



## Journal of Smart Algorithms and Applications JSAA

ISSN: 3070-4189/© 2026 JSAA. (CC BY 4.0) License.

Journal Homepage

<https://pub.scientificirg.com/index.php/JSAA>



# A Critical Review and Methodological Framework for Robust Tabular Learning in Industrial Predictive Maintenance

Arwa Saad<sup>a,1</sup>, Amer Ibrahim<sup>b</sup>, Adriana Burlea Schiopoiu<sup>c</sup>

*a Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt.*

*Email: arwasaad706\_sd@fcis.bsu.edu.eg*

*b Department of Computer Science and Software Engineering, United Arab Emirates University, UAE.*

*Email: Amer.ibrahim@uaeu.ac.ae*

*c Department of Management, Marketing, Business Administration, Faculty of Economy and Business Administration, University of Craiova, Craiova, Romania. Email: Adriana.burlea@gmail.com*

## ABSTRACT

Data-driven predictive maintenance (PdM) relies increasingly on machine learning models trained in sensor and operational data. However, missing measurements, sensor corruption, and distribution shift can complicate the evaluation of predictive reliability under changing industrial conditions. This critical narrative review examines how these data imperfections are addressed in industrial maintenance research and related tabular-learning literature. It critically synthesizes evidence from industrial tabular studies, sequential sensor applications, general-purpose benchmarks, and methodological research, focusing on differences in data characteristics, prediction tasks, stressor mechanisms, and validation procedures. The synthesis identifies important limitations in the available evidence, particularly the lack of comparable controlled factorial evaluations of missingness, measurement corruption, and distribution shift in the industrial tabular PdM studies examined. Consequently, the reviewed findings do not provide a sufficient basis for general conclusions about model robustness under compound imperfections or universal rankings across model families. To support future investigations, this review develops a mechanism-aware conceptual framework distinguishing the three stressors and presents pairwise and three-way loss-interaction contrasts for analyzing potential non-additive effects. It also proposes a leakage-aware evaluation protocol incorporating realistic stressor configurations, appropriate validation strategies, uncertainty estimation, and maintenance-oriented performance measures. The review contributes a critical synthesis and a methodological proposal for future empirical research; neither the proposed framework nor its practical effectiveness is experimentally validated in this study.

## PAPER INFORMATION

### HISTORY

**Received:** 12 August 2026

**Revised:** 19 September 2026

**Accepted:** 25 October 2026

**Online:** 30 October 2026

### MSC:

68T07; 68R10; 94A60;  
68M15

### KEYWORDS

Predictive maintenance;  
Tabular learning;  
Missing data;  
Distribution shift;  
Leakage-aware evaluation.

<sup>1</sup> Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt.

*Email: arwasaad706\_sd@fcis.bsu.edu.eg*

## 1. INTRODUCTION

Data-driven predictive maintenance (PdM) is the use of operational and condition-monitoring data to inform equipment health assessment and maintenance decisions [1]. Machine learning (ML) and deep learning (DL) methods have been studied for several tasks such as fault detection, fault diagnosis, equipment failure prediction, and remaining useful life (RUL) estimation [2].

These tasks differ in prediction targets, temporal requirements, and evaluation standards. In industrial applications, sensor measurements can be incomplete, affected by measurement errors, or collected under operating conditions different from those of the training data [3]–[5]. Reported predictive performance therefore has to be read together with the data conditions under which it was obtained.

Previous studies have examined machine learning (ML) and deep learning (DL) for predictive maintenance and prognostics, and provide overviews of methods, applications and challenges [6]. Surveys of tabular learning compare tree-based methods with neural architectures [7]. Other recent work covers adversarial deep transfer learning for mechanical fault diagnosis [8], attention-enhanced spatiotemporal deep learning for predictive maintenance [9], and out-of-distribution (OOD) detection in fault diagnosis [10].

Research on tabular learning provides further evidence for evaluation under changing data [11], [12]. TableShift is a benchmark of 15 binary classification tasks with distribution shifts across application domains [13]. TabReD introduced eight tabular datasets for studying time-based data partitioning and feature-rich scenarios [14]. Cheng et al. [15] propose TabFSBench to study changes in the set of available input features under feature-shift scenarios. Within predictive maintenance, Autran et al. [16] proposed AI4I-PMDI, a dataset built on the synthetic AI4I 2020 dataset and modified to include missing observations and irregular data acquisition. These studies cover different aspects of data quality, feature availability and distributional variation, but they use different data sources, prediction aims and evaluation methods, so their results do not form one comparable body of evidence.

Missing observations, measurement error and distribution shift are related but different aspects of imperfect data. Missingness concerns the presence of observations and may arise from sensor outages, communication breakdowns or acquisition procedures [17]. Measurement corruption concerns the accuracy of recorded values and includes random noise, systematic bias, sensor drift, and abnormal readings. Distribution shift is a change in the statistical properties of data between training and deployment [18]; the change may affect the input distribution, the relation between inputs and targets, or the target distribution [19], and it may also change the set of available features. The effect of one form of data modification does not necessarily characterize the effect of another [20], so robustness must be assessed for each stressor and for their combinations.

Interpretation of predictive maintenance results also depends on methodological choices made during dataset preparation and evaluation [21]. Information derived from the prediction target may lead to target leakage, and random partitioning of temporally dependent or equipment-specific observations may produce evaluations that do not reflect performance on future operating periods or unseen equipment. The suitability of a validation strategy therefore depends on the intended prediction setting, and published studies differ in fault definitions, prediction horizons, class distributions and units of evaluation [22].

This review analyzes the literature on ML and DL approaches to predictive maintenance (PdM) under imperfect data, with a focus on structured tabular representations. It considers gradient-boosted decision trees (GBDT), deep tabular architectures, missingness-aware learning approaches and tabular foundation models. Sequential sensor data are discussed where they inform the treatment of incomplete or corrupted observations, and evidence from industrial maintenance is kept separate from general-purpose tabular benchmarks and methodological studies.

The main contributions of this critical review are summarized as follows:

- **Mechanism-aware taxonomy:** This review develops a conceptual framework distinguishing sensor missingness (M), measurement corruption (N), and distribution shift (S) according to their underlying mechanisms, operational characteristics, and potential dependencies.

- **Interaction analysis:** This review formalizes pairwise and three-way loss-interaction contrasts to distinguish additive from non-additive responses to compound data imperfections, while clarifying the assumptions and limitations governing their interpretation.
- **Critical evidence synthesis:** This review organizes the existing evidence across industrial tabular applications, sequential sensor studies, general-purpose benchmarks, and methodological research, identifying why the available findings do not establish universal model-family rankings under compound data imperfections.
- **Evaluation and reporting framework:** This review proposes a leakage-aware, maintenance-oriented evaluation protocol integrating realistic stressor configurations, appropriate validation strategies, uncertainty estimation, and maintenance-relevant performance measures.

Together, these contributions provide a critical synthesis and a conceptual framework for future robustness evaluations rather than an empirically validated benchmark.

The remainder of this review is organized as follows. Section 2 outlines the scope of the critical narrative review, the literature selection approach, the research questions, and the evidence classification framework. Section 3 introduces the mechanism-aware conceptual framework, including sensor observation processes, stressor interactions, distributional changes, and maintenance-related feedback. Sections 4–6 examine sensor missingness, measurement corruption, and distribution shift, respectively. Section 7 critically synthesizes the available evidence and identifies the limitations of existing robustness evaluations. Section 8 discusses the evidence and unresolved challenges associated with different machine learning and deep learning model families. Section 9 examines industrial datasets, their suitability for robustness assessment, and potential sources of data leakage. Section 10 proposes a reproducible, leakage-aware protocol for evaluating combined data imperfections. Section 11 discusses the limitations of the present review, and Section 12 concludes the paper by summarizing its main contributions and directions for future research.

## 2. REVIEW SCOPE, QUESTIONS, AND EVIDENCE TRANSPARENCY

### 2.1 Review design and boundaries

This study adopts a critical narrative review approach to examine machine learning (ML) and deep learning (DL) methods relevant to robust tabular prediction in industrial predictive maintenance (PdM). The review draws on primary research articles, benchmark studies, dataset documentation, and relevant methodological literature. Rather than conducting a systematic literature review or quantitative meta-analysis, the study focuses on the critical interpretation of reported findings, methodological assumptions, and differences in experimental settings that influence the assessment and comparability of robustness results [23], [24].

The literature is considered according to its relevance to industrial maintenance prediction, structured tabular learning, missing-data modeling, and robustness under changing data conditions. Attention is given to prediction objectives, data representations, data-imperfection mechanisms, experimental designs, validation protocols, and reported evaluation measures. Foundational methodological studies are considered where they provide relevant theoretical definitions, modeling principles, or evaluation procedures. The review does not claim exhaustive coverage of the available literature, and its conclusions are restricted to the evidence examined.

To distinguish the relevance and applicability of findings across research settings, the reviewed literature is organized into four evidence categories (A–D) and one resource category (R):

**Category A: Industrial tabular maintenance studies.** This category includes studies evaluating maintenance-related prediction tasks using structured operational records or engineered tabular features. Studies employing operational measurements and those relying on simulated or synthetic industrial data are distinguished when interpreting their findings. When temporal sensor measurements are transformed into fixed-length feature vectors, the resulting experiments are considered within this category when the evaluated models operate on tabular representations.

**Category B: Industrial signal-based and sequential studies.** This category covers research addressing fault diagnosis, prognostics, or related maintenance tasks using raw sensor signals or temporal sequences as model inputs. These studies are considered where they provide relevant evidence concerning incomplete measurements, sensor

corruption, temporal dependencies, or changing operating conditions. Their findings are interpreted within the corresponding signal-processing and temporal-modeling settings.

**Category C: General-purpose tabular learning and benchmark studies.** This category comprises investigations of model performance, missing features, data corruption, or distribution shift in structured datasets outside industrial PdM. Such studies provide relevant evidence about the evaluated learning methods and experimental conditions but do not independently establish their performance in industrial maintenance applications.

**Category D: Adjacent methodological studies.** This category includes research on missing-data learning, longitudinal modeling, distribution adaptation, uncertainty estimation, and related methodological approaches. These studies are considered for their theoretical or methodological relevance without assuming that their findings directly establish predictive reliability in industrial PdM settings.

**Category R: Dataset and benchmark resources.** This category covers dataset publications, repository documentation and benchmark descriptions that supply data provenance, acquisition characteristics or evaluation infrastructure. They are treated as resources and not as independent evidence of comparative model robustness. Where a resource publication also reports predictive experiments, as AI4I-PMDI does [16], the resource and its reported experiments are considered separately, and the experiments are assigned to the appropriate category (here, A).

