decision remains explained by physical equations.

3. Problem Definition

The pattern of threshold crossings alone cannot create a ranking among candidates sharing an identical fault signature. The following formalises the condition and then defines the problem addressed in this work.

Fault isolation is the problem of identifying, given the observed measurements, which of a predefined list of faults has actually occurred. Model-based diagnosis solves this problem in three steps. It first forms several residuals from the plant model, then checks whether each residual has crossed its threshold, and finally retains as candidates only those faults that can react to every residual that has crossed. Isolation is complete once the candidate set has been narrowed to one.

Underlying this procedure is knowledge of which faults each residual reacts to. This relation is derived from the structure of the model rather than from simulation: it suffices to check which fault signals appear in the set of equations constituting the residual, and neither the specific forms of the equations nor the parameter values are required. The table organising this relation is the fault signature matrix, as formalised below.

The target system is written in discrete time as follows.

$$\begin{aligned}x(t+1) &= g(x(t), u(t), \theta) + w(t) \\ y(t) &= h(x(t), u(t), \theta) + v(t)\end{aligned}\quad (1)$$

Here, x is the state, u the known input, and y the measurement, while w and v denote process noise and measurement noise, respectively. The vector θ represents the fault state: all of its components are zero in the nominal case, and the corresponding component changes when a fault occurs. The set of fault hypotheses is $H = \{h_0, h_1, \dots, h_{K-1}\}$, where h_0 denotes the no-fault case.

Residuals are formed from part of these equations rather than from all of them. In a structurally overdetermined set of equations, equations remain after all unknowns have been computed, and the degree to which the remaining equation disagrees with the observation becomes the residual. The smallest such unit is called a minimally structurally overdetermined (MSO) set, and structural analysis yields M residuals $r_1, \dots, r_M$. Since each residual reacts only to the faults appearing in the equations it uses, the reaction can be organised as a matrix.

$$FSM[m, k] = \begin{cases} 1, & \text{if } r_m \text{ reacts to } h_k \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Each residual is evaluated as the mean absolute value over a window of the most recent W steps, $\bar{r}_m(t) = \frac{1}{W} \sum_{i=0}^{W-1} |r_m(t-i)|$, and the threshold τ_m is calibrated on nominal data for that quantity. In writing $\alpha_m(t) = \mathbf{1}(\bar{r}_m(t) > \tau_m)$ for whether residual m has crossed its threshold at instant t , the set of fault candidates that are not contradicted by the observation is

$$D(\alpha) = \begin{cases} \{h_0\}, & \text{if } \sum_m \alpha_m = 0 \\ \{h_k \in H : \forall m, \alpha_m = 1 \Rightarrow FSM[m, k] = 1\}, & \text{otherwise} \end{cases} \quad (3)$$

That is, only faults that can react to every alarming residual remain as candidates. A residual that should have reacted but did not raise an alarm is not used as grounds for exclusion, since the fault magnitude may have been too small for detection.

This procedure can narrow the candidate set only when the combination of reacting residuals differs across faults. If some faults exhibit the same reaction over every residual, the corresponding columns of the fault signature matrix coincide exactly, and Equation (3)

then treats those faults in the same way at all times. Generalising this, a fault subset $S \subseteq H$ satisfying the following is called an indistinguishable fault set.

$$FSM[m, i] = FSM[m, j], \quad \forall m, \quad \forall h_i, h_j \in S \quad (4)$$

In the three-tank system, fV_2 , fV_3 , and fT_3 satisfy this condition, and these three are referred to hereafter as the indistinguishable set.

Such a set moves as a single block in the candidate decision: whatever alarm pattern is observed, its members either all remain candidates or are all excluded, so the candidate set never narrows to one of them. This is stated as Proposition 1 and Corollary 1 in Section 5, where it is shown to hold for any alarm pattern the thresholds produce. Separating the interior of S therefore requires information that the pattern of threshold crossings does not contain.

Such information can be obtained from the model already in use for diagnosis. Assuming each candidate h_k to be true, the system model can predict the subsequent measurements, and even with identical fault signatures, the predicted behaviour differs when the fault location and the system dynamics differ. Whereas the alarm pattern carries only whether a residual crossed its threshold, this prediction carries the entire time course of the values, so comparison with the actual measurements reveals differences between candidates that the fault signature cannot provide.

Accordingly, this work addresses the following problem. Given the candidate set D obtained from the existing diagnosis, the measurement history $y_{1:t}$, the system model, and a computational budget T obtain a posterior probability satisfying the three conditions below. The fault hypotheses generated at time t are evaluated at time $t + L$ using the measurements $y(t + 1 : t + L)$. The diagnosis is therefore issued with a latency of L steps, over which the propagated predictions are compared against the measurements that actually arrive. This window is what supplies the temporal evidence that the residual signature does not carry.

1. When the diagnosis has selected a wrong candidate, raise the probability of the true fault.
2. When the diagnosis has already selected the true fault, do not unnecessarily weaken that judgement.
3. Keep the total amount of model computation over the candidates within the given budget T .

Conditions 1 and 2 act in opposite directions. Ignoring the existing diagnosis and using only the result of the model computation improves condition 1 but breaks condition 2, and making use of the existing diagnosis alone has the opposite effect. Condition 3 means that this trade-off must be resolved within limited resources.

4. Proposed Method

The proposed method does not discard the existing diagnostic result but takes it as an initial probability, and it then updates that probability with evidence obtained from model computation under each assumed candidate. The hypotheses formed at time t are evaluated at time $t + L$, once the measurements $y(t + 1 : t + L)$ they predict have been observed. The procedure consists of four steps, namely converting the candidate set into probabilities, allocating computational resources, propagating the model for each candidate, and correcting the posterior probability, as shown in Figure 3.

4.1. Converting the Diagnostic Result into Probabilities

The existing diagnosis provides only whether each fault belongs to the candidate set and gives no ranking among the candidates. Deciding in the next step which candidate

deserves more computation, however, requires a value for each candidate. The candidate set is therefore first mapped into a probability form.

$$b_k \propto \mathbf{1}(h_k \in D) \cdot \frac{1}{1 + \sum_m \text{FSM}[m, k]} \quad (5)$$

Only faults contained in the candidate set receive positive probability, and a candidate reacting to fewer residuals is assigned a relatively higher probability. The value obtained in this way is not the final diagnostic result but an initial value for deciding where the computational resources should be spent.

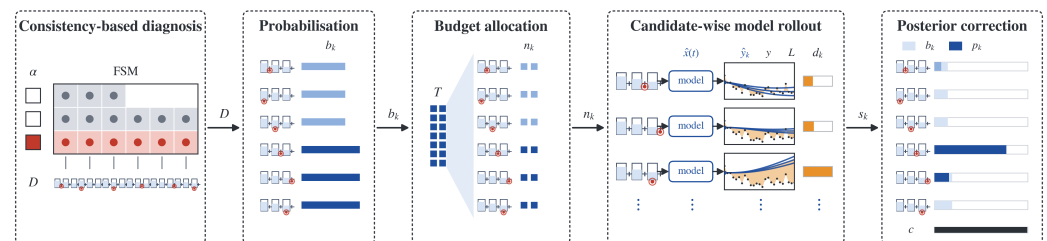


Figure 3. Structure of the proposed method. The candidate set left by the existing diagnosis is converted into probabilities, a limited amount of computation is allocated to the candidates, each candidate is propagated through the model, and the posterior probability is corrected according to the agreement with the actual measurements. All the symbols appearing in the figure are defined in the text of this section.

4.2. Allocating Computational Resources

Because the number of model computations available for one decision is finite, the probabilities obtained above are used to decide how many computations each candidate receives. The allocation is as follows.

$$n_k = \lfloor b_k T \rfloor, \quad k \in D \quad (6)$$

Here, T is the total computational budget. The allocation is restricted to the candidate set D : a hypothesis excluded by the existing diagnosis receives no computation and does not enter the comparison of Equation (8). The floor operation leaves a few computations unassigned, and these are given one at a time to the candidates with the largest b_k so that the budget is spent exactly.

4.3. Propagating the Model for Each Candidate

Once the allocation is complete, each candidate is assumed to be the true fault, and the model computes how the system will subsequently evolve. Comparing the resulting prediction with the actual measurements produces evidence about which fault hypothesis better explains the observation. For candidate h_k , the model is propagated L steps from the estimated state at instant t to obtain predicted measurements $\hat{y}_k(t+i)$, $i = 1, \dots, L$, and the mean absolute error with respect to the actual measurements is defined as follows.

$$d_k = \frac{1}{L} \sum_{i=1}^L \|\hat{y}_k(t+i) - y(t+i)\| \quad (7)$$

Because the propagation injects process noise at every step, each repetition yields a different trajectory. The n_k repetitions assigned in the previous step are therefore carried out, and d_k is taken as their average, which is where the allocation enters the procedure:

a candidate that received more repetitions obtains a less noisy estimate of d_k . Since d_k compares the prediction with $y(t+1:t+L)$, it is evaluated at instant $t+L$.

The smaller d_k is, the better the corresponding fault hypothesis explains the actual measurements. Since this value may be on a different scale for each candidate, it is not used directly; instead, the error is inverted and normalised within the set of candidates that received computation, $A = \{k : n_k > 0\}$, to give an agreement score s_k .

$$s_k = \frac{\max_{j \in A} d_j - d_k}{\max_{j \in A} d_j - \min_{j \in A} d_j}, \quad s_k = 0, \quad k \in D \setminus A \quad (8)$$

Equation (8) expresses not the absolute prediction error of a single candidate but the relative explanatory power of the candidates for the same observation. For candidates in D that receive no computation, s_k is set to zero, because no additional model-based evidence is available. When $|A| = 1$ the normalisation is not defined by a comparison, and the single propagated candidate is assigned $s_k = 1$; no inter-candidate evidence is available in that case, and the allocation of Equation (6) is designed to avoid it. The denominator also vanishes when every propagated candidate yields the same distance; the same score is then assigned to all of A , so the correction leaves the ranking of the existing diagnosis unchanged. Distributing the computation over several candidates is therefore not a choice made to improve performance but a condition for this quantity to be informative at all.

Performing the normalisation only within A also matters. Including candidates that received no computation would make their placeholder value the minimum, compressing the differences among the candidates that were actually propagated.

The propagation starts from a state estimated from the measurements rather than from the true state. In the three-tank system, p_2 and p_3 are not measured directly, so using their true values would amount to exploiting information that the comparison methods cannot access.

4.4. Correcting the Posterior Probability

The final step combines the probability from the existing diagnosis and the agreement score from the model propagation, weighting them according to how confident the diagnosis is. In this way, the instants that the diagnosis already judged correctly make use of the existing result, and only the instants where it is uncertain are corrected. Concretely, the two pieces of information are combined as follows: one is the probability provided by the existing diagnosis, and the other is the agreement obtained from the model propagation. Since the two should not carry equal weight when the existing diagnosis is already confident and when several candidates remain, the strength of the combination is adjusted according to the confidence of the diagnosis.

$$\log p_k \propto (1 - c) \log b_k + \beta(1 + \gamma c) s_k \quad (9)$$

Here, $\beta > 0$ sets how strongly the agreement score moves the probability, and $\gamma \geq 0$ sets how that strength grows as the diagnosis becomes less certain. Equation (9) is evaluated only for k such that $b_k > 0$; hypotheses with $b_k = 0$ retain $p_k = 0$. If the alarm pattern leaves $D = \emptyset$, no single fault explains the observation; the weights of Equation (5) are then taken uniform over H , and the instant is recorded as one at which the diagnosis has made no decision.