The categories describe the primary application context, data representation, and purpose of the evidence rather than a hierarchy of methodological quality. A publication may contribute to more than one category when it reports distinct experiments involving different data representations or application settings. Category R is a resource category and not a level of evidence.

The scope of the review primarily encompasses structured industrial records and engineered tabular representations used for maintenance-related prediction. Studies involving raw sensor signals or sequential observations are included where they inform the discussion of missing measurements, sensor corruption, temporal dependencies, or distributional changes. Fault detection, fault classification, failure prediction, and remaining useful life estimation are distinguished according to their prediction objectives, temporal requirements, and evaluation criteria.

## ***2.2 Literature Search and Selection***

Publications were identified through a purposive, topic-oriented examination of literature relevant to industrial predictive maintenance, tabular machine learning, and robustness under data imperfections. Selection was guided by the review's research questions, with emphasis on maintenance prediction, fault diagnosis, remaining useful life estimation, sensor missingness, measurement corruption, feature availability, and distribution shift. Priority was given to studies providing direct evidence from industrial maintenance applications, while sequential sensor research, general-purpose tabular benchmarks, and methodological contributions were included where they offered relevant insights into modeling approaches, robustness mechanisms, or evaluation practices. The selection was not intended to be exhaustive or systematic.

The selected publications are critically examined in terms of their research objectives, data representations, prediction tasks, data-imperfection mechanisms, model families, validation strategies, and reported evaluation measures. Attention is given to whether robustness is evaluated under individual stressors or combined data imperfections, and whether the experimental conditions support comparisons across studies.

To preserve the distinction between application-specific evidence and broader methodological findings, the synthesis follows the evidence categories defined in the preceding subsection. Findings from industrial tabular applications, sequential sensor studies, general-purpose tabular benchmarks, and methodological research are interpreted within their respective experimental contexts.

### 2.3 Research Questions

This review is guided by five research questions addressing the characterization of data imperfections and distributional changes, the empirical evidence on model robustness, and methodological considerations relevant to the evaluation of industrial predictive maintenance (PdM) systems.

**RQ1.** What are the characteristics of sensor missingness, measurement corruption, and distribution shift in industrial tabular data in terms of their mechanisms, severity, temporal persistence, and relationships with equipment operating conditions and health status?

**RQ2.** What does the existing literature reveal about the robustness of tabular machine learning and deep learning methods under individual data imperfections, and what limitations constrain comparisons across model families?

**RQ3.** To what extent have the combined effects and potential interactions of sensor missingness, measurement corruption, and distribution shift been empirically investigated? Under what conditions can findings from isolated-stressor evaluations inform assessments of compound conditions?

**RQ4.** How do dataset properties, prediction targets and horizons, data preparation procedures, information leakage, and validation strategies influence the validity and comparability of reported robustness evaluations?

**RQ5.** What methodological considerations should guide the evaluation of predictive robustness under imperfect industrial data, particularly regarding performance metrics, uncertainty estimation, and maintenance-related consequences?

### 2.4 Evidence Classification and Comparative Assessment

The interpretation of robustness findings requires careful consideration of the conditions under which the evidence was obtained and the extent to which those conditions correspond to industrial tabular predictive maintenance (PdM). Studies investigating similar learning challenges may differ substantially in their prediction objectives, data representations, observation mechanisms, and validation procedures. For example, industrial fault-diagnosis research may evaluate adaptation across operating regimes using vibration signals, whereas general-purpose tabular benchmarks may investigate shifts between predefined populations. Dataset publications may instead document acquisition characteristics or irregular observations without establishing comparative robustness. These contributions provide different forms of evidence and should not be treated as interchangeable demonstrations of model reliability in industrial PdM.

This review distinguishes **evidence applicability** from **quality of inference**. Evidence applicability refers to the correspondence between a study's experimental setting and the industrial tabular prediction problem under consideration, including data provenance, representation, prediction target, observation mechanisms, and operating conditions. Quality of inference concerns whether the experimental design, validation procedures, and reported results adequately support the conclusions drawn within the original study. A methodologically rigorous experiment may provide only indirect evidence for industrial maintenance, while a study conducted on maintenance data may still have limitations that constrain claims about deployment robustness.

**Table 1** summarizes the evidence categories defined in **Section 2.1**, their supported interpretations, and their principal applicability limits.

**Table 1. Evidence categories, supported interpretations, and applicability limits.**

| Representative references (context only) | Evidence category                         | Supported interpretation                                                                         | Principal limitation                                                                                                                                                                     |
|------------------------------------------|-------------------------------------------|--------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [16], [25]–[27]                          | A. Industrial tabular maintenance         | Performance on the evaluated maintenance task and specified data irregularities                  | Synthetic-derived evidence does not establish field-deployment reliability. AI4I-PMDI is also a Category R resource; only its reported experiments count as Category A.                  |
| [5], [28]–[30]                           | B. Industrial signal-based and sequential | Behavior under evaluated signal conditions and operating regimes                                 | Cannot directly establish robustness rankings for tabular maintenance predictors                                                                                                         |
| [13]–[15], [19], [31]                    | C. General tabular and non-PdM benchmarks | Findings on natural shifts, temporal evaluation, feature-set changes, or tabular prediction      | Does not directly establish robustness to industrial sensor failures. TabReD [14] is industry-grade but drawn from non-maintenance application domains, so it is classified as C, not A. |
| [32]–[36]                                | D. Adjacent methodological studies        | Design options for masks, observation times, imputation and uncertainty under stated assumptions | A method or theoretical result is not a PdM stress-test outcome; GRU-D [32], for example, was evaluated on clinical time series.                                                         |
| [37]–[41]                                | R. Dataset and benchmark resources        | Dataset provenance, acquisition characteristics and temporal structure                           | Documentation alone does not establish comparative robustness                                                                                                                            |

### 3. A MECHANISM-AWARE CONCEPTUAL FRAMEWORK

#### 3.1 Sensor Observation and Data Imperfection Mechanisms

Evaluating predictive robustness under imperfect industrial data requires distinguishing the underlying operational quantities from the processes through which sensor measurements are generated, observed, and supplied to learning algorithms [42]. Let  $X^*$  denote the underlying sensor and operational feature values,  $Y$  the maintenance-related prediction target,  $Z$  the operating regime,  $M$  a binary observation mask, and  $E$  the measurement error. Depending on the application,  $Y$  may represent a fault class, a failure event within a specified prediction horizon, or a remaining useful life (RUL) quantity. For temporal prediction tasks, variables are indexed by time, and the predictive information set is restricted to observations available at the decision time.

For sensor channel  $j$  at time  $t$ , let  $M_{j,t} = 1$  indicate an available measurement and  $M_{j,t} = 0$  an unavailable measurement.

$$W_{j,t} = h_j(X_{j,t}^*, E_{j,t}, Z_t)$$

where  $h_j$  describes the measurement process. A special case applicable to certain numerical sensors is the additive measurement-error model,

$$W_{j,t} = X_{j,t}^* + E_{j,t}$$

More general formulations may be required to represent systematic bias, sensor drift, saturation, clipping, quantization and other distortions.

$$X_{j,t}^{obs} = \begin{cases} W_{j,t}, & \text{if } M_{j,t} = 1 \\ \text{NA}, & \text{if } M_{j,t} = 0 \end{cases}$$

The observation mask may also be retained as an explicit predictive input when the learning architecture and preprocessing procedure support its use. To keep the notation unambiguous,  $X^{obs}$  always denotes the recorded features with NA entries, which is what native missing-value routing and imputation operate on, and the mask-augmented model input is written  $X = (X^{obs}, M)$  in this paper (after imputation,  $X$  contains imputed values and the mask).

**Notation and terminology.**  $M_{j,t}$  is the observation mask. The italic capitals  $M$ ,  $N$  and  $S$  name the three stressors (missingness, measurement corruption, distribution shift) in figures and in the subscripts of the loss  $L$ , and the lowercase  $m$ ,  $n$ ,  $s$  are the corresponding binary design indicators. The stressor  $N$  acts through the error term  $E_{j,t}$  and the measurement function  $h_j$ . We use *measurement corruption* for any departure of a recorded value from the underlying quantity, including random noise (the zero-mean stochastic special case), bias, drift, saturation, and stuck-at readings; *sensor corruption* in the title is a synonym, and *measurement disturbance* is not used as a separate term.  $Z$  is the operating regime, a variable of the process that exists in every environment and can be conditioned on, for example, load level.  $S$  is a difference between two environments, training and deployment, in the distribution of  $(Z, X^*, M, E)$  or in  $P(Y|X)$ . Thus  $Z$  says where within an environment the equipment operates, and  $S$  says how one environment differs from another.

This representation separates three related sources of data imperfection. Missingness describes the availability of measurements, measurement corruption concerns the correspondence between recorded values and the underlying quantities, and distribution shift refers to changes in the statistical properties of the underlying or observed data-generating processes between training and deployment environments [20]. The three are conceptually distinct but may share dependencies on operating conditions, equipment state, or sensor characteristics, and an observed change in feature distributions does not necessarily imply a change in the physical process or in the relationship between equipment condition and maintenance outcomes [43].

**Figure 1** shows that basic equipment and operational parameters are converted into measurable features by sensor measurement and observation techniques. The lack affects the usability of the measurements, and the distortion affects the precision of the recorded values. Distribution shift is a change in the operational setting from the training environment. This can include changes in operating conditions, equipment demography, or measurement and observation techniques. For failure prediction, remaining useful life estimation, or fault detection, the observed data is used by machine learning and deep learning models, possibly through missingness indicators or features extracted during the pre-processing phase. The feedback loop indicates how decisions made in maintenance can influence the future state of the equipment and the control processes. The illustration is theoretical and does not depict a system that has been empirically validated.

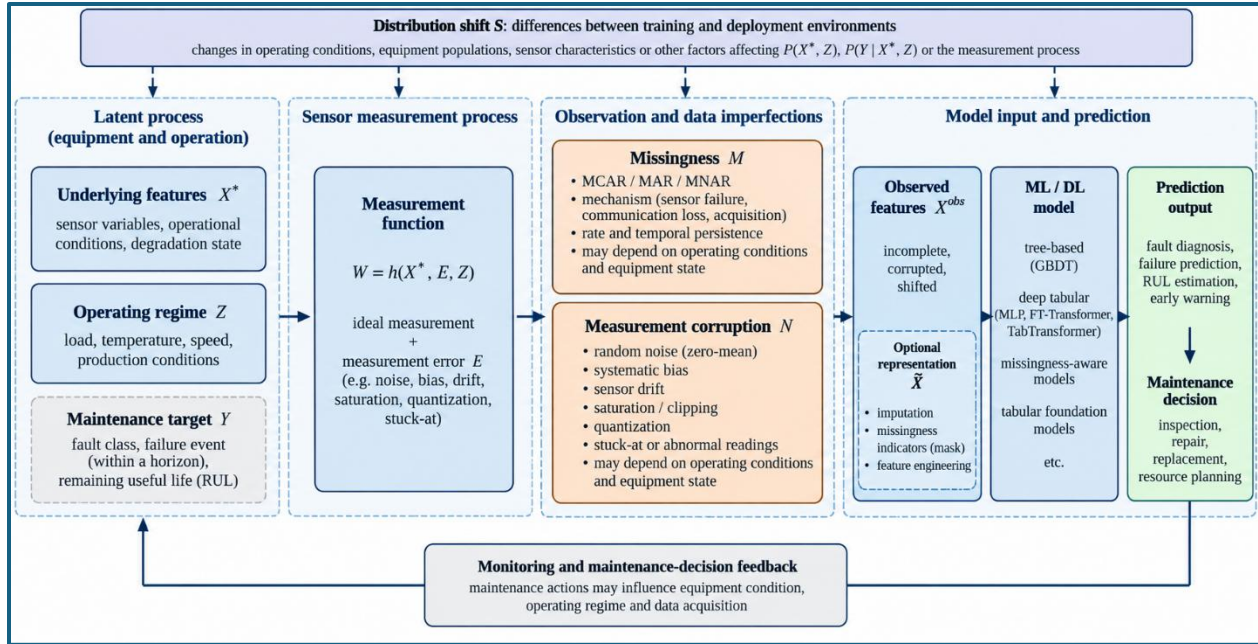


**Figure 1. Mechanism-aware observation and prediction framework.**

Sensor availability and measurement quality may also depend on operating conditions. Changes in equipment load, for example, may coincide with increased sensor variability or communication interruptions, while progressive sensor deterioration may produce biased or unstable measurements before a channel becomes unavailable [44], [45]. Such dependencies motivate the consideration of multiple data imperfections within a common evaluation framework.

Missing-data mechanisms are commonly described using the distinctions between missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [46]. Under MCAR, the missingness pattern is independent of the complete data values. Under MAR, the probability of a given missingness pattern depends only on the values that are observed under that pattern and not on the values that are missing, as formalized by Rubin [46], [47]. For example, missing readings from one sensor may be compatible with MAR when their absence depends on temperature measurements reliably available from another channel. By contrast, missingness may be MNAR when a sensor becomes unavailable with a probability that depends on its own unobserved extreme measurement [45]. The classification depends on the information observed and the assumptions governing the missingness mechanism. MNAR cannot be tested from the observed data alone without further assumptions, because the distribution of the unobserved values is not identified by the observed entries [46]. Informative missingness, meaning that the mask is predictive of the outcome  $Y$ , is a different property from MNAR with respect to the features: a mask can predict failure while the missing values are MAR given the observed channels. When the mask is informative about  $Y$  and the mechanism is unchanged between training and deployment, imputation combined with the mask can remain consistent for supervised prediction [34]. An observed association between missing measurements and subsequent equipment failure is therefore neither sufficient to identify MNAR nor evidence against MAR [48].

Missingness that depends on the outcome behaves differently when the outcome is observed in training but absent when the prediction is made. In training data  $Y$  is known, so a mask that depends on  $Y$ , for example, a sensor that drops out before failure, can be learned as a predictor. At deployment  $Y$  is unobserved, so the model can use the mask only through the association learned in training, and that association breaks if the maintenance process, sensor hardware or acquisition policy changes; the resulting error is then a joint effect of  $M$  and  $S$ . Outcome-dependent missingness also cannot be identified from observed values alone. Evaluations should therefore state whether the mask is an input, whether its dependence on  $Y$  is assumed, and whether the deployment mask distribution equals the training one.

For sequences, the decision unit is a window of length  $w$  that ends at the decision time  $t$ , with prediction horizon  $h$  (the lead time), and the features are summaries of the window such as means, slopes and counts. Stressors act on the

raw channel-time entries before windowing, so missingness inside a window changes the window features (for example, a mean over fewer entries), and its temporal persistence (isolated entries versus outages of mean length  $l$ ) must be reported. For remaining-useful-life and time-to-event targets, the loss is defined on window-label pairs with censoring-aware labels. Windows from the same asset or run are never treated as independent, and the units of clustered inference are assets or runs.

### 3.2 Stressor Interactions, Non-Additivity and Attribution

The joint effects of sensor missingness, measurement corruption, and distribution shift can be examined through changes in predictive loss. However, an increase in loss under combined stressors does not necessarily indicate a statistical interaction. Identifying non-additive responses requires a predefined loss measure, comparable evaluation conditions, and explicitly specified perturbation mechanisms.

Consider a fixed predictive model evaluated under four conditions: unperturbed observations, missingness alone, measurement corruption alone, and their combination. Let  $L_{mns}$  denote the expected loss with stressor indicators  $m$  (missingness),  $n$  (measurement corruption) and  $s$  (shift), each equal to 1 when the stressor is present and 0 when absent. In this subsection, the shift factor is absent,  $s = 0$  and larger values of  $L$  indicate poorer performance. For controlled test-time perturbations, the losses should be evaluated on the same underlying test population, with consistent prediction targets, loss definitions, and treatment of randomness.

$$I_{MN} = L_{110} - L_{100} - L_{010} + L_{000}$$

where  $L_{000}$  is the baseline loss,  $L_{100}$  and  $L_{010}$  the losses under missingness alone and corruption alone, and  $L_{110}$  the loss under both. A positive contrast indicates that the combined loss exceeds the additive reference derived from the individual loss changes, whereas a negative contrast indicates less-than-additive degradation.

The interpretation also depends on whether the predictive model remains fixed across conditions. When controlled perturbations are applied to identical underlying test units without retraining, the contrast describes the response of a fixed prediction system to those input modifications. If models are independently retrained, recalibrated, or adapted under different conditions, the comparison additionally reflects changes in the learning procedure. Such results should be reported separately from fixed-model robustness evaluations.

The contrasts of this paper are defined within a test track and never across tracks. In the fixed-model track, one clean-trained model  $f$  is evaluated in all eight cells, and  $L_{mns}$  is the expected loss of  $f$  on the perturbed inputs. In a retrained track, a model  $f_{mns}$  is fitted for each cell, so  $L_{mns}$  is the loss of a train-and-deploy procedure and the contrast mixes learning-procedure effects with the response to the stressors; it is interpretable as the interaction of the procedure and is reported as such. The robust training track is therefore a legitimate but separate estimand. A comparison between tracks is a comparison of procedures, reported as a difference of losses or contrasts with their own clustered interval, and is not an interaction contrast.

**Figure 2** illustrates three patterns of predictive loss under sensor missingness ( $M$ ), measurement corruption ( $N$ ), and their joint occurrence. The dashed line is the additive reference  $L_{100} + L_{010} - L_{000}$ . The bar heights are arbitrary, and the axis values are omitted so that they are not read as data. In the additive response **Figure 2A**, the increase in loss under combined imperfections equals the sum of the individual increases relative to the unperturbed baseline, so the interaction contrast is zero ( $I_{MN} = 0$ ). In a synergistic response **Figure 2B** the combined loss exceeds this additive reference ( $I_{MN} > 0$ ); in an antagonistic response, **Figure 2C** it is below the additive reference ( $I_{MN} < 0$ ). The terms super-additive and sub-additive are used below only as synonyms for these two cases of the loss response, not in the sense of a property of a set function. Antagonism does not mean that the combination is harmless: the combined loss can still exceed either single-stressor loss. A greater loss under compound stressors therefore does not by itself constitute a statistical interaction, and the contrast depends on the chosen loss scale, on consistent evaluation conditions and on how the test population is treated. The values in the figure are arbitrary (baseline 1, missingness alone 2, corruption alone 1.8, hence an additive reference of 2.8), they are not shown on the axis, and they do not represent results from any study.

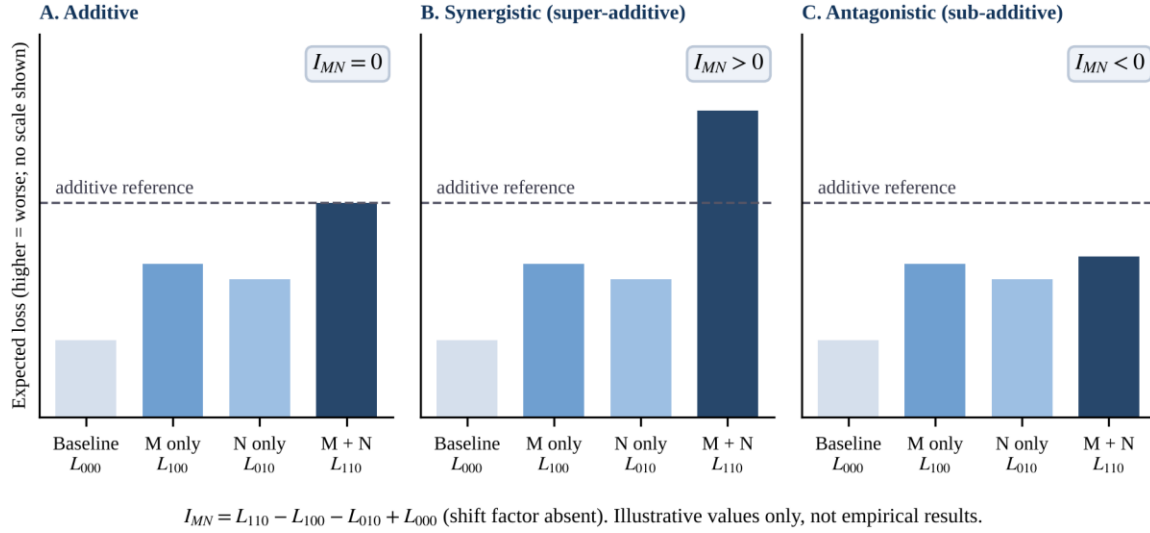


**Figure 2. Illustrative expected loss under individual and combined sensor missingness ( $M$ ) and measurement corruption ( $N$ ), with the shift factor absent. (A) Additive response. (B) Synergistic (super-additive) response. (C) Antagonistic (sub-additive) response**

For an illustrative observation model, distributional or measurement-system changes are defined first, recorded values may then be corrupted, and the observation mask finally determines availability. Under value-dependent missingness, the order can change the observed distribution and hence the interpretation of any proposed stressor contrast. This is a conceptual specification for future studies, not an implemented perturbation generator.