Here, c expresses how uncertain the existing diagnostic result is and controls how much the existing diagnosis is made use of and how much the additional model computa-

tion is reflected. It is defined concretely as the normalised entropy of the initial diagnostic distribution b .

$$c = - \frac{\sum_k b_k \log b_k}{\log K} \quad (10)$$

Equation (10) is the Shannon entropy [17] of b , with the convention $0 \log 0 = 0$, normalised by its maximum $\log K$ over the K hypotheses of H [18], so that $c \in [0, 1]$; the normalisation is what makes the weight $1 - c$ of Equation (9) lie in that range irrespective of how many candidates remain. When the probability is concentrated on a single candidate, c becomes small; when several candidates hold similar probabilities, c becomes large. Accordingly, when the diagnosis is already strongly confident, the existing result is made use of, and when confidence is low because several candidates remain, the result of the model propagation carries relatively more weight. This term jointly controls the trust placed in the existing diagnosis and the influence of the additional model computation. In the three-tank experiments, the measured c was 0.672 over the interval where the diagnosis selected a different candidate and 0.307 over the interval where it selected the true fault.

An epistemic measure such as a Dirichlet concentration parameter or the predictive variance of a Bayesian neural network is read from a trained model, and no such model is available here. The entropy of Equation (10) is computed from b alone and therefore demands nothing beyond what Equation (5) already provides.

The two quantities combined in Equation (9) are of different kinds: b carries the structural information already available before the propagation, whereas s is formed from the measurements that arrive during the window, a distinction that corresponds to the epistemic and aleatoric components in the uncertainty quantification literature [19]. In this way, the existing diagnosis is not entirely replaced; additional model information is used to adjust the ranking of the candidates only for the part that the existing diagnosis could not decide.

Consolidating the four steps gives Algorithm 1. At every decision instant, the existing diagnostic result is mapped into probabilities, the computational budget is divided among the candidates, the prediction under each assumed candidate is compared with the actual measurements to obtain an agreement score, and that score is combined with the existing probability to yield the final posterior probability.

Algorithm 1: Posterior correction using candidate-wise model propagation

Input :Candidate set D at instant t , system model, measurements $y_{1:t+L}$, budget T , horizon L , coefficients β, γ

Output :Corrected posterior probability p for instant t , issued at $t + L$

- 1 $b \leftarrow$ convert D into probabilities by Equation (5)
- 2 $c \leftarrow$ Equation (10)
- 3 **for** $k \in D$ **do**
- 4 $n_k \leftarrow$ Equation (6)
- 5 **end**
- 6 allocate the $R = T - \sum_k n_k$ remaining computations to the R candidates with the largest b_k , one each
- 7 $A \leftarrow \{k : n_k > 0\}$
- 8 $\hat{x}(t) \leftarrow$ state estimator($y_{1:t}$)
- 9 **for** $k \in A$ **do**
- 10 **for** $r = 1, \dots, n_k$ **do**
- 11 $\hat{y}_k^{(r)} \leftarrow$ propagate the model L steps from $\hat{x}(t)$ assuming h_k
- 12 **end**
- 13 $d_k \leftarrow$ average the n_k trajectories, each by Equation (7)
- 14 **end**
- 15 $s \leftarrow$ normalise within A by Equation (8)
- 16 $p \leftarrow$ normalise Equation (9)
- 17 **return** p

5. Theoretical Analysis

This section formalises the argument of Section 3 and establishes properties of the quantities used by the method. Proposition 1 states that an indistinguishable set moves as a single block in the candidate decision, and Corollary 2 transfers this to the weights of Equation (5). Proposition 2 states what structural information those weights use. Proposition 3 bounds the loss incurred by the floor operation in Equation (6), and Proposition 4 bounds the propagation error of Equation (7) in the horizon L . Proposition 5 shows that Equation (8) preserves the order of the propagation distances. Proposition 6 identifies Equation (9) as a tempered geometric combination of the initial weights and the rollout evidence, and Proposition 7 shows that the candidate selected by Equation (9) inside an indistinguishable set does not depend on β or γ . Proposition 8 gives a sufficient condition, on the separation of the candidate predictions, under which the true fault is selected.

Throughout, $S \subseteq H$ denotes a set satisfying Equation (4), $D = D(\alpha)$ the candidate set of Equation (3), and $A = \{k : n_k > 0\}$ the set of candidates that receive computation.

Proposition 1 (Closure of an indistinguishable set). *Let $S \subseteq H$ satisfy Equation (4). Then, for every alarm vector α with $\sum_m \alpha_m \geq 1$ and every $h_i, h_j \in S$,*

$$h_i \in D(\alpha) \iff h_j \in D(\alpha).$$

Proof. Assume $h_i \in D(\alpha)$. By Equation (3), $FSM[m, i] = 1$ for every m with $\alpha_m = 1$. By Equation (4), we have $FSM[m, j] = FSM[m, i]$ for every m , hence $FSM[m, j] = 1$ for every m with $\alpha_m = 1$, and therefore $h_j \in D(\alpha)$. Exchanging i and j gives the converse. $\square$

Corollary 1. *Let α satisfy $\sum_m \alpha_m \geq 1$. If one member of S belongs to $D(\alpha)$ then $|D(\alpha)| \geq |S|$, so for $|S| \geq 2$, the candidate set does not have size one. Moreover, any ranking that is a function of α and of the signature columns takes the same value on every member of S , since those columns coincide. The proof uses only the signature columns, so the conclusion is unchanged if the thresholds are adjusted or the alarm pattern is post-processed, provided at least one residual still alarms.*

Proof. By Proposition 1, every member of S belongs to $D(\alpha)$ whenever one of them does, so $S \subseteq D(\alpha)$. By Equation (4), the columns $FSM[:, i]$ and $FSM[:, j]$ are equal for all $h_i, h_j \in S$, so a function of α and of those columns cannot take different values on h_i and h_j . $\square$

Corollary 2. *Let $h_0 \notin S$, which holds whenever $|S| \geq 2$ since the column of h_0 is zero. Then, for every $h_i, h_j \in S$ the weights of Equation (5) satisfy $b_i = b_j$, so b is constant on S .*

Proof. If $\sum_m \alpha_m = 0$ then $D = \{h_0\}$ and, since $h_0 \notin S$, $b_i = b_j = 0$ for every $h_i, h_j \in S$. Otherwise, by Proposition 1, the indicators $\mathbf{1}(h_i \in D)$ and $\mathbf{1}(h_j \in D)$ coincide, and by Equation (4), the column sums coincide: $\sum_m FSM[m, i] = \sum_m FSM[m, j]$. Both factors of Equation (5) therefore agree, and the normalising constant is common to all hypotheses. $\square$

Remark 1. *The proof of Corollary 2 uses only that the two column sums coincide. Replacing the weight $1/(1 + \sum_m FSM[m, k])$ in Equation (5) by $f(\sum_m FSM[m, k])$ for any strictly positive function f leaves the conclusion unchanged. Equation (5) therefore induces no ordering within S for any such choice, which is the sense in which the pattern of threshold crossings alone cannot separate the set.*

Proposition 2 (Properties of the initial weights). *Let $D \neq \emptyset$ and let b be given by Equation (5). Then*

1. $b_k = 0$ for every $h_k \notin D$, and $\sum_k b_k = 1$;

2. For $h_k \in D$, b_k depends on the column $\text{FSM}[:, k]$ only through the sum $\sum_m \text{FSM}[m, k]$. In particular, two members of D whose columns have the same number of non-zero entries receive the same weight whether or not their columns coincide. Membership in D itself is decided by Equation (3) and depends on the whole column.

Proof. Both statements are immediate from Equation (5): the indicator $\mathbf{1}(h_k \in D)$ gives the first part together with the normalisation, and on D , that indicator equals one, so the remaining factor is a function of $\sum_m \text{FSM}[m, k]$ alone. $\square$

Remark 2. The second part of Proposition 2 states exactly how much of the signature structure Equation (5) uses—the breadth of a column, and not which residuals the fault reacts to. The weights are therefore a structural initialisation for the allocation of Equation (6) and not an estimate of the probability that h_k has occurred; the quantity $\sum_m \text{FSM}[m, k]$ carries no such information. The underlying design rule is that a fault reacting to more residuals is compatible with more alarm patterns, so that an observation constrains it less and it is given a smaller share of the budget. Equation (5) is one function satisfying that rule and is not claimed to be optimal. By Corollary 2, the choice does not affect the decision inside an indistinguishable set, since the weights are equal there for any function of the column sum; it affects only how the budget of Equation (6) is divided between an indistinguishable set and the candidates outside it.

Proposition 3 (Rounding loss of the allocation). Let n_k be the allocation of Equation (6) after the residual $R = T - \sum_{k \in D} \lfloor b_k T \rfloor$ units have been distributed, at most one to each candidate. Then,

$$0 \leq R \leq |D| - 1, \quad \left| \frac{n_k}{T} - b_k \right| \leq \frac{1}{T} \quad \text{for every } k \in D.$$

Proof. For each k , $b_k T - 1 < \lfloor b_k T \rfloor \leq b_k T$. Summing over $k \in D$ and using $\sum_{k \in D} b_k = 1$ gives $T - |D| < \sum_{k \in D} \lfloor b_k T \rfloor \leq T$, so $0 \leq R < |D|$; since R is an integer, $R \leq |D| - 1$. Each candidate receives at most one additional unit, hence $\lfloor b_k T \rfloor \leq n_k \leq \lfloor b_k T \rfloor + 1$ and therefore $b_k T - 1 < n_k \leq b_k T + 1$. Dividing by T gives the second claim; the upper end is attained when $b_k T$ is an integer and k receives one of the residual units. $\square$

Remark 3. The second bound does not depend on $|D|$. At the reference budget $T = 14$, it reproduces the target proportions to within $1/14 \approx 0.0715$.

Remark 4 (Cost of one decision instant). The allocation of Equation (6) spends exactly $\sum_{k \in D} n_k = T$ propagations, each of L steps, so one decision instant costs $O(\text{TLC})$, where C is the cost of evaluating a single model step, and the working memory is $O(n_x + L)$ with n_x being the state dimension. The factor C is model-dependent and is not in general quadratic in n_x : for a dense model that is linear in the state, a step is a matrix–vector product, and $C = O(n_x^2)$, whereas the three-tank model used here evaluates a fixed sparse set of scalar expressions, for which C is linear in n_x . The groups of propagations belonging to different candidates are mutually independent and can be evaluated in parallel, in which case the latency is that of a single group of $\max_k n_k$ propagations.

When the fault magnitude is unknown and a grid of G candidate magnitudes is searched inside each propagation, the cost becomes $O(G \text{ TLC})$, which is linear in G , while the budget T itself is unchanged, since Equation (6) distributes the budget over fault hypotheses, whereas the magnitude search takes place within the propagation of a single hypothesis.

These costs were also measured directly at the reference setting $L = 25$, $T = 14$, in single-threaded Python 3.12.3 with NumPy 2.5.1 on an AMD Ryzen Threadripper PRO 5995WX. One propagation took 0.41 ms, and one decision instant 5.75 ms, which is 5.8% of the 100 ms control period used in the experiments. With the five-point magnitude grid, the same decision

took 28.76 ms, or 28.8% of the control period, a factor of 5.0 that matches the linear dependence on G stated above. Evaluating the candidate groups in parallel reduced the latency to 2.05 ms, which exceeds the $\max_k n_k$ propagations of a single group because the state estimate, the allocation, and the normalisation are carried out once per instant regardless. The implementation is a direct transcription of Algorithm 1 and is not optimised, so these figures are an upper bound on the cost attainable for this model.

These figures scale with the plant through C alone. The budget T and the horizon L are design variables that do not grow with the state dimension, so the cost of one decision instant follows the cost of a single model evaluation rather than the number of faults or of candidates. A hard real-time guarantee lies outside what is established here. The propagated predictions are compared against measurements that arrive over the following L steps, so a decision formed at instant t is issued at $t + L$, and the horizon fixes that latency. The method is accordingly placed in a monitoring layer running alongside the control loop, where the decision is issued on the horizon rather than within a control period.