When missingness is independent of feature values (MCAR), or depends only on variables unaffected by corruption, applying missingness before or after measurement corruption produces equivalent observed inputs. In contrast, when the missingness mechanism depends on recorded values affected by corruption, the order of operations may change which measurements are observed under the combined-stressor condition. Consequently,  $L_{110}$  may differ between the two application orders, while  $L_{100}$  remains unchanged because no corruption is applied in the missingness-only condition. The interaction contrast is therefore order-dependent through  $L_{110}$ , potentially reflecting both the predictive response to combined imperfections and changes in the resulting observation mask. Accordingly, the order of operations and the variables governing the missingness mechanism must be explicitly specified when interpreting the contrast.

Industrial PdM datasets frequently contain dependent observations because of repeated measurements, overlapping windows, or shared equipment histories. Row-level resampling underestimates uncertainty when these dependencies are ignored, and cluster-aware resampling at the equipment, run, or site level is needed where the number and structure of independent units support valid inference. Perturbation severity and temporal persistence must also be documented, because identical missingness fractions or standardized noise magnitudes can represent very different physical disturbances or losses of predictive information.

**Figure 3** summarizes the proposed framework. It distinguishes sensor missingness ( $M$ ), measurement corruption ( $N$ ), and distribution shift ( $S$ ) and separates their evaluation into controlled and naturally occurring settings. In the controlled setting, a three-factor factorial design considers the eight binary stressor combinations, which allows individual effects and pairwise and three-way interaction contrasts to be assessed under comparable conditions. Naturally occurring shifts are examined through within-domain evaluations of missingness and corruption, followed by observational comparisons across operating environments. The controlled branch illustrates a possible eight-cell factorial design and theoretical interaction contrasts; the natural-domain branch illustrates within-domain contrasts and descriptive cross-domain comparison. Neither branch represents an experiment performed in this review.

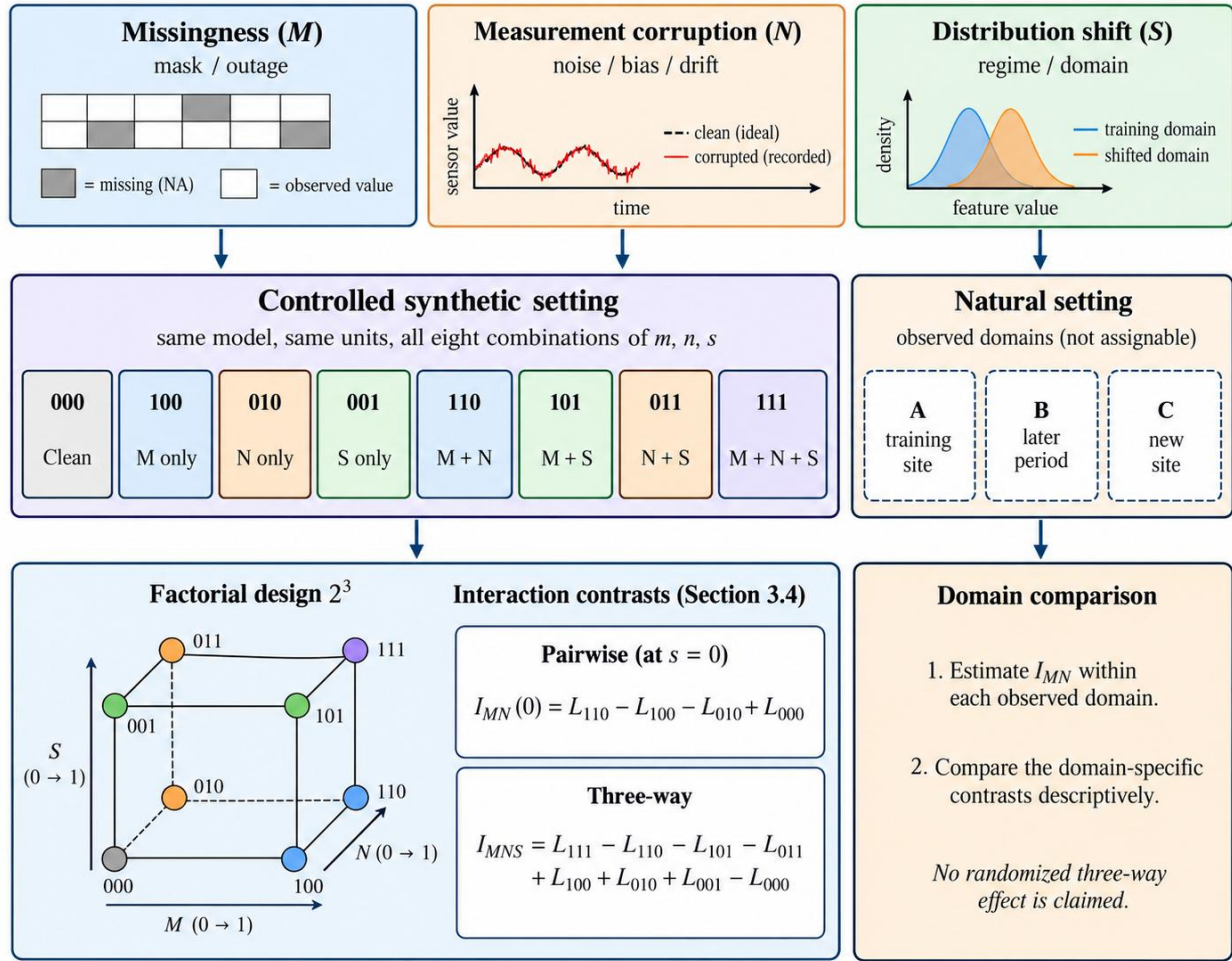


**Figure 3. Conceptual framework for future evaluation of sensor missingness, measurement corruption and distribution shift.**

### 3.3 Operational boundary between corruption and shift

The rule used here has three parts. (R1) A perturbation that changes the value, availability, or timing of individual measurements while training and test units share the same measurement system is  $M$  if it acts on the mask and  $N$  if it acts on the recorded value. (R2) A perturbation that changes the measurement system, operating regime, population or outcome mechanism consistently for all units of a domain, so that training and deployment environments differ, is  $S$ . (R3) An additive offset is  $N$  when it is a fault of an otherwise calibrated sensor that affects some units or episodes, and  $S$  when it is a persistent calibration change applied to the whole deployment domain. The severity, persistence, and unit of application of every offset must be reported. These categories are operational labels for proposed evaluation settings, not mutually exclusive physical mechanisms.

The operational distinction between measurement corruption and distribution shift depends on the scope and context of the perturbation rather than its mathematical form alone. Localized measurement faults are classified as corruption, whereas persistent changes between training and deployment environments are treated as distribution shifts. These categories may overlap in their underlying mechanisms. **Table 2** summarizes the resulting classification and the conditions under which prediction labels can be preserved.

**Table 2. Operational classification of data imperfections in industrial predictive maintenance.**

| Perturbation | Factor | Operational scope | Label preservation |
|--------------|--------|-------------------|--------------------|
|--------------|--------|-------------------|--------------------|

|                                                                        |   |                                                                           |                                                            |
|------------------------------------------------------------------------|---|---------------------------------------------------------------------------|------------------------------------------------------------|
| <b>Missing readings, channel outages, and temporal gaps [46], [47]</b> | M | Measurement availability within the evaluated environment                 | Yes, if only observations are removed                      |
| <b>Noise, spikes, bias, drift, and stuck-at readings [42], [44]</b>    | N | Recorded values affected by measurement errors or localized sensor faults | Yes, if the underlying state and target remain unchanged   |
| <b>Domain-wide recalibration or sensor replacement [49]</b>            | S | Persistent changes in the measurement function between environments       | Yes, if only the measurement process changes               |
| <b>Changes in operating regimes or equipment populations [43]</b>      | S | Differences in operating conditions or population distributions           | Not necessarily; depends on the target and shift mechanism |
| <b>Changes in sensor suites or acquisition policies [17]</b>           | S | Domain-level changes in feature availability or observation mechanisms    | Yes, if only data acquisition changes                      |
| <b>Changes in maintenance policies or degradation dynamics [19]</b>    | S | Changes in the outcome-generating mechanism                               | Not necessarily; target relationships may change           |

Note: M = missingness; N = measurement corruption; S = distribution shift. Categories are operational rather than mutually exclusive. The same measurement offset may be classified as N when it represents a localized sensor fault or as S when it represents a persistent measurement-system difference between environments. Label preservation refers to whether the original prediction targets remain valid under the specified perturbation.

### 3.4 Three-Factor Estimands and Identifiability

**Figure 3** refers to pairwise and three-way contrasts; this subsection defines them and separates three quantities that are easily confused. Let  $L_{mns}$  denote the expected loss, evaluated on common eligible decision units, under prespecified binary interventions on missingness ( $m$ ), measurement corruption ( $n$ ) and a synthetically manipulable shift operation ( $s$ ), where 1 denotes that the stressor is present and 0 that it is absent. Thus  $L_{000}$  is the clean baseline. Larger values indicate poorer performance, and  $L$  is the expected value of one primary loss scale (below).

**Total excess loss.** The damage caused by all three stressors together, relative to the clean baseline, is

$$D_{111} = L_{111} - L_{000}$$

The  $M$ – $N$  interaction at shift level  $s$  is

$$I_{MN}(s) = L_{11s} - L_{10s} - L_{01s} + L_{00s}$$

so that  $I_{MN}(0)$  equals the contrast  $I_{MN}$ . The other pairs are defined analogously with the third factor absent:

$$I_{MS}(0) = L_{101} - L_{100} - L_{001} + L_{000} \text{ and } I_{NS}(0) = L_{011} - L_{010} - L_{001} + L_{000}.$$

$I_{MN}(0)$  is the simple interaction at the clean level of the third factor. The averaged interaction,

$$\bar{I}_{MN} = \frac{1}{2}[I_{MN}(0) + I_{MN}(1)] = I_{MN}(0) + \frac{1}{2}I_{MNS},$$

averages over the two shift levels with equal weights. The simple version is the primary pairwise quantity because the clean domain is the reference condition of the protocol and the quantity reads directly as the interaction before shift. The averaged version answers a different question and depends on equal weights that have no operational justification when deployment has a single shift level. Both come from the same eight cells, and the choice is declared in advance.

The three-way interaction contrast is

$$I_{MNS} = L_{111} - L_{110} - L_{101} - L_{011} + L_{100} + L_{010} + L_{001} - L_{000}$$

which equals the difference between the  $M$ – $N$  interaction under shift and without shift:

$$I_{MNS} = I_{MN}(1) - I_{MN}(0)$$

A positive  $I_{MNS}$  therefore indicates that the  $M$ – $N$  interaction is more positive when the shift operation is present than when it is absent, on the chosen loss scale. It does not by itself imply that the total excess loss exceeds the sum of the three single-stressor excess losses.