Proposition 4 (Propagation error bound). *Let the plant be linear in the state, $x(t+1) = Ax(t) + Bu(t) + w(t)$ with $\|w(t)\| \leq \bar{w}$ and $y(t) = Cx(t)$. Let the propagation of Equation (7) be $\hat{x}(t+i) = A\hat{x}(t+i-1) + Bu(t+i-1) + \hat{w}(t+i-1)$ with the injected process noise bounded by the same $\|\hat{w}(t)\| \leq \bar{w}$ and the propagation carried out under the hypothesis that generated the data, and let $e(0) = \hat{x}(t) - x(t)$ be the initial state estimation error. If $\rho(A) < 1$, so that $\|A^i\| \leq \kappa \rho^i$ for some $\kappa > 0$ and $\rho \in (\rho(A), 1)$. Then, for every $i \leq L$*

$$\|\hat{y}(t+i) - y(t+i)\| \leq \kappa \|C\| \rho^i \|e(0)\| + 2\kappa \|C\| \bar{w} \frac{1 - \rho^i}{1 - \rho},$$

and, since $\rho^i \leq 1$ and $1 - \rho^i \leq 1$ for every $i \geq 0$, the right-hand side is bounded above by $\kappa \|C\| (\|e(0)\| + 2\bar{w}/(1 - \rho))$ uniformly in i . For a noise-free propagation, $\hat{w} \equiv 0$, the factor $2\bar{w}$ is replaced by $\bar{w}$.

Proof. The error $e(i) = \hat{x}(t+i) - x(t+i)$ satisfies $e(i) = Ae(i-1) + \hat{w}(t+i-1) - w(t+i-1)$, so $e(i) = A^i e(0) + \sum_{j=0}^{i-1} A^j (\hat{w}(t+i-1-j) - w(t+i-1-j))$. Taking norms and applying $\|A^j\| \leq \kappa \rho^j$ and $\|\hat{w} - w\| \leq 2\bar{w}$ together with the triangle inequality gives $\|e(i)\| \leq \kappa \rho^i \|e(0)\| + 2\kappa \bar{w} \sum_{j=0}^{i-1} \rho^j$. Summing the geometric series and multiplying by $\|C\|$ yields the stated bound; bounding ρ^i and $1 - \rho^i$ by one gives the uniform bound. $\square$

Remark 5. For the three-tank model, the discretised transition matrix has spectral radius $\rho(A) = 0.9804$, so the hypothesis of Proposition 4 is satisfied and the prediction error does not grow without bound in L . The decay over the reference horizon is nevertheless modest, at $\rho^{25} \approx 0.609$, so the initial estimation error is not negligible at $t + L$. Its effect on the comparison is limited instead by the fact that every candidate rollout in Algorithm 1 starts from the same $\hat{x}(t)$: by Corollary 4 the decision depends on the differences $d_k - d_l$. Writing A_k for the transition matrix propagated under hypothesis h_k , the predictions differ at step i by $C(A_k^i - A_l^i)\hat{x}(t)$ plus the fault terms, so what separates them is governed by $\|C(A_k^i - A_l^i)\|$, which is small precisely when the two fault hypotheses induce similar dynamics. Since d_k is a mean of absolute errors, the shared initial error does not cancel exactly in $d_k - d_l$. Rather, the same initial error is propagated differently under A_k and A_l , and its differential effect is governed by $C(A_k^i - A_l^i)e(0)$.

Proposition 5 (Order preservation). *Assume $\max_{j \in A} d_j > \min_{j \in A} d_j$. Then for all $k, l \in A$,*

$$s_k > s_l \iff d_k < d_l,$$

and in particular $\arg \max_{k \in A} s_k = \arg \min_{k \in A} d_k$.

Proof. Write $d_{\max} = \max_{j \in A} d_j$ and $d_{\min} = \min_{j \in A} d_j$. By Equation (8), $s_k = (d_{\max} - d_k) / (d_{\max} - d_{\min})$ with $d_{\max} - d_{\min} > 0$, which is a strictly decreasing affine function of d_k . $\square$

Remark 6. Proposition 5 uses only strict monotonicity. Replacing Equation (8) by any strictly decreasing map of d_k , such as a softmax of $-d_k$ leaves $\arg \max_{k \in A} s_k$ unchanged.

Proposition 6 (Equation (9) as a tempered geometric combination). Let q be the distribution on D given by $q_k = \exp(s_k) / \sum_{j \in D} \exp(s_j)$ for $k \in D$, with $s_k = 0$ for $k \in D \setminus A$ as in Equation (8), and set $w_1 = 1 - c$, $w_2 = \eta(c) = \beta(1 + \gamma c)$, $\tau = w_1 + w_2$ and $\lambda = w_2 / \tau$. Then, Equation (9) is equivalent to

$$p \propto \left(b^{1-\lambda} q^\lambda \right)^\tau.$$

It then follows that p is the point at parameter λ on the exponential geodesic joining b and q , tempered with inverse temperature τ .

Proof. Writing $Z = \sum_{j \in D} \exp(s_j)$, which does not depend on k , we have $\log q_k = s_k - \log Z$ and hence $w_2 s_k = w_2 \log q_k + w_2 \log Z$. Substituting into Equation (9) and absorbing the terms that are constant in k ,

$$\log p_k = w_1 \log b_k + w_2 \log q_k + \text{const} = \tau [(1 - \lambda) \log b_k + \lambda \log q_k] + \text{const},$$

which is the stated form after normalisation. $\square$

Remark 7. Proposition 6 places Equation (9) in three standard settings at once. Both b and q are members of the exponential family over the finite hypothesis set, and a weighted geometric combination of two such members remains in that family, so p does as well. The bracket is the logarithmic opinion pool of b and q [20], which in information geometry is the point at parameter λ on the e -geodesic joining them [21]. Setting $w_1 = w_2 = 1$ gives $p \propto bq$, which has the product form of a Bayesian update if q is interpreted as a normalised evidence term. This algebraic correspondence does not imply Bayesian calibration; Proposition 6 characterises the form of the update rather than its calibration properties. Equation (9) generalises this form by varying the mixing parameter λ and by tempering, both of which are standard in generalised Bayesian updating [22].

The mixing parameter admits a direct reading. Since $\lambda = 1 / ((1 - c) / \eta(c) + 1)$ and $(1 - c) / \eta(c)$ is decreasing in c for every $\beta > 0$ and $\gamma \geq 0$, the parameter λ increases with c : the less certain the existing diagnosis, the further along the geodesic towards the rollout evidence the corrected posterior is placed. This is the role of Equation (10) in the combination.

Proposition 7 (Invariance of the decision in the combination coefficients). Let $S' \subseteq A$ be a subset on which b is constant. Then, for every $\beta > 0$, $\gamma \geq 0$ and every $c \in [0, 1]$,

$$\arg \max_{k \in S'} p_k = \arg \max_{k \in S'} s_k,$$

where p is given by Equation (9).

Proof. For $k \in S'$ the term $(1 - c) \log b_k$ takes the same value, say μ . By Equation (9), $\log p_k = \mu + \beta(1 + \gamma c) s_k$ up to an additive constant common to all hypotheses. Since $\beta > 0$, $\gamma \geq 0$ and $c \geq 0$, the coefficient $\beta(1 + \gamma c)$ is strictly positive, so $k \mapsto \log p_k$ is a strictly increasing affine function of s_k on S' . $\square$

Corollary 3 (Scope of the combination coefficients). By Corollary 2, b is constant on $S \cap A$. Thus, the candidate that Equation (9) selects within $S \cap A$ is independent of β and γ . At an instant whose maximum-probability candidate over D lies in $S \cap A$ and stays there as the coefficients vary, the contribution to R_{det} is therefore unaffected by them. Where the candidate set is not contained in an indistinguishable set, as on the three-tank plant, the conclusion does not follow from this corollary; Remark 8 records what is observed there.

Remark 8. Proposition 7 concerns the ordering of p , not its values. The metrics P_{ih} and P_{nm} are probability masses assigned to the true fault, and β and γ do change them: β sets how strongly the agreement score moves the probability, and γ sets how that strength grows as the diagnosis becomes less certain. The balance they set is the trade-off between conditions 1 and 2 of Section 3. Where the candidate set extends beyond an indistinguishable set, the prior is not constant and that balance enters the ranking; on a set where it is constant, the hypothesis of Proposition 7 holds, and the two coefficients move the probability mass while the selection stays where Proposition 5 puts it. The ablation of the entropy term isolates this effect: fixing $c = 0$, which removes the γ pathway, lowers P_{ih} from 0.6735 to 0.6060 while raising P_{nm} from 0.8274 to 0.8523. The flatness of the coefficient grid over $\beta \in [3, 24]$ and $\gamma \in [0, 16]$ extends beyond the instants Corollary 3 covers, and the reason is the scale on which the two terms meet: with the weights of Equation (5) the spread of $(1 - c) \log b_k$ across the candidate set is at most 0.19, whereas $\beta(1 + \gamma c)$ is at least 3 everywhere on the grid, so the selected candidate is $\arg \max_k s_k$ throughout it, while the probability mass continues to move.

Corollary 4. Combining Proposition 5 and Proposition 7, for $S \cap A$ with $|S \cap A| \geq 2$ and $\max_{j \in A} d_j > \min_{j \in A} d_j$,

$$\arg \max_{k \in S \cap A} p_k = \arg \min_{k \in S \cap A} d_k.$$

Within an indistinguishable set, the decision is therefore determined by the propagation of Equation (7) alone.

Proposition 8 (Separation condition). Let the plant satisfy the hypotheses of Proposition 4, let S be an indistinguishable set, let b be constant on $S \cap A$, and let $\hat{y}_k$ denote the noise-free prediction of Equation (7) under hypothesis h_k , that is the propagation with $\hat{w} \equiv 0$. Write

$$\sigma_{kl} = \frac{1}{L} \sum_{i=1}^L \|\hat{y}_k(t+i) - \hat{y}_l(t+i)\|, \quad \varepsilon = \kappa \|C\| \left(\|e(0)\| + \frac{\bar{w}}{1-\rho} \right) + \bar{v},$$

with $\bar{v}$ a bound on the measurement noise and the second quantity the uniform bound of Proposition 4. If the true fault is h_k , $h_k \in S \cap A$, and

$$\sigma_{kl} > 2\varepsilon \quad \text{for every } l \in S \cap A, l \neq k,$$

then $d_k < d_l$ for every such l , and consequently $\arg \max_{j \in S \cap A} p_j = h_k$ for every $\beta > 0$, $\gamma \geq 0$ and $c \in [0, 1]$.

Proof. Since the measurements are generated by h_k , Proposition 4 together with the measurement noise bound gives $\|\hat{y}_k(t+i) - y(t+i)\| \leq \varepsilon$ for every $i \leq L$, hence $d_k \leq \varepsilon$. For $l \neq k$, the triangle inequality gives $\|\hat{y}_l(t+i) - y(t+i)\| \geq \|\hat{y}_l(t+i) - \hat{y}_k(t+i)\| - \varepsilon$, so $d_l \geq \sigma_{kl} - \varepsilon > \varepsilon \geq d_k$. The conclusion on $\arg \max p$ follows from Corollary 4. $\square$

Remark 9 (Scope of the guarantee). Proposition 8 is a sufficient condition on the separation of the predicted trajectories relative to the combined estimation and noise level, and it is stated on a set on which b is constant. That restriction costs nothing on the sets this method addresses: by

Corollary 2, every indistinguishable set carries a constant b , so the condition applies to each of them as it stands. The scale on which the two terms of Equation (9) meet is worth recording alongside it. By Equation (8), the agreement score satisfies $s_k \in [0, 1]$, so the evidence term is bounded above by $\beta(1 + \gamma)$. The prior term is bounded as well, since Equation (5) gives $b_k \geq 1 / ((1 + M)|D|)$ for every $k \in D$ with M the number of residuals; once a residual alarms h_0 leaves the candidate set, on the three-tank plant $|D| \leq 6$ and $b_k \geq 0.028$, while at the evaluated instants $b_k \geq 0.2174$ and $(1 - c)|\log b_k| \leq 1.53$. The two terms therefore meet on scales that the plant fixes, and on this plant, the evidence term is the larger by more than an order of magnitude at the reference coefficients. Proposition 8 is stated on the noise-free propagation; the quantity Algorithm 1 forms averages n_k realisations of it and approaches it as n_k grows.