That comparison is a different quantity,

$$A_{111} = D_{111} - (D_{100} + D_{010} + D_{001}) = L_{111} - L_{100} - L_{010} - L_{001} + 2L_{000}$$

where  $D_{100} = L_{100} - L_{000}$  and so on.  $A_{111}$  combines all three pairwise interactions with the three-factor term:

$$A_{111} = I_{MN}(0) + I_{MS}(0) + I_{NS}(0) + I_{MNS}$$

Consequently  $D_{111}$ ,  $A_{111}$  and  $I_{MNS}$  answer different questions and can differ in size and sign. A large total loss  $D_{111}$  may arise entirely from large single-stressor effects with no interaction ( $A_{111} = 0$ ), and a nonzero  $I_{MNS}$  may coexist with  $A_{111} = 0$  when pairwise and three-way terms offset one another. None of these algebraic identities imply a nonzero empirical effect.

Interaction contrasts quantify non-additivity; they do not attribute the total excess loss to the stressors. An attribution requires decomposition. Treating the three stressors as players and the excess loss  $v(T) = L_T - L_{000}$  of a coalition  $T$  as the value, the Shapley value [50] assigns to missingness

$$\varphi_M = D_{100} + (1/2)I_{MN}(0) + (1/2)I_{MS}(0) + (1/3)I_{MNS}$$

These contrasts have a causal reading only for interventions that can be assigned to matched units, with consistent target definitions, preserved outcome semantics, sufficient support and controlled nuisance variation. When  $S$  denotes a historical period, plant, asset population or maintenance policy, its apparent effect is confounded with unit composition and time and is an observational domain comparison, not an intervention. The defensible design then estimates the  $M$ – $N$  contrast separately within each observed domain and compares the domain-specific contrasts descriptively, without interpreting the between-domain difference as a randomized three-factor effect **Figure 3**, right branch. Likewise, a synthetic transformation of input columns is not automatically concept drift: a genuine change in  $P(Y|X)$  requires a change in the conditional outcome mechanism, which generally needs outcome data to establish.

The synthetic shift arm is restricted to two operators that preserve  $P(Y|X^*, Z)$ . (S1) Sensor recalibration offset for every unit of the deployment domain, a prespecified offset  $\kappa\sigma_j$ , with  $\sigma_j$  the training standard deviation, is added to a stated set of channels. The latent state and the labels are unchanged, so this is a shift in the measurement function **Table 2**, and it changes  $P(Y|X^{obs})$  by construction. (S2) Operating-regime covariate shift: deployment units are reweighted by  $q(z)/p(z)$  or subsampled so that the regime mix differs from training (for example, a higher share of high-load operation), with labels and features retained; it requires regime labels, such as the six operating conditions of C-MAPSS FD002 and FD004, and sufficient support in the training data. Neither operator represents concept drift or a change of maintenance policy, which are only available as natural domains. Every report states which operator was used and does not describe the result as robustness to distribution shift in general.

The contrasts are defined on expected loss, which suits per-decision proper scoring rules such as log loss and the Brier score; these decompose over decision units and are the primary scale here. The area under the precision–recall curve (PR-AUC), event recall at a fixed false-alarm budget, and cost-based metrics do not decompose by unit in this way. They are reported as secondary summaries, and an interaction found on one of them is labelled as scale-specific. PR-AUC also depends on class prevalence, which confounds comparisons across domains with different failure rates; across-domain comparisons should therefore report prevalence next to PR-AUC or use a prevalence-robust quantity such as recall at a fixed false-alarm rate.

### 3.5 Maintenance feedback, censoring and policy-dependent labels

Maintenance actions change the process that generates the labels. A component replaced after an alarm or on a schedule does not fail, so the failure the model was meant to predict is censored by the action, and the labels reflect the combination of equipment degradation and the maintenance policy in force. Three consequences follow. First,  $P(Y|X)$  depends on the policy: a change of inspection frequency, alert threshold, or maintenance contract between training and deployment is a shift in the outcome mechanism ( $S$  acting on  $P(Y|X)$ ) that cannot be created by perturbing inputs. Second, a model may suppress its own evidence: successful alerts lead to interventions that remove failures, so recall

on observed failures can fall without any deterioration of the model. Third, retraining on post-intervention labels can teach the policy instead of physics, when interventions are recorded among the features.

For evaluation, maintenance actions should be recorded with timestamps and post-intervention information excluded from the features; interventions should be treated as censoring events in time-to-event and RUL tasks, with censoring-aware metrics; targets should be defined by condition states or degradation thresholds that do not depend on the intervention where possible; and the policy in force should be reported for each domain. Counterfactual statements about what would have happened without an alert are outside the scope of an observational log unless an explicit causal design supports them.

## 4. MISSING FEATURES AND SENSOR AVAILABILITY

Missingness may occur as isolated unavailable measurements, intermittent interruptions, simultaneous losses across correlated channels, or persistent sensor outages. These patterns differ in their implications for learning algorithms [45]. Isolated missing values may leave sufficient information in neighboring observations or related variables, whereas prolonged channel unavailability can eliminate an entire measurement stream. Furthermore, missingness patterns may vary between training and deployment because of changes in acquisition procedures, sensor configurations, or operating environments. The proportion, duration, and location of missing observations should therefore be distinguished when interpreting robustness experiments [17].

Conventional approaches to incomplete tabular data include complete-case analysis, constant or statistical imputation, nearest-neighbor methods, iterative multivariate imputation, and model-based procedures [51]–[53]. These approaches address different objectives and rely on different assumptions. Complete-case analysis excludes observations with missing features and may reduce the available sample or alter the population represented by the retained observations. Mean or median imputation provides a simple replacement strategy but does not reconstruct the original joint distribution of the missing variables. More elaborate approaches use relationships among observed features to estimate unavailable values [52], [53].

Missingness can also be incorporated into prediction without explicitly reconstructing every unavailable measurement. Some learning methods can process missing entries through model-specific mechanisms. For example, the sparsity-aware split procedure in XGBoost learns default split directions for missing feature values [54]. Alternatively, missingness indicators can be included as additional predictors, allowing a model to use information about observation availability. For temporal data, approaches such as GRU-D incorporate missingness masks and elapsed time from observations, although the evidence supporting such mechanisms should be interpreted within the sequential prediction settings in which they were evaluated [32].

These alternatives do not establish that native missing-value handling or missingness-aware architectures are universally preferable to imputation-based pipelines. Their performance depends on the prediction objective, model specification, missingness mechanism, and information available during training and deployment. A predictive relationship associated with missingness patterns in the training data may not remain stable if sensor failures, acquisition policies, or maintenance procedures change. Comparisons should therefore distinguish the information represented by observed measurements from that conveyed by missingness patterns themselves [34].

Temporal dependencies introduce additional considerations when missing measurements arise in sequential industrial records [17]. Forward filling, interpolation, and other time-dependent reconstruction procedures may use neighboring observations, but their suitability depends on the prediction setting. For real-time failure prediction, imputations must not use future measurements that would be unavailable at the prediction time [55]. Similarly, preprocessing parameters and imputation models should be estimated using the appropriate training data rather than information from held-out evaluation periods or equipment units [55]. These restrictions are relevant to avoiding information leakage and ensuring that evaluated prediction pipelines reflect the intended deployment conditions.

#### ***4.1 Missingness-aware deep learning versus imputation-based prediction***

Deep learning approaches to incomplete observations differ in how missing information is represented and incorporated into prediction. Relevant strategies include reconstructing unavailable values before downstream learning, predicting directly from observed features and missingness indicators, and exploiting temporal dependencies alongside observation-availability patterns. These categories are not mutually exclusive, since some architecture combines imputation and prediction within a shared learning framework. Their comparison therefore requires consideration of learning objectives, statistical assumptions, and the information available at training and inference [34], [35].

Generative imputation methods seek to reconstruct unavailable measurements using relationships learned from incomplete data. GAIN uses adversarial learning, which means that a generator guesses missing values and a discriminator uses a hint mechanism to tell the difference between observed and imputed values [36]. MIWAE instead uses deep latent-variable modeling and an importance-weighted objective to learn from incomplete observations under a missing-at-random assumption, supporting both single and multiple imputations [33].

NeuMiss networks address supervised prediction directly. They are motivated by the Bayes predictor of a linear model with Gaussian features and missing values, which they approximate with a truncated Neumann series implemented as a differentiable network whose weights are shared across missingness patterns; the authors report that the approach is robust to the missingness mechanism in their experiments [35]. This is a supervised, mask-conditional design and differs from imputation followed by a separate learner.

Josse et al. [34] study the consistency of supervised learning with missing values. Their results indicate that, asymptotically and for sufficiently flexible learners, imputation by a constant such as the mean, combined with the missingness mask where missingness is informative, can approach the Bayes-optimal predictor, and they analyses tree methods that handle missing values natively. These are consistency results. They do not establish finite-sample performance, and they assume that the missingness mechanism at deployment equals the one in training, so they do not cover a change of the mask distribution.

Taking together, these sources define a comparison that industrial practice rarely makes simple imputation with a mask, native missing-value routing in boosted trees, mask-conditional architectures such as NeuMiss, and deep generative imputation (GAIN, MIWAE) followed by a downstream learner. The consistency results suggest that simple imputation plus a mask is a strong control that more expensive imputers have to beat. The generative methods optimize reconstruction under a missing-at-random assumption, which does not by itself imply better sensitivity to rare failures. Whether the ordering of these options is stable under realistic sensor outages in maintenance data has not been tested in the evidence examined.

#### ***4.2 Industrial evidence on missingness mechanisms and severity (RQ1)***

RQ1 asks about the mechanisms, severity, and links to equipment health of missingness in industrial data. The evidence examined answers the conceptual part better than the empirical part. Natural missingness is present in the APS Scania data, but the variables are anonymized, and the study that reports its preprocessing treats the gaps as a cleaning step: variables with more than 10% missing values are dropped, and the rest are imputed [25]. The MetroPT-3 repository describes the released file as having no missing values; some valid operating states are encoded by zeros and must not be treated as unavailable measurements [37], [38]. AI4I-PMDI injects missing and irregular observations into a synthetic dataset [16], and sensor-failure experiments on C-MAPSS zero out channels or readings [29]. Severity is therefore chosen by the experimenter rather than measured, and none of the examined studies relate observed missing rates or patterns to equipment condition or time to failure. General sources note that sensor degradation and imperfect measurements affect availability [44], [45], and one recent study proposes an imputer for informative (MNAR) missingness in maintenance data [48], but it does not identify the mechanism in a deployed system.

The empirical characterization of missingness mechanisms and severity in industrial data is thus an open gap (identified as a gap in this review), and findings from the controlled experiments reviewed here should be read as sensitivity analyses over assumed mechanisms, not as descriptions of field conditions.

## 5. MEASUREMENT CORRUPTION AND SENSOR INTEGRITY

In industrial predictive maintenance (PdM), sensor data reliability is a crucial problem since the measured values may not match the actual physical quantities because of sensor failures, calibration errors, acquisition failures or communication problems. Unlike missingness, measurement corruption can result in the observation of a recorded value, but with reduced precision or reliability. It should also be distinguished from true changes in equipment performance which can produce unusual but physically valid results. Hence, the detection of an anomaly can be a sign of reduced measurement accuracy, changes in operating conditions, device wear or a combination of these factors [44] - [45]

These distinctions are important for understanding predictive resilience. A model's response to artificially degraded inputs does not necessarily determine its behavior under naturally occurring sensor failures. The significance of documented results depends on the corruption mechanism, its physical plausibility, its temporal characteristics, and its link to the core prediction task. This part analyzes these factors and distinguishes resilience to artificial feature changes and resilience to disturbance of measurements in industrial sensing systems.

### *5.1 Mechanisms of measurement corruption and physical plausibility*

Measurement distortion may arise from stochastic noise, systematic bias, sensor drift, spikes, saturation, quantization errors or fixed readings. These disruptions have different statistical properties and physical origin. Measurement corruption can include stochastic noise, systematic bias, and temporally persistent sensor faults. In an industrial environment, the degree of errors can be affected by signal strength, load on the operation, temperature and environment, resulting in heteroscedastic variations. Errors may also be correlated across sensor channels due to shared acquisition equipment, common calibration problems, or environmental influences. Continued sensor drift and occasional instability are not like independently generated errors, and temporal dependencies add another level of complexity.

Acquisition problems rather than the numerical precision of individual measurements may compromise temporal synchronization or data access (timestamp inaccuracies, inconsistent sampling, communication lags). These disruptions may impact the data that feeds into the predictive models but should not be considered the same type of corruption mechanisms. The outcome depends on the affected channels, the intensity and duration of the disturbance, and the operational environment [56].

This is especially important in the test of robustness where physical realizability of synthetic disturbances is crucial. The corruption mechanisms should specify the engineering units, accuracy of sensors, operational range, sampling characteristics and known measurement limitations. Feature semantics also constrain allowable transformations. Continuous measurements (e.g. temperature, pressure) are suitable for disturbances but not for categorical operating modes, cumulative counters or variables constrained by physical limitations. Furthermore, directly injecting corruption into designed tabular features does not necessarily mimic the effect of changing the raw sensor signals before feature extraction [44].

For ongoing measurements, the observed value can be expressed as the true quantity plus temporal bias and stochastic error in an additive error model. But this formulation is not general enough for all sensor faults. Saturation, clipping, quantization, stuck-at behavior and temporal misalignment for example, may need different measurement functions. The intensity of perturbation should be described by engineering metrics or reference statistics computed from training data alone, including cross-channel and temporal dependencies when physically justified.

Label preservation requires careful consideration. If artificial disturbances are to be understood as inaccuracies of measurement and not as changes in the fundamental condition of the apparatus, then it may be appropriate to retain the original prediction objective. In contrast, changes in the actual operating load, speed, temperature or degradation patterns will affect failure probabilities or the remaining useful life. Such interventions cannot be indiscriminately labeled as label-preserving measurement distortion. Understanding robustness results requires distinguishing changes in recorded quantities from changes in the underlying physical process.

### *5.2 Robustness of learned representations versus robustness to sensor faults*

Robustness to synthetic feature perturbations during representation learning should be distinguished from robustness to physical sensor faults in industrial environments. Self-supervised tabular learning methods often employ feature masking or corruption to construct training objectives. For example, VIME combines mask estimation and value imputation as self-supervised pretext tasks [57], while SCARF learns contrastive representations through random feature corruption [58]. TabTransformer has also been evaluated on general-purpose tabular classification tasks involving missing or corrupted features [59]. Although these studies provide insights into model behavior under the evaluated perturbations, their findings do not directly establish robustness to persistent sensor drift, stuck-at readings, temporal misalignment, or correlated measurement errors in industrial predictive maintenance.

Additional evidence is provided by research on fault diagnosis of industry. Tang [29] studies simulated sensor failures in C-MAPSS, such as channel zeroing, sporadic measurement corruption, bias and drift, using a generative repair phase and a CNN-LSTM model for remaining useful life prediction. The research demonstrates the use of controlled sensor-fault injections for sequential prognostics. However, the results should not be generalized to naturally occurring anomalies or tabular model classes without proper evaluations. Similarly, domain-adversarial fault diagnosis in different bearing operating conditions is associated with a different evaluation framework than measurement degradation in designed industrial tabular attributes [28].

The differences highlight the need for a proper fit of corruption frameworks to the context of their intended implementation. Robustness tests should consider whether the disruptions are during pretraining applied to raw sensor data or directly imposed on designed features. They need to distinguish between inaccuracies in a single measurement and continuous or correlated faults, and to assess whether the disturbances preserve physically valid inputs and prediction objectives.

At present, the evidence supports the need for assessment of robustness against specific mechanisms, rather than the blanket assertion that models with feature masking, noise-augmentation, or representation-learning objectives are robust to industrial sensor malfunctions. The examination of resilience during genuine measurement corruption, especially when corruption overlaps with absence and distributional changes, remains a significant agenda for future research.

## **6. DISTRIBUTION SHIFT AND CONCEPT DRIFT IN INDUSTRIAL PREDICTIVE MAINTENANCE**

Distribution shift is an important challenge in industrial predictive maintenance (PdM), as the statistical properties of operational and sensor data may change between model development and deployment. Such changes can arise from variations in operating conditions, equipment characteristics, degradation processes, maintenance practices, or measurement systems. From a statistical perspective, these changes may affect the distribution of input features, the distribution of maintenance outcomes, or the relationships between them. Their implications for predictive performance depend on the underlying shift mechanism and the stability of the relationship between observed measurements and maintenance outcomes.

Importantly, changes in sensor measurements do not necessarily indicate changes in equipment health or failure mechanisms. Observed distributional differences may originate from the physical process, the measurement system, or both. In industrial settings, distinguishing these sources is essential for interpreting performance degradation and determining whether data correction, model recalibration, or adaptation may be appropriate. The following subsections connect the statistical definitions of distribution shift to changes in industrial operating regimes and discuss the limitations of identifying shift mechanisms from observational data.

### ***6.1 Types of distribution shift and changes in operating regime***

Distribution shift refers to differences between the joint data distributions of training and deployment environments. Covariate shift occurs when the input distribution changes while the conditional distribution of outcomes given inputs remains stable. Label shift involves changes in the outcome distribution while the conditional distribution of inputs

given outcomes remains unchanged. Concept drift, also referred to as conditional shift, involves changes in the relationship between input features and prediction targets. These distinctions provide a statistical basis for interpreting predictive performance across environments.

In industrial PdM, such changes may arise from variations in equipment loads, environmental conditions, degradation states, maintenance interventions, or differences across machines and production sites. Their effects may be reflected in feature distributions, dependencies among sensor measurements, or relationships between observed conditions and future failures. However, changes in operating regimes do not necessarily imply concept drift, since the underlying relationship between equipment condition and maintenance outcomes may remain stable.

Evidence from general-purpose tabular benchmarks illustrates the importance of accounting for these differences in evaluation design. TabReD examines eight industry-grade tabular datasets and demonstrates the relevance of time-based evaluation and feature-rich representations for comparisons of approaches including gradient-boosted decision trees and MLP-like neural architectures [14]. WhyShift investigates naturally occurring distribution shifts across five tabular datasets, highlighting the importance of conditional distribution changes beyond differences in input distributions alone [19]. Although these studies provide methodological insights into evaluating distribution shift, their findings should be interpreted within their respective prediction tasks and cannot be directly generalized to industrial sensor failures or maintenance-specific deployment conditions.

## ***6.2 Changes in Observation Systems Versus Changes in Failure Mechanisms***

An important distinction in industrial PdM is whether observed distributional changes originate from the measurement system or from changes in the underlying equipment and failure processes. Sensor replacement, recalibration, acquisition-system modifications, or changes in feature extraction may alter the relationship between the underlying equipment state and recorded measurements without necessarily changing the physical degradation process [49]. Similarly, changes in sensor availability or observation patterns may modify the distribution of recorded features without corresponding changes in equipment health.

In contrast, changes in equipment utilization, degradation dynamics, operating conditions, or maintenance policies may alter the relationship between observed conditions and future maintenance outcomes. These mechanisms can produce overlapping changes in feature distributions, making their effects difficult to distinguish from changes in the observation process. Furthermore, measurement-system changes and physical-process changes may occur simultaneously, complicating the interpretation of model performance under distribution shift [45].

Identifying the underlying shift mechanism is particularly challenging when labeled deployment data are unavailable or delayed. Although changes in observed feature distributions can be detected using unlabeled measurements, these observations alone generally cannot establish whether the conditional relationship between inputs and failure outcomes has changed. Distinguishing label shift from other distributional changes also requires additional assumptions. These limitations are especially relevant to industrial PdM, where failures may be rare, equipment histories incomplete, and observed outcomes influenced by maintenance interventions.

Consequently, detecting distribution shift should not be equated with identifying its physical or operational cause. Statistical drift detection can indicate deviations from a reference distribution, but it cannot independently determine whether they result from sensor faults, operating-regime changes, or altered failure mechanisms. Reliable interpretation therefore requires complementary evidence, including equipment metadata, sensor calibration records, operating histories, maintenance information, and subsequent failure observations. Such evidence is also important when deciding whether data correction, recalibration, or model adaptation is justified.

## **7. WHAT EXISTING EVIDENCE ESTABLISHES AND WHAT IT DOES NOT**

The reviewed literature provides evidence from complementary but methodologically heterogeneous research settings. TableShift, TabReD, and TabFSBench examine different aspects of distributional and feature-set changes [13]–[15], while AI4I-PMDI investigates irregular observations in a synthetic-derived maintenance setting [16]. Other studies

address test-time adaptation [31], [60], natural distribution shifts [19], or robustness to missing and corrupted features [59]. However, differences in prediction tasks, data representations, stressor definitions, and validation protocols restrict direct comparisons across these findings. The following synthesis therefore examines the applicability and inferential limits of the available evidence rather than combining reported results into a common model ranking.