The quantity σ_{kl} grows with the fault magnitude θ , since the two predictions differ only through the fault terms, while ε grows with $\bar{w}$ and $\bar{v}$. The condition is therefore a statement about both quantities rather than about θ alone. Appendix B reports the measured behaviour across twelve combinations of θ and σ_w . The proposed method is above the baseline in every one of them, and the ratio θ/σ_w separates the twelve without error at 12.5: the six conditions at or above that value gain between +0.34 and +0.42, and the six at or below 10 gain at most +0.15. That value is an empirical boundary for this plant and is not implied by Proposition 8, whose condition is stated on σ_{kl} , a quantity that depends on which pair of hypotheses is being separated and not on θ alone. The agreement between the two is therefore evidence that σ_{kl} scales with θ on this plant, not a derivation of the threshold.

6. Experimental Setup

The comparison methods were implemented so that each prior method can also achieve its original performance on the present plant. The architecture, hyperparameters, preprocessing and training procedure of each method follow the original paper, and the inputs are formed from the residual set of the three-tank system. The detailed settings of each method are summarised in Appendix A.

6.1. Experimental Environment

The experiments use the three-tank model included in the Fault Diagnosis Toolbox [4]. Three tanks are connected in series through valves, and six faults are considered, namely valve leaks fV_1 through fV_3 and tank leaks fT_1 through fT_3 . Including the no-fault case, there are seven fault classes in total. The measurements are the pressure p_1 of the first tank and the flow q_2 through the second valve, while p_2 and p_3 are not measured directly. The inflow q_0 enters as a known input. Structural analysis produced five residuals, under which fV_2 , fV_3 , and fT_3 share an identical fault signature.

Each simulation runs for 400 steps of 100 ms with a single fault introduced at step 150. The fault magnitude was set to 0.15, the standard deviation of the process noise to 0.01, and that of the measurement noise to 0.02. The residual window was $W = 20$ steps. Evaluation starts 40 steps after fault onset and uses 12 seeds and 7 fault classes.

The model propagation of the proposed method starts from a state estimated from the measurements. Since the three-tank model is linear in the state, a Kalman filter was used for state estimation. Because this filter integrates the state with the nominal model, the estimate diverges from the true value once a fault occurs; the measured error was 0.4% of the state magnitude for the directly measured p_1 and 2.7% and 5.5% for the unmeasured p_2 and p_3 , respectively.

The main experiments use a reference setting selected on validation seeds separate from those used for reporting. The computation horizon was set to 25 steps and the computational budget to 14, with combination coefficients $\beta = 6$ and $\gamma = 8$. The grounds for selecting each design variable are summarised in Appendix C. The fault magnitude

is fixed at 0.15 and taken as known in these experiments. Where it is not known, each propagation evaluates a fixed grid of five candidate magnitudes $\{0.05, 0.10, 0.15, 0.25, 0.40\}$ identically for every candidate, with the grid containing the true magnitude of these experiments. The budget T is unaffected, since Equation (6) distributes it over fault hypotheses while the magnitude search lies inside the propagation of a single hypothesis; by Remark 4, the cost of one decision instant is instead linear in the grid size, and it was measured to rise from 5.75 ms to 28.76 ms.

6.2. Comparison Methods

The comparison methods comprise two families, model-based and learning-based, as listed in Table 1.

Table 1. Methods used for comparison, with their sources and fault-data requirements.

Family	Method	Source	Fault Data
Model-based	Consistency-based diagnosis	Frisk et al. [4]	Not required
Model-based	Residual sensitivity-based isolation	Murata et al. [9]	Not required
Model-based	Uncertainty-based diagnosis	Mohammadi et al. [8]	Nominal only
Learning-based	BG-CNN, ANN, RF, KNN, SVM	Dash et al. [12]	Required

For consistency-based diagnosis, it was verified that the candidate sets agree with the reference implementation of the Fault Diagnosis Toolbox over all $2^5 = 32$ possible alarm patterns. The convolutional neural network of Dash et al. [12] consists of 32 kernels of size 2×2 , 2×2 max pooling with stride 2, and a fully connected layer of 128 nodes, trained for 200 epochs with the Adam optimiser at a learning rate of 0.001. The accompanying ANN uses hidden layers of 128-64-32, RF uses $\text{max_depth} = 50$ and $\text{n_estimators} = 20$, KNN uses $\text{n_neighbors} = 2$, and SVM uses an RBF kernel with $C = 1$.

The number of input residuals and the number of fault classes are formed to match the present plant. All methods are evaluated on the same three-tank plant.

6.3. Evaluation Metrics

The evaluation uses two metrics that separate the case where the existing diagnosis missed the true fault from the case where it identified it, together with an auxiliary metric indicating whether the correction actually reversed the decision. This is because, for structurally indistinguishable faults, plain accuracy does not adequately express the additional discriminative ability. The metrics of Table 2, corresponding to the three conditions of Section 3, were therefore used.

P_{ih} and P_{nm} indicate how the final probability is maintained when the existing diagnosis missed and selected, respectively, the true fault, and R_{det} indicates how much the proposed method corrected the cases in which the existing diagnosis selected wrongly. The metrics are defined as follows. At each evaluation instant, the instants are divided into two sets according to whether the maximum-probability candidate of the distribution b output by the existing diagnosis coincides with the true fault. Let I denote the set of instants where it does not coincide and N the set where it does. The set I consists of the instants at which the existing diagnosis pointed to a candidate other than the true fault, and it is the target of the correction in this work.

Table 2. Evaluation metrics. Arrows indicate that higher values are better.

Metric	Definition	Direction
P_{ih}	Mean of the corrected probability p assigned to the true fault over instants in I	↑
P_{nm}	Mean of the corrected probability p assigned to the true fault over instants in N	↑
R_{det}	Fraction of instants in I at which the true fault became the maximum-probability candidate after correction	↑

P_{ih} corresponds to condition 1 of Section 3 and P_{nm} to condition 2. Both are probability masses averaged over instants and over all seven faults, whereas the accuracies reported later are top-1 decisions for one fault at a time; the two families of numbers are therefore not directly comparable. Performance comparison between methods uses these two metrics together with the diagnostic accuracy.

R_{det} is not a metric for comparing methods but an auxiliary metric expressing the corrective ability of the proposed method. Its denominator is the size of I and its numerator is the number of instants in I at which the true fault became the top candidate after correction, so $R_{det} = 0.686$ means that the true fault returned to the top rank at 68.6% of the instants where the existing diagnosis had pointed elsewhere. It should be noted that for methods performing no correction, we have $p = b$, and since I is defined by $\arg \max(b)$, the true fault can never become the top candidate; R_{det} is therefore not defined for such methods and is left blank in the tables. This metric therefore does not indicate superiority over existing methods and is used only to show how much the proposed method actually reversed the decisions of the existing diagnosis.

The evaluation was restricted to instants at which the maximum probability of the existing diagnosis exceeds 0.25, since below that value, the diagnosis has effectively made no decision. Applying the criterion to that probability rather than to each method's own output ensures that all compared methods are evaluated on the same set of instants. For the comparison over all seven faults, top-1 accuracy was used. Statistical tests were carried out as paired t -tests with the seed as the pairing unit.

7. Experimental Results

The proposed method substantially raised the probability of the true fault at the instants where the existing diagnosis missed it, and this change led to an improvement in the actual decisions both on the indistinguishable set and on the full fault set. The following presents in turn the change in probability, the change in decisions, the role of each component, and the comparison with approaches that use fault data.

7.1. Comparison with the Existing Diagnosis

The proposed method raised the probability of the true fault from 0.2547 to 0.6735 in the situation where the existing diagnosis selected a different candidate. For structurally indistinguishable faults, the ranking among candidates is itself the object of evaluation, so this change in value directly expresses the discriminative ability that has been added. Table 3 compares the existing model-based diagnosis methods with the proposed method.

The differences in P_{ih} against consistency-based diagnosis and against the residual sensitivity-based method were 0.4188 and 0.4483, respectively, with both being significant ($p < 0.001$). Over the interval where the diagnosis had already selected the true fault, the values were also higher by 0.1882 and 0.1998, so the two regimes were improved together rather than traded against each other. The uncertainty-based diagnosis achieved the highest P_{nm} of 0.9983, but this results from withholding a decision at most instants, and its P_{ih} was only 0.0080. The gap $|P_{nm} - P_{ih}|$ makes this difference clear: the existing methods show gaps ranging from 0.38 to 0.99, indicating a sharp divide between the regimes where the

diagnosis was right and wrong, whereas the proposed method has the smallest gap at 0.1539 and therefore behaves evenly across both.

The proposed method restored the true fault to the top rank at 68.6% of these instants, so the increase in P_{ih} did not remain a change in probability values but carried through to the final decision.

Table 3. Comparison with model-based methods. P_{ih} and P_{nm} are probability masses assigned to the true fault, not top-1 decision rates; they are computed on a different basis from the accuracies of Table 4 and are not directly comparable with them. Values are the mean and standard deviation over 12 seeds, and the best value in each column is shown in bold. Arrows indicate the direction of improvement: $\uparrow$ means higher is better and $\downarrow$ means lower is better. $|P_{nm} - P_{ih}|$ is the gap between the two regimes, and R_{det} is not defined for methods that perform no correction. The highest P_{nm} of the uncertainty-based diagnosis results from withholding a decision at most instants, as confirmed by the gap column. The last row reports the setting in which the fault magnitude is not known, and a fixed grid of five candidate magnitudes is searched inside each propagation, which is the setting that applies in deployment.

Method	$P_{ih} \uparrow$	$P_{nm} \uparrow$	$ P_{nm} - P_{ih} \downarrow$	$R_{det} \uparrow$
Consistency-based diagnosis	0.2547 $\pm$ 0.0159	0.6392 $\pm$ 0.0235	0.3845	—
Residual sensitivity-based isolation	0.2252 $\pm$ 0.0101	0.6276 $\pm$ 0.0151	0.4024	—
Uncertainty-based diagnosis	0.0080 $\pm$ 0.0030	0.9983 $\pm$ 0.0000	0.9903	—
Proposed method	0.6735 $\pm$ 0.1055	0.8274 $\pm$ 0.0337	0.1539	0.686
Proposed method (fault magnitude unknown)	0.6513 $\pm$ 0.0819	0.8169 $\pm$ 0.0358	0.1656	0.654

Table 4. Accuracy over the full fault set. The overall accuracy is a top-1 decision rate over the full fault set; it is computed on a different basis from the probability masses of Table 3 and is not directly comparable with them. Values are sample-weighted over 1040 evaluated instants; since the number of instants differs across faults, they do not equal the unweighted mean of the per-fault accuracies in Table 5.

Method	Overall Accuracy $\uparrow$	Indistinguishable Set $\uparrow$	Other Four Faults $\uparrow$
Consistency-based diagnosis	0.5942	0.3148	0.7538
Proposed method	0.7673	0.5265	0.9048
Difference	+0.1731	+0.2116	+0.1511

Table 5. Isolation accuracy per fault. The best value in each column is shown in bold. The higher value of the consistency-based diagnosis on fV_2 results from tie breaking, as confirmed in Table 6.

Method	NF	f_{V1}	f_{T1}	f_{T2}	f_{V2}	f_{V3}	f_{T3}
Consistency-based diagnosis	1.000	0.089	1.000	0.706	0.983	0.000	0.000
Proposed method	1.000	0.605	1.000	0.918	0.099	0.460	0.977

Table 6. Verification of tie breaking. Only the order of the fault list was reversed, leaving the probability values unchanged.