### 7.1 Primary-Evidence Audit

**Table 3** presents a curated evidence map of selected studies relevant to data imperfections and predictive robustness in industrial maintenance and related tabular learning. The studies are classified according to the evidence categories defined in Section 2.1: (A) industrial tabular maintenance studies, (B) industrial signal-based and sequential studies, (C) general-purpose tabular learning studies, (D) adjacent methodological studies, and (R) dataset and benchmark resources.

For each study, the table summarizes the data or prediction task, the evidence examined, the scope of the data investigated imperfections, and the principal limitations affecting the interpretation of its findings. These distinctions help separate application-specific results from broader methodological contributions and dataset documentation. Experiments involving artificial feature masking, missing-value imputation, or individual sensor perturbations should not be interpreted as evidence of robustness under combined industrial data imperfections unless such conditions were explicitly evaluated. The table provides a qualitative synthesis rather than a quantitative meta-analysis or a standardized assessment of methodological quality.

**Table 3. Critical evidence map of data imperfections and robustness evaluation in industrial predictive maintenance and related tabular learning.**

| Study     | Cat. | Data / task                   | Evaluated evidence                                                                          | Stressor scope                | Limits of inference                                                                                                                               |
|-----------|------|-------------------------------|---------------------------------------------------------------------------------------------|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| [16]      | A/R  | Synthetic-derived PdM data    | Missing observations, irregular acquisition, and classification experiments                 | Multiple irregularities       | Injected irregularities in synthetic-derived data limit direct generalization to naturally occurring industrial acquisition failures.             |
| [26]      | A    | AI4I 2020                     | Effects of leakage-aware versus leakage-prone evaluation on reported predictive performance | Evaluation methodology        | Findings concern the influence of validation design on reported performance rather than sensitivity to sensor imperfections.                      |
| [27]      | A    | APS Scania                    | Random forest prediction with engineered features and cost-sensitive evaluation             | Existing dataset conditions   | Cost-sensitive evaluation supports maintenance-related assessment, but anonymized features restrict interpretation of sensor-specific mechanisms. |
| [25]      | A    | APS Scania / SECOM            | Missing-feature preprocessing and cost-sensitive Transformer-based prediction               | Existing missing data         | Feature removal and imputation are preprocessing choices; their individual contributions to predictive performance are not isolated.              |
| [37]-[38] | R    | Compressor sensor time series | Real asset measurements and maintenance-event documentation                                 | Data resource                 | Provides operational data and event records, but documentation alone cannot establish comparative predictive performance.                         |
| [28]      | B    | Bearing vibration signals     | Cost-sensitive domain adaptation across operating conditions                                | Operating-condition variation | Evidence is specific to signal-based diagnosis under changing operating conditions rather than engineered tabular predictors.                     |
| [29]      | B    | C-MAPSS FD001                 | Evaluation of sensor zeroing, bias, and drift at different severities                       | Injected sensor faults        | Controlled fault injections support severity analysis, but their                                                                                  |

|      |   |                                |                                                                         |                                  |                                                                                                                                                           |
|------|---|--------------------------------|-------------------------------------------------------------------------|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
|      |   |                                |                                                                         |                                  | correspondence to naturally occurring sensor degradation is uncertain.                                                                                    |
| [5]  | B | C-MAPSS / N-CMAPSS             | Evaluation of combined temporal irregularities and sensor perturbations | Multiple injected imperfections  | Combined perturbation experiments do not separately quantify pairwise and three-way effects; conclusions remain specific to sequential RUL prediction.    |
| [30] | B | MetroPT-3 sequences            | Anomaly detection using recurrent and autoencoder-based approaches      | Anomaly detection                | Anomaly-detection performance does not independently characterize sensitivity to specific missingness or corruption mechanisms.                           |
| [13] | C | General binary tabular tasks   | Evaluation under naturally occurring distribution shifts                | Natural distribution shift       | Natural domain shifts support distributional robustness analysis but do not represent industrial measurement faults.                                      |
| [14] | C | Non-maintenance tabular tasks  | Temporal evaluation and feature-rich benchmarking                       | Temporal evaluation              | Time-based evaluation improves temporal relevance, but the tasks differ from industrial failure prediction and maintenance decisions.                     |
| [15] | C | General tabular datasets       | Model evaluation under feature-set changes                              | Feature shifts                   | Feature-set changes characterize input availability rather than the physical processes responsible for industrial sensor degradation.                     |
| [31] | C | General tabular tasks          | Test-time adaptation under decreasing feature availability              | Feature removal                  | Adaptation performance depends on the evaluated feature-removal settings; deployment constraints and maintenance costs require separate assessment.       |
| [19] | C | Natural-shift tabular datasets | Analysis of distribution changes and conditional-shift effects          | Natural distribution shift       | Observed population and conditional shifts do not identify the underlying physical causes of industrial measurement changes.                              |
| [59] | C | General tabular classification | Model evaluation with noisy or missing features                         | Feature corruption / missingness | Artificial feature perturbations provide task-specific evidence without establishing correspondence to sustained industrial sensor faults.                |
| [57] | C | General tabular learning       | Self-supervised learning using feature masking and value estimation     | Pretraining perturbations        | Training-time masking primarily supports representation learning and does not directly characterize deployment-time sensor outages.                       |
| [58] | C | General tabular learning       | Contrastive learning using artificial feature corruption                | Pretraining perturbations        | Random feature corruption is a representation-learning operation rather than a physically specified measurement-fault process.                            |
| [61] | C | General tabular tasks          | In-context prediction and benchmark performance                         | General benchmarking             | Reported benchmark performance does not characterize model behavior under persistent sensor unavailability or maintenance-specific evaluation conditions. |

|      |   |                                              |                                                               |                          |                                                                                                                                               |
|------|---|----------------------------------------------|---------------------------------------------------------------|--------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| [62] | C | Larger general tabular tasks                 | Extended tabular modeling capabilities                        | General benchmarking     | Expanded modeling capabilities do not provide direct evidence about robustness to industrial sensor degradation.                              |
| [63] | C | Large-scale tabular tasks                    | Scalable in-context tabular learning                          | General benchmarking     | Scalability findings do not address the effects of incomplete or unreliable industrial measurements on maintenance prediction.                |
| [35] | D | Supervised learning with incomplete features | Missingness-aware predictive architecture                     | Missingness              | Results concern prediction with incomplete features; stability under changing industrial observation mechanisms requires separate evaluation. |
| [34] | D | Missing-data learning theory                 | Theoretical results concerning prediction with missing values | Methodological theory    | Theoretical consistency results do not determine finite-sample performance when missingness mechanisms change between environments.           |
| [36] | D | Incomplete tabular datasets                  | Adversarial missing-value imputation                          | Missingness / imputation | Reconstruction quality does not directly determine the preservation of failure-relevant information in downstream prediction.                 |
| [33] | D | Incomplete datasets                          | Latent-variable modeling and imputation under MAR assumptions | Missingness / imputation | Its missing-data assumptions constrain interpretation when industrial observation mechanisms are informative or change across environments.   |
| [32] | D | Clinical multivariate time series            | Prediction using missingness masks and observation intervals  | Temporal missingness     | Clinical sequence results depend on domain-specific observation processes and cannot be directly transferred to industrial maintenance tasks. |

Note: A = industrial tabular maintenance; B = industrial signal-based or sequential; C = general-purpose tabular learning; D = adjacent methodological research; R = dataset or benchmark resource. The stressor-scope column describes the type of evidence rather than a standardized interaction test. Multiple perturbations within a study do not necessarily establish factorial interactions. Categories indicate evidence applicability, not methodological quality.

## 7.2 Studies on the benchmark maintenance datasets

Maintenance datasets are widely used, but the studies examined here rarely test the stressors of this review. For AI4I 2020, classifier comparisons are common; Qahtani [26] compares gradient boosting, random forest and logistic regression under leakage-aware cross-validation, compares estimates obtained under leakage-aware and leakage-prone validation, underscoring the importance of avoiding information leakage. The dataset has no missing values by construction, so such studies evaluate clean data. For APS Scania, the challenge cost metric motivated random forests with feature engineering [27] and cost-sensitive deep models [25]; the missing entries are natural and their treatment, dropping variables with more than 10% missing values and imputing the rest in one study [25], is a preprocessing choice and not an experimental factor. For C-MAPSS, Tang [29] injects channel loss, intermittent zeroing, bias and drift at three severities and reports that a generative repair stage followed by a CNN–LSTM keeps remaining-useful-life error low.

Sequence-based studies on C-MAPSS and MetroPT-3 provide additional industrial evidence. Zhuang et al. [5] evaluate a mask-aware continuous-time RUL model under injected observation irregularities in C-MAPSS and N-CMAPSS. A MetroPT-3 study [30] evaluates anomaly-detection approaches on compressor sensor sequences; its reported findings should be interpreted within the study’s preprocessing and evaluation setting.

**Table 4** summarizes the application context, evaluated conditions, and principal inferential limitations of these industrial studies. Differences in prediction targets, data representations, preprocessing choices, and validation

procedures prevent direct numerical comparisons across studies. Overall, the reviewed evidence supports task-specific conclusions and the importance of leakage-aware evaluation, but it does not establish a controlled three-factor robustness assessment of missingness, measurement corruption, and distribution shift in industrial tabular predictive maintenance.

**Table 4. Critical assessment of industrial predictive maintenance studies addressing data imperfections**

| Study | Data / Task                           | Evaluated condition                                  | Models/ approach                                      | Reported evidence                                    | Main limitation                                                                                                                                           |
|-------|---------------------------------------|------------------------------------------------------|-------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| [26]  | AI4I 2020 / Failure classification    | Leakage-aware versus leakage-prone validation        | Gradient boosting, RF, logistic regression            | Evaluation design influences reported discrimination | Findings concern validation sensitivity in synthetic data rather than sensor-level robustness.                                                            |
| [27]  | APS Scania / Failure classification   | Existing dataset characteristics and challenge costs | RF with engineered features                           | Cost-oriented failure prediction                     | Anonymized measurements restrict physical interpretation of individual sensor mechanisms.                                                                 |
| [25]  | APS Scania and SECOM / Classification | Missing-value preprocessing and class imbalance      | Cost-sensitive Transformer, imputation and resampling | Cost-sensitive evaluation and component ablation     | Missing-data handling is examined within the modeling pipeline rather than through independently controlled missingness mechanisms and severities.        |
| [29]  | C-MAPSS FD001 / RUL                   | Simulated channel faults, bias and drift             | Generative repair with CNN-LSTM                       | Reported RUL prediction under sensor disturbances    | Injected faults permit controlled evaluation, but their physical correspondence to naturally occurring sensor degradation remains uncertain.              |
| [5]   | C-MAPSS and N-CMAPSS / RUL            | Simulated irregular observations and data scarcity   | PC-MambaSDE and comparative baselines                 | RUL evaluation under irregular sensing conditions    | Findings concern sequential RUL prediction; the separate contributions of interacting irregularity mechanisms are not established by factorial contrasts. |
| [30]  | MetroPT-3 / Anomaly detection         | Reported preprocessing and anomaly detection         | LSTM, autoencoder and hybrid approaches               | Anomaly-detection evaluation                         | Reported anomaly-detection performance does not independently characterize sensitivity to controlled missingness and corruption mechanisms.               |

### 7.3 Recurring inferential errors

Four category errors occur. First, randomly masking features for contrastive or denoising pretraining is not evidence of robustness to sustained outages at test time. Second, smaller imputation error is not necessarily better rare-failure sensitivity, because an imputer can smooth away mechanically informative extremes. Third, improvements reported for convolutional, recurrent or adversarial-adaptation models on raw bearing signals cannot rank attention models against boosted trees on static maintenance tables. Fourth, an accuracy change between naturally shifted domains is not a causal three-way corruption interaction. Each claim requires task-matched held-out outcomes and a documented stressor mechanism.