Item	Original Order	Reversed Order
Max–min difference of the three probabilities		0.000000
Fraction of instants sharing the maximum		98.7%
Selected fault	f_{V2}	f_{T3}
Fraction of instants at which it was selected	98.7%	98.7%

These results show that additional model computation can correct the ranking of the candidates when the existing diagnosis has selected a wrong candidate. The following examines in which fault candidates this difference actually appears.

7.2. Isolation Results on Indistinguishable Faults

The isolation accuracy on the indistinguishable set improved from 31.5% to 52.7%, that is, the preceding change in probability carried through to a change in the actual fault decision. To confirm this, the decisions were examined for the three faults fV_2 , fV_3 , and fT_3 alone.

Under the existing consistency-based diagnosis the accuracies of these three faults were 0.983, 0.000, and 0.000, respectively. Because the three share an identical fault signature, the diagnosis cannot distinguish them and concentrates its decisions on one of them. As a result, even when fV_3 or fT_3 actually occurred, the decision was fV_2 in almost every case.

Under the proposed method, the accuracies of the same three faults became 0.099, 0.460, and 0.977. Decisions appeared for the two faults that had never been selected by the existing diagnosis, and fT_3 in particular was correctly isolated in almost all samples. This follows from comparing the future behaviour under each assumed candidate, which creates a new ranking among candidates that share an identical residual signature.

Whether this change is biased towards a particular fault is examined together with the full fault set in the next subsection.

7.3. Isolation Results on the Full Fault Set

The effect of the proposed method was not confined to the indistinguishable set. The accuracy over all seven faults increased from 0.5942 to 0.7673, and for the remaining four faults outside the indistinguishable set, it increased from 0.7538 to 0.9048. The results are given in Table 4.

The per-fault results are given in Table 5 and Figure 4.

Examining the results per fault, only the accuracy of fV_2 decreases. A separate analysis was carried out to determine whether this decrease reflects a loss of performance or arises from another cause.

By Equation (5) the probability of a candidate is determined by the number of residuals that fault reacts to. All three faults of the indistinguishable set react to four residuals, so their probabilities are equal at 0.2609, whereas fV_1 , which reacts to all five residuals, is lower at 0.2174. That the three faults share the maximum means precisely that the existing diagnosis has no basis for choosing among them.

This was confirmed empirically. As summarised in Table 6, the difference between the maximum and the minimum of the probabilities assigned to the three faults was 0.000000 at every evaluated instant, i.e., they were identical even at the level of floating-point representation. Instants at which two or more faults shared the maximum accounted for 98.7% of the total, and at those instants, the existing diagnosis pointed to fV_2 without exception.

To verify directly that tie breaking is the cause, the same experiment was repeated with the order of the fault list reversed. Although only the order was changed while the probability values remained the same, the selected fault shifted entirely from fV_2 to fT_3 . The high accuracy observed for fV_2 under the existing diagnosis is therefore not the result of identifying that fault but of the leading entry being selected under a tie, and which fault is selected is determined by the order of the list rather than by performance.

The decrease in fV_2 under the proposed method is therefore not a loss of performance but the consequence of the tie being resolved and the decisions being distributed over several candidates with similar signatures. Over the set as a whole, the accuracy improves, as reported in Table 4.

Building on these results, the role of each component of the proposed method in the performance improvement was analysed.

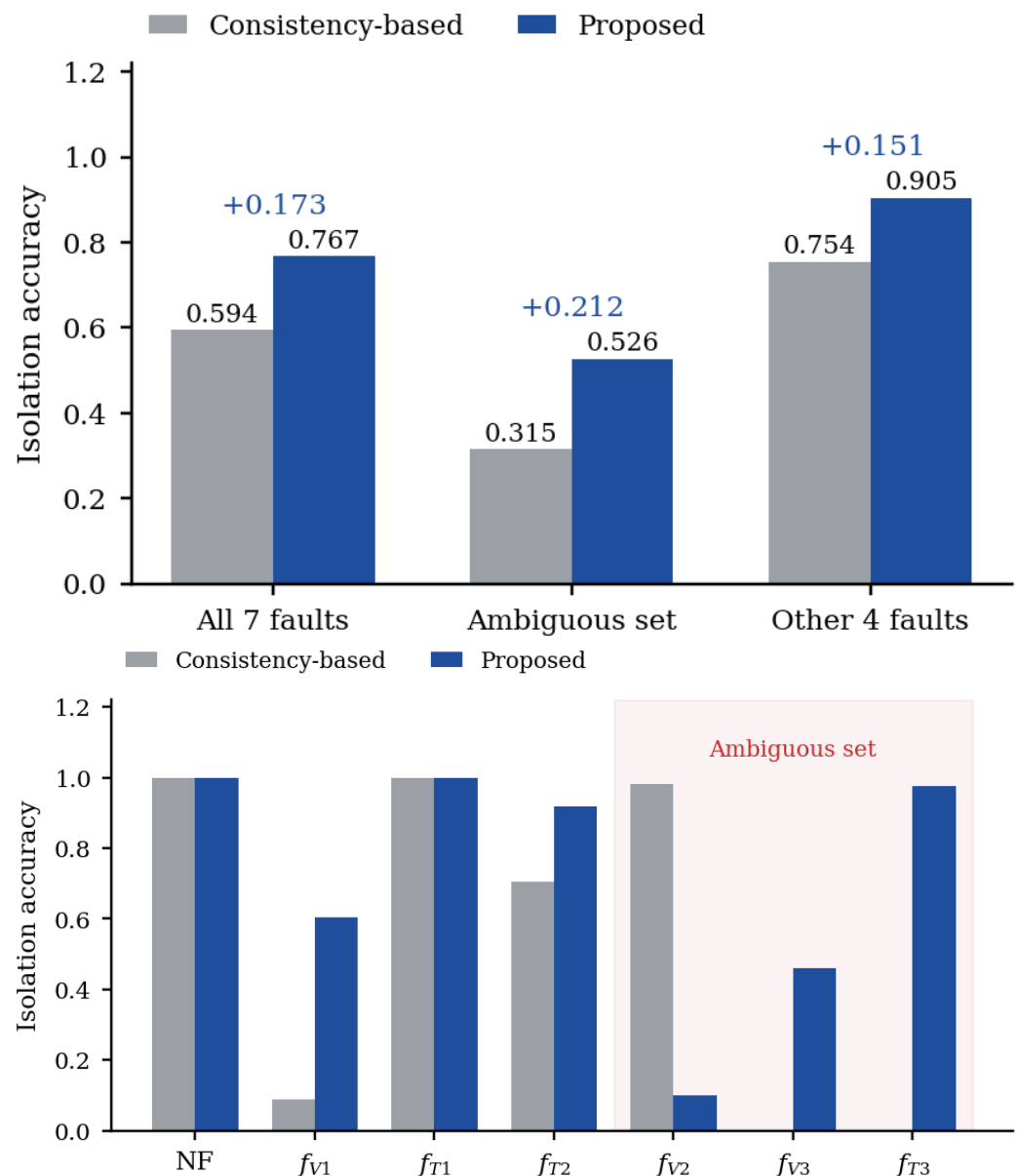


Figure 4. Accuracy over the full fault set (top) and per-fault accuracy (bottom). The red region marks the indistinguishable set.

7.4. Results by Component

The performance improvement originates in the computational allocation. In the four configurations from which the allocation was removed, P_{ih} remained at or below 0.0024, whereas the configurations including it rose to at least 0.6057. To confirm this, the three components were removed step by step, with the results given in Table 7.

Without distributed computation across multiple candidates, the agreement score carries no comparative information, which is why those four configurations cannot correct at all.

The remaining two components govern different regimes. Adding the combination with the diagnosis to the allocation raises P_{nm} from 0.8389 to 0.8523, while adding the entropy-based correction raises P_{ih} from 0.6057 to 0.6738. The former therefore acts on the

interval where the true fault was already selected and the latter on the interval where it was not.

Table 7. Results by component. The best value in each column is shown in bold. R_{det} is not defined for configurations that perform no correction, and two configurations attain the same best value.

Configuration	$P_{ih} \uparrow$	$P_{nm} \uparrow$	$R_{det} \uparrow$
None	0.0024	0.9853	—
Computational allocation	0.6057	0.8389	0.688
Combination with the diagnosis	0.0023	0.9977	—
Entropy-based correction	0.0001	0.9952	—
Allocation + combination	0.6060	0.8523	0.670
Combination + entropy correction	0.0000	1.0000	—
Allocation + entropy correction	0.6738	0.8225	0.688
Full configuration	0.6735	0.8274	0.686

The grounds for adopting the full configuration are as follows. The configuration with $P_{nm} = 1.0000$ has $P_{ih} = 0.0000$ and thus corrects nothing in the interval that was judged wrongly. The configuration with the highest P_{ih} , namely allocation with entropy correction, is statistically indistinguishable from the full configuration at 0.6738 against 0.6735 ($p = 0.213$), yet differs in P_{nm} at 0.8225 against 0.8274. The full configuration, which loses nothing in either regime, was therefore adopted.

One further variant was measured in the same context. When all computation was concentrated on a single candidate, no inter-candidate comparison was available. Although the sole propagated candidate received the maximum agreement score by convention, increasing the budget did not provide additional discriminative information, and P_{ih} remained at 0.0023.

The above concerns the case of using additional model computation. The following compares the same problem with approaches that solve it using fault data.

7.5. Comparison with Learning-Based Methods

Learning-based methods taking the same residuals as input did not exceed an isolation accuracy of 0.30 on the indistinguishable set, even with 512 training samples per fault, whereas the proposed method achieved 0.5000 without using any fault data. Figure 5 shows how the accuracy changes as the training set grows; the experimental configuration of Dash et al. [12] was followed, and Table 8 reports the results at the largest size.

Table 8. Comparison with learning-based methods at 512 training samples. Every step in the evaluation window is scored here, and the 0.25 candidate gate is not applied, so the values are computed on a different basis from those of Table 4, where the corresponding figures are 0.5942 and 0.7673.

Method	Fault Data $\downarrow$	Overall Accuracy $\uparrow$	Indistinguishable Set $\uparrow$	In-Set F1 $\uparrow$
Consistency-based diagnosis	0	0.5622	0.3272	0.4887
BG-CNN	512	0.5930	0.2531	0.5878
ANN	512	0.5922	0.2665	0.5898
RF	512	0.5926	0.2387	0.5870
KNN	512	0.5961	0.2274	0.5805
SVM	512	0.5944	0.2747	0.5656
Proposed method	0	0.7288	0.5000	0.6999

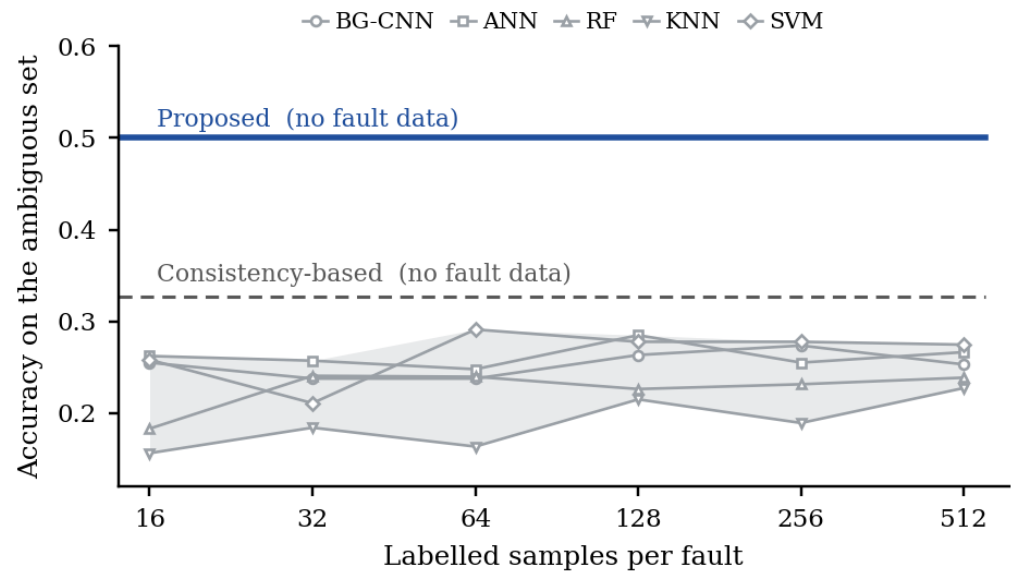


Figure 5. Isolation accuracy on the indistinguishable set as a function of the number of training samples. None of the five methods taking residuals as input reached the level of the proposed method as the number of samples increased. The shaded band spans the lowest and the highest of the five learning-based methods at each sample size.

The comparison covers learning-based methods that use only the same residual-based information. Because these methods do not produce a candidate-set probability, the gate of 0.25 used elsewhere cannot be applied to them, and Table 8 therefore evaluates every step in the window. The proposed method obtained 0.5000 in the set and 0.7288 overall. Measuring the accuracy on the training data separately gave 0.937 for BG-CNN, 0.964 for ANN, and 0.999 for RF, indicating that these models fitted the training data itself sufficiently well.

For faults sharing an identical residual signature, the values and temporal patterns of the residuals do not separate the candidates far enough for a classifier to exploit. The proposed method secures the inter-candidate differences that are hard to observe directly in the residuals through additional model computation.

All of the above results were obtained under the reference setting; results for conditions departing from it are reported in Appendix B.

7.6. Second Benchmark: Electric Motor

The experiments of Sections 6 and 7 are repeated on a second plant.

The second benchmark is the direct-current series motor distributed with the Fault Diagnosis Toolbox [4]. Its measured variables are the supply voltage and the current, speed, and torque sensors, and its faults are an increase f_R in the winding resistance together with biases f_i , f_w and f_T on the three sensors. The model is used as distributed, with no sensor removed and no fault added. Parameter values are not supplied with it, so unit values are used throughout, as in the three-tank case. Unlike the three-tank model, this one is not linear in the state: the torque is quadratic in the current, and the back electromotive force is a product of current and speed. The propagation of Equation (7) integrates these equations as they stand, and the state at the decision instant is taken from the current and speed sensors rather than from a Kalman filter.

Structural analysis of this configuration yields three MSO sets, from which the toolbox generates three residuals. All four faults are detectable, and the fault signature matrix is given in Table 9.

Table 9. Fault signature matrix of the electric motor. The columns of f_R and f_i coincide, so consistency-based diagnosis returns the two together.

Residual	Fault-Free	f_R	f_i	f_w	f_T
r_0	0	0	0	1	1
r_1	0	1	1	1	0
r_2	0	1	1	0	1

The columns of f_w and f_T differ from each other and from the remaining two, so these two faults are isolated on their own. The columns of f_R and f_i are identical, and $S = \{f_R, f_i\}$ is therefore an indistinguishable set in the sense of Equation (4). It has two elements, where the set of the three-tank benchmark has three.

The two faults of this set are quantities of different physical kinds: f_R is an increment on a resistance, and f_i is a bias on a current measurement. Giving them the same numerical magnitude would make their residual responses differ in size, and the pair would then be separated by the magnitude of the response rather than by its structure. The magnitudes are therefore matched so that the residual root mean square of each fault, after the fault-free noise floor has been removed in power, takes the same value. At the reference setting, f_R is 0.15 and f_i is 0.0872, for which the two responses are 0.0857 and 0.0854. The horizon $L = 25$, the budget $T = 14$, the ambiguity threshold 0.25, and the 12 evaluation seeds are those of Section 6. Results are given in Table 10.

Table 10. Comparison on the electric motor, over twelve seeds. The columns carry the quantities of Table 3 with the in-set top-1 accuracy in place of the gap column.

Method	$P_{ih} \uparrow$	$P_{nm} \uparrow$	In-Set $\uparrow$	$R_{det} \uparrow$
Consistency-based diagnosis	0.3523 ± 0.0491	0.8067 ± 0.0234	0.4323	—
Residual-sensitivity isolation	0.4659 ± 0.0551	0.5581 ± 0.0208	0.4323	—
Uncertainty-based diagnosis	0.4976 ± 0.0510	0.7542 ± 0.0321	0.4375	—
Proposed method	0.7110 ± 0.1145	0.9820 ± 0.0229	0.7760	0.708

The proposed method is ahead of all three comparison methods. Against the consistency-based diagnosis, the gain in P_{ih} is $+0.3587$ ($p = 1.7 \times 10^{-8}$) and in P_{nm} it is $+0.1753$ ($p = 1 \times 10^{-10}$); against residual-sensitivity isolation, it is $+0.2450$ ($p = 2.9 \times 10^{-7}$), and against the uncertainty-based diagnosis, $+0.2134$ ($p = 5.2 \times 10^{-5}$). The corresponding gain over the consistency-based diagnosis on the three-tank benchmark is $+0.4188$ for a set of three candidates, against $+0.3587$ here for a set of two. Removing the correction of Equation (9) while leaving the allocation and propagation in place lowers P_{ih} by 0.0545 ($p = 2.7 \times 10^{-3}$).

The behaviour of the consistency-based diagnosis inside the set is the same as the one examined in Section 7. Table 11 records the two probabilities of the pair over the evaluated instants together with the candidate selected when the order of the fault list is reversed.

Table 11. Treatment of ties on the electric motor, over 888 evaluation instants. The probabilities are identical in the two cases, and only the order of the fault list differs.

Quantity	Listed Order	Reversed Order
Spread of the two probabilities of the pair	0.000000	0.000000
Share of instants at which the maximum is shared	82.9%	82.9%
Fault selected most often	f_R	f_i
Share of instants at which it is selected	82.9%	52.0%

The two probabilities coincide to six decimal places, and the maximum is shared at 82.9% of the evaluated instants. Reversing the order of the fault list moves the selection from f_R to f_i while leaving every probability unchanged. The share at which the new winner is selected is lower, at 52.0%, because at some instants, the shared maximum extends beyond the pair and the reversal hands those instants to a third candidate. Whatever the consistency-based diagnosis attains inside the set is therefore produced by the order in which the candidates are listed, which is the situation that Corollary 1 describes.

The comparison with learning-based methods uses the configuration of Section 7.5: the same five methods with the hyperparameters of Dash et al. [12], residual windows of length 20 as input, a training pool of 60 episodes, 3 repetitions averaged at each training-set size, and no candidate gate. Results are given in Table 12.

Inside the indistinguishable set, the five methods reach between 0.5077 and 0.5856, against 0.5000 for a selection made at random within a set of two. The proposed method reaches 0.8244 on the same basis without using fault data, which is a margin of +0.2388 over the best of the five. The corresponding margin on the three-tank benchmark is +0.2253 for a set of three candidates.

Figure 6 shows the in-set accuracy against the number of training samples. Over the last three doublings, the best method gains +0.034, +0.022 and +0.017, and closing the remaining +0.2388 at the last of these rates would take a further 14 doublings, that is, a training set of the order of 10^7 samples per fault, to reach the value the proposed method attains with none. Accuracy on the training data itself is 0.9105 for BG-CNN, 0.9275 for ANN, and 0.9991 for RF, so the models absorb the training data they are given here as they do on the three-tank benchmark.

Results for conditions departing from the reference setting on this benchmark are reported in Appendix B alongside those of the three-tank system.

The proposed method outperformed the existing diagnosis on every evaluation metric and, without using fault data, achieved higher isolation performance than the learning-based methods. The results so far are summarised in Table 13, and the figures of the two benchmarks are placed side by side in Table 14.

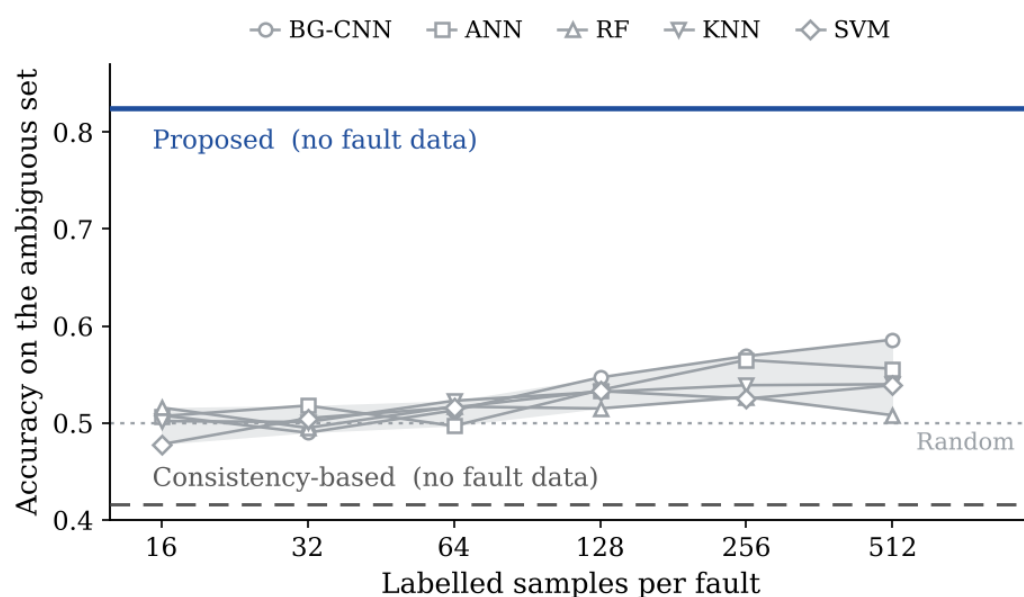


Figure 6. Isolation accuracy on the indistinguishable set of the electric motor as a function of the number of training samples per fault. The solid line is the proposed method and the dashed line the consistency-based diagnosis, neither of which uses fault data; the dotted line is selection at random within the two-element set. The shaded band spans the lowest and the highest of the five learning-based methods at each sample size.

Table 12. Comparison with learning-based methods on the electric motor at 512 training samples per fault. The columns and the evaluation basis are those of Table 8. Random selection within the two-element set corresponds to 0.5000.

Method	Fault Data ↓	Overall Accuracy ↑	Indistinguishable Set ↑	In-Set F1 ↑
Consistency-based diagnosis	0	0.6774	0.4167	0.6015
BG-CNN	512	0.8188	0.5856	0.8184
ANN	512	0.8080	0.5556	0.8081
RF	512	0.7852	0.5077	0.7860
KNN	512	0.8012	0.5401	0.7907
SVM	512	0.8040	0.5394	0.7943
Proposed method	0	0.9095	0.8244	0.9114

Table 13. Summary of the main experimental results on the three-tank benchmark. Arrows indicate that higher values are better, and the better value of each item is shown in bold.

Item	Consistency-Based	Proposed	Difference	Reference
$P_{ih} \uparrow$	0.2547	0.6735	+0.4188	Table 3
$P_{nm} \uparrow$	0.6392	0.8274	+0.1882	Table 3
$R_{det} \uparrow$	—	0.686	—	Table 3
Overall accuracy ↑	0.5942	0.7673	+0.1731	Table 4
Indistinguishable set ↑	0.3148	0.5265	+0.2116	Table 4
Other four faults ↑	0.7538	0.9048	+0.1511	Table 4
Fault data required	0	0	—	Table 8
Best learning-based in-set result ↑	0.2747	0.5000	+0.2253	Table 8

Table 14. The same quantities on the two benchmarks. The indistinguishable set has three elements on the three-tank system and two on the electric motor, so the value reached by selection at random within the set differs and is given alongside.

Item	Three-Tank	Electric Motor
Size of the indistinguishable set	3	2
Random selection within the set	0.3333	0.5000
P_{ih} , consistency-based	0.2547	0.3523
P_{ih} , proposed	0.6735	0.7110
Gain in P_{ih}	+0.4188	+0.3587
In-set accuracy, best learning-based	0.2747	0.5856
In-set accuracy, proposed	0.5000	0.8244
Margin over the best learning-based	+0.2253	+0.2388
Fault data used by the proposed method	0	0

8. Conclusions

This work addressed the problem of structural indistinguishability, in which different faults share an identical residual signature because of the sensor configuration and the system structure. In such cases, the existing consistency-based diagnosis keeps the true fault in the candidate set but produces no ranking among candidates that share the same signature.

To resolve this limitation, the proposed method does not discard the existing diagnostic result but assumes each candidate left by the diagnosis to be the true fault and propagates the model forward. By combining the agreement obtained from comparing each candidate prediction with the actual measurements with the posterior probability of the existing diagnosis, the ranking of the candidates was adjusted using temporal information that the

residual signature does not contain. The limited amount of computation was allocated over several candidates to secure comparable evidence.

On the three-tank benchmark, the isolation accuracy over the three faults sharing an identical residual signature improved from 31.5% to 52.7%. The improvement was not confined to those faults: the accuracy over all seven faults rose in the same experiment, and the true fault was restored to the top candidate at most of the instants where the existing diagnosis had pointed elsewhere. On a second benchmark, the direct-current series motor distributed with the same toolbox, the probability assigned to the true fault rose from 0.3523 to 0.7110.

Learning-based methods supplied with the same residuals did not reach this level even with 512 training samples per fault, whereas the proposed method uses no fault data at all.

Because the proposed method performs additional computation on the candidates obtained under the current model and sensor configuration, it is distinguished from approaches that acquire new fault data or apply a separate input to the plant. It can therefore be regarded as a way of using the model already at hand to generate additional discriminative information in situations where existing model-based diagnosis cannot narrow the candidates sufficiently.

The results reported above are obtained on two low-order simulated plants with a single abrupt fault at a time, under stationary process noise and at the reference setting, and the conclusions are limited to those conditions. Five limitations follow.

First, both plant models are of low order, and the error bound of Proposition 4 is stated for a plant linear in the state with $\rho(A) < 1$; the propagation of Equation (7) itself integrates whichever model is supplied, as the motor of Section 7.6 shows. Second, a single fault is active at a time, so the candidate space does not grow combinatorially, and the budget of Equation (6) is not strained; simultaneous and cascading faults enlarge that space beyond what a fixed budget T can cover, and the method does not apply to them. Third, Propositions 4 and 8 assume bounded disturbances, whereas the experiments, including the sweep of Appendix B, use stationary Gaussian noise, which does not satisfy that assumption exactly. The robustness results under Gaussian noise are therefore to be read empirically, and non-stationary or unmodelled disturbances remain outside the scope of the present analysis. Fourth, the main experiments assume the fault magnitude to be known. In a deployed setting it is not, and the reported figures are therefore to be read as an upper bound, the measured drop being $R_{det} = 0.686 \rightarrow 0.654$ under a search over five candidate magnitudes. Fifth, the correction requires the candidate predictions to separate beyond the combined estimation and noise level. Proposition 8 gives a sufficient separation condition, and Appendix B reports twelve combinations of fault magnitude θ and process-noise standard deviation σ_w . The method is above the existing diagnosis in all twelve, but a substantial gain is obtained only for $\theta/\sigma_w \geq 12.5$; in the six conditions at or below $\theta/\sigma_w = 10$, the gain does not exceed 0.15. This boundary is empirical and specific to this plant.

The comparison with learning-based methods applies to methods taking the same residuals as input. The behaviour of the method on an industrial plant remains to be established.

Author Contributions: Conceptualization, Y.E.M. and M.-J.K.; methodology, Y.E.M.; software, Y.E.M.; validation, M.-J.K.; formal analysis, Y.E.M.; investigation, Y.E.M.; resources, M.-J.K.; data curation, Y.E.M.; writing—original draft preparation, Y.E.M.; writing—review and editing, M.-J.K.; visualization, Y.E.M.; supervision, M.-J.K.; project administration, M.-J.K.; funding acquisition, M.-J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the IITP grant funded by the Korean government (MSIT) (RS-2023-00256629, RS-2024-00437718, IITP-2026-RS-2026-25619158, AI Star Fellowship Support Program); and by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2025-25398164).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Detailed Settings of the Comparison Methods

The settings used when implementing each comparison method on the present plant are summarised in Table A1. The architecture and hyperparameters presented in the original papers were retained, and the input dimension and the residual set are formed to match the three-tank system.

Table A1. Implementation settings of the comparison methods

Method	Setting in the Source	Setting in this Work
Consistency-based diagnosis	Residuals and fault signature from structural analysis	Five three-tank MSOs, thresholds calibrated at the 99th percentile on nominal data
Residual sensitivity-based isolation	Sensitivity index from the signal-to-noise ratio, minimum-cardinality subset selection	The same procedure applied to the indistinguishable setting of this benchmark, selecting the minimum-cardinality subset that preserves the achievable isolability; r_1 , r_2 , r_3 maximise SSI
Uncertainty-based diagnosis	Ensemble of probabilistic neural networks trained on nominal data only, decision withheld under epistemic uncertainty	Single-layer LSTM with 16 hidden units predicting mean and variance, ensemble of 10, trained on 6 nominal seeds
BG-CNN	32 kernels of 2×2 , 2×2 max pooling with stride 2, fully connected 128	Same architecture, input is a five-channel residual window
ANN	Hidden layers 128-64-32	Same
RF	max_depth 50, n_estimators 20	Same
KNN	n_neighbors 2	Same
SVM	RBF kernel, $C = 1$	Same

The BG-CNN and ANN were trained for 200 epochs using the Adam optimiser with a learning rate of 0.001. The number of training samples per fault was varied from 16 to 512. Data generated from different seeds were used for training and for evaluation.

Appendix B. Results under Varying Conditions

To confirm whether the proposed method operates only under a particular reference setting, the fault magnitude and the process noise were varied. The results are given in Table A2.

The gain is positive in all twelve conditions, and the ratio θ/σ_w orders them. In the six conditions with $\theta/\sigma_w \geq 12.5$, the gain lies between +0.3424 and +0.4188; in the six with $\theta/\sigma_w \leq 10$, it does not exceed +0.1494. The horizontal rule in Table A2 marks the boundary.

When the fault is small relative to the noise, the differences between the candidate predictions do not separate beyond the noise level, so the agreement score carries no discriminative information. This is the behaviour described by Proposition 8, whose separation condition compares the distance between the candidate predictions with the combined

estimation and noise level. The value 12.5 is measured on this plant; the proposition is stated on σ_{kl} , which takes its own value for each pair of hypotheses. Figure A1 shows the trend.

Table A2. Results for varying fault magnitude and process noise. The rows are ordered by θ/σ_w . The last column is the gain of the proposed method over the consistency-based diagnosis, and the reference setting is shown in bold.

θ	σ_w	θ/σ_w	Consistency-Based $\uparrow$	No Correction $\uparrow$	Proposed $\uparrow$	Gain
0.05	0.020	2.5	0.0097	0.0114	0.0111	+0.0014
0.05	0.010	5.0	0.0354	0.0501	0.0500	+0.0146
0.10	0.020	5.0	0.0335	0.0452	0.0448	+0.0113
0.15	0.020	7.5	0.0753	0.1422	0.1477	+0.0724
0.05	0.005	10.0	0.0961	0.1576	0.1605	+0.0644
0.10	0.010	10.0	0.1370	0.2710	0.2864	+0.1494
0.25	0.020	12.5	0.2442	0.5759	0.6391	+0.3949
0.15	0.010	15.0	0.2547	0.6060	0.6735	+0.4188
0.10	0.005	20.0	0.2561	0.5949	0.6612	+0.4051
0.25	0.010	25.0	0.2454	0.5783	0.6384	+0.3930
0.15	0.005	30.0	0.2440	0.5771	0.6276	+0.3836
0.25	0.005	50.0	0.2475	0.5838	0.5899	+0.3424

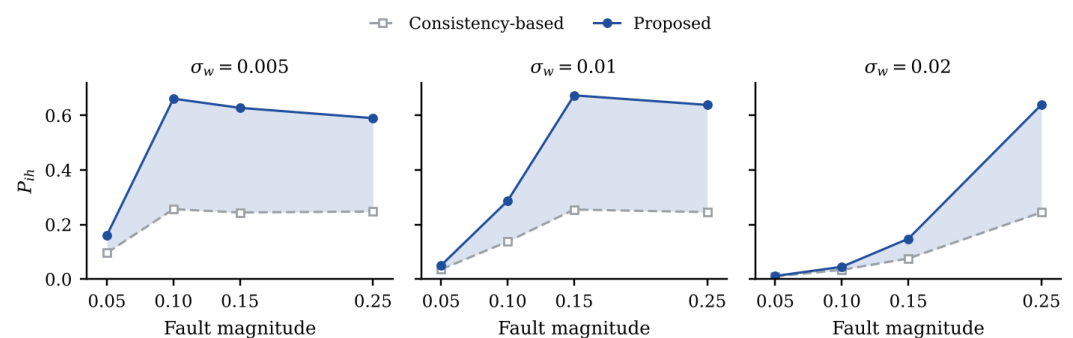


Figure A1. P_{th} as a function of fault magnitude and process noise. The shaded region marks the gap between the two methods.

The same sweep was carried out on the electric motor of Section 7.6, with the fault magnitude varied against a measurement noise of 0.020 and the magnitudes of the two members of the indistinguishable set kept matched throughout. The results are given in Table A3.

Table A3. Results for varying fault magnitude on the electric motor over the twelve seeds of Section 6. The ratio is that of the fault magnitude to the measurement noise. The reference setting is shown in bold and reproduces the values of Table 10.

Magnitude	Ratio	Consistency-Based $\uparrow$	Proposed $\uparrow$	Gain	p
0.02	1.0	0.0015	0.0039	+0.0024	0.21
0.05	2.5	0.0131	0.0251	+0.0120	0.0096
0.08	4.0	0.0645	0.1264	+0.0618	4.5×10^{-5}
0.15	7.5	0.3523	0.7110	+0.3587	1.7×10^{-8}
0.25	12.5	0.4994	0.9553	+0.4559	1.5×10^{-12}
0.40	20.0	0.5000	0.9896	+0.4896	6.6×10^{-16}

The gain grows monotonically with the ratio on this benchmark, and the three conditions at or above the reference ratio gain between +0.3587 and +0.4896, while the three below it gain at most +0.0618. The ratio here is formed against the measurement noise,

whereas the three-tank sweep varies the process noise, so the two reference ratios are not directly comparable; that they differ is in any case consistent with the separation condition of Proposition 8 being stated on σ_{kl} , a quantity that takes its own value on each plant and each pair of hypotheses. What the two sweeps have in common is the direction: the correction adds information once the fault is large relative to the noise and adds progressively less as that ratio falls.

To examine how the amount of information given to the model propagation affects the result, the state used as the starting point and the knowledge of the fault magnitude were varied. The results are given in Table A4.

The results are shown in Figure A2. Performance was higher under the ideal condition in which the true state is available, but the improvement of the proposed method was maintained when the estimated state was used in accordance with the actual setting. Even when the fault magnitude is not known in advance, searching over candidate magnitudes maintained $R_{det} = 0.654$.

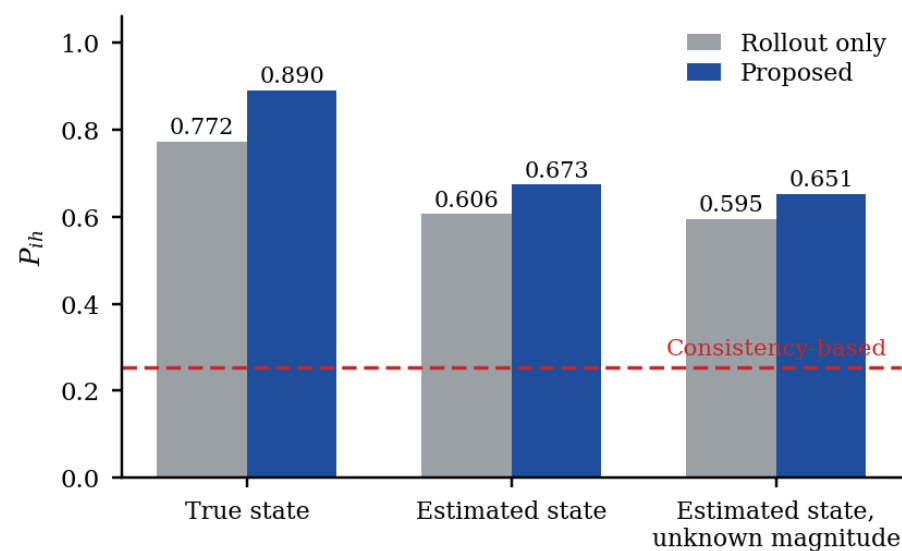


Figure A2. P_{th} when the information given to the model propagation is restricted. The red dashed line is the existing consistency-based diagnosis.

Table A4. Results for varying information access.

Condition	No Correction ↑	Proposed ↑	R_{det} ↑
True state, known magnitude	0.7719	0.8904	0.893
Estimated state, known magnitude	0.6060	0.6735	0.686
Estimated state, unknown magnitude	0.5946	0.6513	0.654

Appendix C. Design Variables

The reference setting used in the main text was selected on validation seeds that do not overlap with the seeds used for reporting the results. Each table varies a single design variable and keeps the others at the reference setting.

Appendix C.1. Computation Horizon

As shown in Table A5 and Figure A3, a setting of 25 steps was used as the reference. With a short horizon, the trajectory differences between candidates do not separate sufficiently, while with a long horizon, the state estimation error accumulates. Performance higher than the existing consistency-based diagnosis was maintained at the other horizons as well.

Table A5. Results for varying computation horizon L . The evaluation window ends L steps before the end of the run, so each row is scored on its own set of instants, and the difference is taken against the consistency-based diagnosis evaluated at the same horizon.

L	Relative Cost ↓	No Correction ↑	Proposed ↑	Difference
5	0.2	0.5906	0.6418	+0.3888
10	0.4	0.6237	0.6681	+0.4144
25	1.0	0.6060	0.6735	+0.4188
50	2.0	0.5830	0.6611	+0.4073
75	3.0	0.5712	0.6137	+0.3593

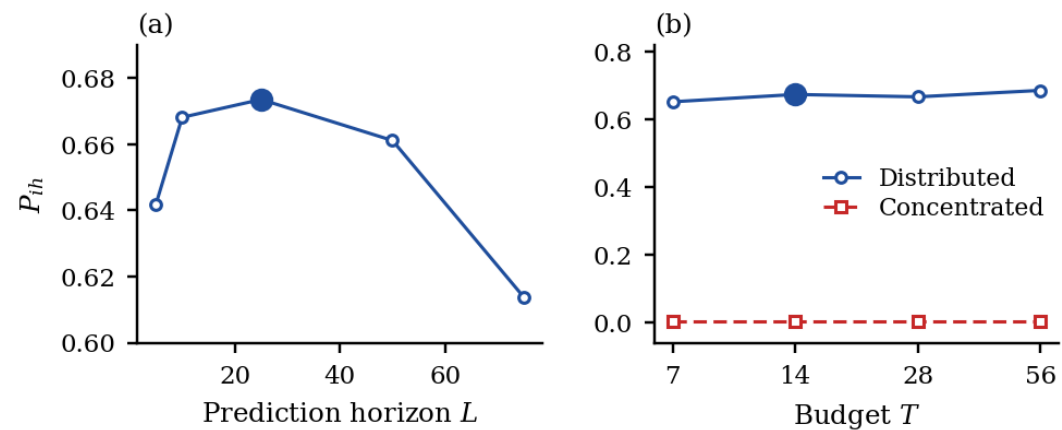


Figure A3. Influence of the design variables. The filled marker indicates the reference setting used in the main text.

Appendix C.2. Computational Budget

In Table A6, R_{det} rises by 0.030 when T is increased from 7 to 14 and varies by less than 0.02 beyond that. The highest performance appeared at $T = 56$, but the difference from $T = 14$ is only 0.005 in R_{det} while the amount of computation is four times larger. Considering performance and computational cost together, $T = 14$ was used as the reference setting, which gives 3 to 4 computations to each candidate when 4 candidates remain.

Table A6. Results for varying computational budget T .

T	Concentrated ↑	No Correction ↑	Proposed ↑	R_{det} ↑
7	0.0023	0.5929	0.6517	0.656
14	0.0023	0.6060	0.6735	0.686
28	0.0023	0.5989	0.6664	0.673
56	0.0023	0.6079	0.6855	0.691

When all computation was allocated to the single candidate with the highest probability, on the other hand, the result did not change from 0.0023 even when the budget was increased eightfold. Equation (8) then has no comparison to make and assigns $s_k = 1$ to that candidate, so the correction reinforces whichever candidate the existing diagnosis already favoured, and no amount of computation produces discriminative evidence.

Appendix C.3. Combination Coefficients

In Table A7, the maximum value appears identically for several combinations, and the performance varies little over a wide region. This is because the combination in Equation (9) takes the form $\beta(1 + \gamma c)$, so the two coefficients do not act independently; it is the combined quantity $\beta(1 + \gamma c)$, not either coefficient on its own, that sets the effective strength. It is

therefore more appropriate to regard the performance of the method as stable over that range than to interpret a particular combination as the unique optimum. Two mechanisms account for this flatness and fix what the grid search settles. By Corollary 3, the candidate selected within an indistinguishable set is independent of β and γ . Beyond that set, the selection is held in place by the scale at which the two terms of Equation (9) meet; with the weights of Equation (5), the spread of $(1 - c) \log b_k$ over the candidate set is at most 0.19, whereas $\beta(1 + \gamma c)$ is at least 3 everywhere on the grid, so the decision follows $\arg \max_k s_k$ throughout it. The probability masses are a separate matter, and the grid search does select among them; over the whole grid, the entry ranges from 0.6601 to 0.7254, a spread of 0.0653, and with the column $\gamma = 0$ set aside, it ranges from 0.7107 to 0.7254, a spread of 0.0147. The reference pair $(\beta, \gamma) = (6, 8)$ attains 0.7254. The quantities reported in this paper that are probability masses, namely P_{ih} and P_{nm} , are accordingly read within a band of 0.0147.

Table A7. Grid search over the combination coefficients β and γ on the validation seeds, which do not overlap with the seeds used for reporting. Values are the mean of P_{ih} and P_{nm} and are therefore not comparable with Table 3.

$\beta \backslash \gamma$	0	2	4	6	8	12	16
3	0.6601	0.7107	0.7190	0.7220	0.7236	0.7249	0.7253
6	0.7075	0.7217	0.7243	0.7252	0.7254	0.7251	0.7246
9	0.7170	0.7242	0.7253	0.7254	0.7251	0.7243	0.7236
12	0.7206	0.7251	0.7254	0.7250	0.7245	0.7235	0.7227
18	0.7237	0.7254	0.7248	0.7240	0.7233	0.7223	0.7217
24	0.7249	0.7251	0.7241	0.7232	0.7225	0.7217	0.7212

Appendix C.4. Evaluation Threshold

In Table A8 P_{ih} and R_{det} remain within a narrow range from 0.15 to 0.35. As the threshold rises, more instants fail it because several candidates remain, so the evaluated sample shrinks quickly. At 0.50, only 15 instants pass, so that row reflects the size of the remaining sample rather than a change in behaviour.

Table A8. Results for various evaluation thresholds.

Threshold	Samples	$P_{ih} \uparrow$	$R_{det} \uparrow$	Correction Gain $\uparrow$
0.15	655	0.6570	0.656	+0.0691
0.20	647	0.6586	0.660	+0.0704
0.25	422	0.6735	0.686	+0.0675
0.35	64	0.6649	0.683	+0.0601
0.50	15	0.1667	0.167	+0.0005

References

1. Blanke, M.; Kinnaert, M.; Lunze, J.; Staroswiecki, M. *Diagnosis and Fault-Tolerant Control*, 2nd ed.; Springer: Berlin, Germany, 2006.
2. Qin, S.J. Survey on data-driven industrial process monitoring and diagnosis. *Annu. Rev. Control* **2012**, *36*, 220–234.
3. Krysander, M.; Åslund, J.; Nyberg, M. An efficient algorithm for finding minimal overconstrained subsystems for model-based diagnosis. *IEEE Trans. Syst. Man Cybern. Part A* **2008**, *38*, 197–206.
4. Frisk, E.; Krysander, M.; Jung, D. A toolbox for analysis and design of model based diagnosis systems for large scale models. *IFAC-PapersOnLine* **2017**, *50*, 3287–3293.
5. De Kleer, J.; Williams, B.C. Diagnosing multiple faults. *Artif. Intell.* **1987**, *32*, 97–130.
6. Nyberg, M.; Krysander, M. Combining AI, FDI, and statistical hypothesis-testing in a framework for diagnosis. *IFAC Proc. Vol.* **2003**, *36*, 813–818.
7. Jung, D.; Ng, K.Y.; Frisk, E.; Krysander, M. Combining model-based diagnosis and data-driven anomaly classifiers for fault isolation. *Control Eng. Pract.* **2018**, *80*, 146–156.

8. Mohammadi, A.; Krysander, M.; Jung, D. Consistency-based diagnosis using data-driven residuals and limited training data. *Control Eng. Pract.* **2025**, *159*, 106283.
9. Murata, R.; Marzat, J.; Piet-Lahanier, H.; Boujnah, S.; Belleoud, P. Fault detection and isolation based on structural analysis: Application to a multi-engine propulsion cluster. *Sensors* **2025**, *25*, 1054.
10. Jung, D. Isolation and localization of unknown faults using neural network-based residuals. In Proceedings of the Annual Conference of the PHM Society, Scottsdale, AZ, USA, 21–26 September 2019; Volume 11.
11. Lundgren, A.; Jung, D. Data-driven fault diagnosis analysis and open-set classification of time-series data. *Control Eng. Pract.* **2022**, *121*, 105006.
12. Dash, B.M.; Ould Bouamama, B.; Boukerdja, M.; Pekpe, K.M. Bond graph-CNN based hybrid fault diagnosis with minimum labeled data. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107734.
13. Sankavaram, C.; Kodali, A.; Pattipati, K.R.; Singh, S. Incremental classifiers for data-driven fault diagnosis applied to automotive systems. *IEEE Access* **2015**, *3*, 407–419.
14. Nikoukhah, R. Guaranteed active failure detection and isolation for linear dynamical systems. *Automatica* **1998**, *34*, 1345–1358.
15. Šimandl, M.; Punčochář, I. Active fault detection and control: Unified formulation and optimal design. *Automatica* **2009**, *45*, 2052–2059.
16. Scott, J.K.; Findeisen, R.; Braatz, R.D.; Raimondo, D.M. Input design for guaranteed fault diagnosis using zonotopes. *Automatica* **2014**, *50*, 1580–1589.
17. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
18. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
19. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **2021**, *110*, 457–506.
20. Genest, C.; Zidek, J.V. Combining probability distributions: A critique and an annotated bibliography. *Stat. Sci.* **1986**, *1*, 114–135.
21. Amari, S. *Information Geometry and Its Applications*; Applied Mathematical Sciences 194; Springer: Tokyo, Japan, 2016.
22. Bissiri, P.G.; Holmes, C.C.; Walker, S.G. A general framework for updating belief distributions. *J. R. Stat. Soc. B* **2016**, *78*, 1103–1130.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.