### 7.4 Distinguishing Established Findings from Future Research Directions

The findings of this review can be interpreted at three levels that collectively address the five research questions. First, the literature provides a conceptual understanding of sensor missingness, measurement corruption, and distribution shift (RQ1), although empirical evidence linking naturally occurring data imperfections, their severity, and equipment health remains limited. Existing studies also provide task-specific evidence on model robustness (RQ2), but differences in datasets, prediction objectives, and evaluation conditions prevent reliable comparisons across model families.

Second, the reviewed studies highlight the methodological challenges of evaluating combined data imperfections (RQ3). The evidence examined does not establish controlled three-factor interaction effects in industrial tabular PdM. To support future investigations, this review proposes a mechanism-aware framework and pairwise and three-way loss contrasts for examining potential non-additive responses. These are analytical tools for future evaluation rather than empirically validated findings.

Third, dataset characteristics, information leakage, and validation strategies influence the interpretation and comparability of reported results (RQ4). Reliable evaluation also requires attention to predictive uncertainty, calibration, event-level performance, and maintenance-related costs (RQ5). The proposed leakage-aware protocol integrates these considerations into a structured approach for future research. Its practical feasibility and effectiveness remain subject to empirical investigation.

## 8. MODEL-FAMILY COMPARISON UNDER IMPERFECT INDUSTRIAL DATA

Model comparison must use task-matched baselines. Grinsztajn et al. identify inductive-bias advantages of tree ensembles on conventional tabular benchmarks [11], and McElfresh et al. show that dataset properties and tuning conditions complicate universal tree-versus-neural claims [12]. TabReD reports that, under time-based splits and feature-rich pipelines, gradient-boosted decision trees and simple MLP-like models are among the strongest tested approaches [14]. That is a meaningful but benchmark-specific comparison; it does not establish which family withstands a jointly missing, corrupted and shifted industrial sensor stream.

XGBoost uses a sparsity-aware split algorithm that learns a default direction for missing values [54], and CatBoost uses ordered boosting and ordered target statistics that can mitigate certain target-leakage-related biases [64]. CatBoost's ordered boosting addresses prediction shift caused by target statistics computed on the training rows themselves; it is a learner-internal bias correction and does not prevent the temporal or target leakage of Section 10.1, which depends on split design and feature construction. Neither study establishes superior behavior for permanently absent industrial sensors, MNAR mechanisms or changing feature schemas. The value of a native missing-value branch depends on the training mask distribution, signal redundancy and schema consistency, and it should be compared with an imputer-plus-indicator baseline and with training under matched perturbations.

MLP and ResNet-style networks provide flexible smooth representations but typically need tuned preprocessing and regularization. TabNet uses sequential attentive feature selection [65], FT-Transformer builds contextual feature interactions through feature tokenization [66], and TabTransformer focuses on contextual categorical embeddings [59]. Attention mechanisms are not certificates of corruption tolerance or interpretability, and feature tokenization does not reconstruct a deleted critical channel. TabTransformer's missing- and noisy-feature experiments are in general tabular tasks and should be cited as such.

TabPFN reports strong accuracy and favorable speed-to-tune trade-offs on small standard tabular benchmarks, with a primary evaluation of datasets of up to 10,000 examples and 500 features and with disclosed memory-scaling and inference-time limits [61]. These limits are version specific. The later TabPFN-2.5 report states support for up to 50,000 rows and 2,000 features [62], and TabICL is pretrained on synthetic datasets of up to 60,000 samples and is reported to scale to 500,000 samples [63]. Drift-Resilient TabPFN extends the in-context approach to temporal distribution shift [67], which is a different problem from sensor outage or corruption. These models are applicability data points, not evidence of fragility or strength under sustained industrial outages. Each model version should be named and evaluated separately, none should be presented as validated for streaming, heavily missing or novel-fault settings without task-specific experiments, and the version, date and access route of every foundation model evaluated should be recorded, because the field changes quickly. No foundation model was run in this paper; the version

statements above refer to the cited reports, the first of which describes the model published in Nature (TabPFN v2) and the later one TabPFN-2.5.

Fair comparisons across model families require consistent prediction targets, available information, validation procedures, preprocessing, and hyperparameter-tuning budgets. Clean-trained models, robustness-trained approaches, and online adaptation methods should be evaluated in clearly distinguished tracks because they involve different training assumptions and access to deployment data. Temporal models must use only information available at the prediction time, and preprocessing or imputation procedures must avoid information leakage. Importantly, the reviewed evidence does not establish a universal robustness ranking between gradient-boosted trees and deep learning under compound industrial data imperfections. Accordingly, **Table 5** summarizes supported methodological properties and unresolved evaluation questions rather than comparative performance claims.

**Table 5. Model-family evidence, unresolved robustness questions, and recommended evaluation controls.**

| Model family                                             | Established evidence                                                           | Unresolved robustness questions                                                    | Recommended controls                                                         |
|----------------------------------------------------------|--------------------------------------------------------------------------------|------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| <b>GBDT — XGBoost [54], CatBoost [64]</b>                | Strong general tabular baselines; native missing-value handling in XGBoost     | Changing missingness mechanisms, sustained outages, calibration under shift        | Native missing handling versus imputation with indicators; matched tuning    |
| <b>MLP / ResNet-like [66]</b>                            | Competitive performance in selected tabular benchmarks, including TabReD [14]  | Dependence on critical features, scaling and corruption severity                   | Matched features, preprocessing, masks and tuning budgets                    |
| <b>Attention-based tabular DL [59], [65], [66]</b>       | Feature-selection and contextual representation mechanisms                     | Robustness under compound industrial imperfections                                 | Identical tasks, splits, available information and calibration procedures    |
| <b>Tabular foundation models [61]–[63]</b>               | Strong reported performance in supported tabular benchmark settings            | Persistent missing sensors, unseen regimes, latency and deployment constraints     | Version-specific comparisons with tuned GBDTs; matched resources             |
| <b>Imputation / Missingness-aware DL [33], [34]–[36]</b> | Methods for reconstructing missing values or predicting with incomplete inputs | Whether reconstruction improves rare-failure prediction under changing missingness | Native missing handling; simple imputation plus masks; downstream evaluation |
| <b>Adaptation / Signal-domain DL [28], [31], [60]</b>    | Methods targeting specified distribution or feature-availability changes       | Negative adaptation, novel faults and unreliable target observations               | Fixed target-data access, leakage controls and no-adaptation baseline        |

## 9. INDUSTRIAL DATASETS: SUITABILITY, LEAKAGE AND TRANSFER LIMITS

The datasets examined in this review do not establish a validated joint evaluation of all three imperfections and their interactions. A credible evaluation distinguishes naturally occurring missingness from injected corruption, and synthetic process variation from genuinely observed fleet-level regime variation. Datasets should be stratified by task (failure classification, multiclass diagnosis, anomaly detection, RUL forecasting), because their evaluation units (records, runs, intervals, assets) are not interchangeable. **Table 6** summarizes suitability and leakage guards.

The UCI AI4I 2020 dataset is a synthetic, 10,000-record benchmark with five continuous process measurements, a categorical product-type feature, identifiers and several failure-related targets [39]. It is useful for controlled sensitivity experiments because clean measurements can be perturbed reproducibly, but it is not a substitute for longitudinal field validation. The columns for individual failure modes (TWF tool wear failure, HDF heat dissipation failure, PWF power failure, OSF overstrain failure and RNF random failure) are targets or target-derived indicators; when predicting the aggregate machine-failure label they must not enter the predictor matrix, and identifier fields should not be treated

as predictive signals. AI4I-PMDI addresses irregularities omitted from the original simulated benchmark and is a relevant intermediate case, but it inherits the synthetic process [16].

APS Failure at Scania Trucks provides anonymized operational variables with genuine missing entries and asymmetric misclassification costs [40]. It is valuable for incomplete high-dimensional industrial tables, but anonymization limits physically meaningful perturbations, and its predefined split should not be assumed to be a verified chronological deployment scenario. MetroPT contains real multivariate observations from a metro air-production unit with reported maintenance events [37], [38]. The repository documentation of MetroPT-3 states that there are no missing values and lists 15 signals (seven analog and eight digital) for about 1.5 million records from February to August 2020. It has no GPS signal. It does describe operating-state zeros, for example the DV pressure signal reads zero when the compressor is under load and the motor current is near zero when the compressor is off, so such zeros are states and not missing data and must not be imputed. Researchers should verify the sampling characteristics, event timestamps, and failure definitions against the relevant dataset documentation before constructing prediction labels. Although temporal and event-level evaluation is feasible, the limited number of documented failure events may constrain the statistical precision of the results.

TableShift [13] and TabReD [14] supply non-PdM controls that help separate domain-specific effects from general tabular-learning behavior. A benchmark suite should combine at least one synthetic controlled maintenance dataset, one naturally incomplete industrial table and one longitudinal real-sensor dataset, with results reported per dataset and artificially corrupted data labeled separately from naturally imperfect data.

The leakage analysis for AI4I 2020, where the individual failure-mode columns must not enter the predictor matrix, extends to the other two datasets. For APS Scania the predefined training and test split is kept, variable removal (for example dropping variables above a missing-value threshold), imputation, scaling and any feature selection are fitted on training data only and inside the training folds, and, because the variables are anonymized and carry no timestamps, the pipeline cannot check for target-derived content or chronological order and should say so. For C-MAPSS the partition is by engine, never by window, so that windows of one engine do not appear on both sides; the capped RUL label used in many studies is a training-label convention that must be stated and applied identically when comparing with published results; the cycle index is a strong predictor in run-to-failure data and its use as a feature must be declared; normalization statistics, including per-regime statistics, are fitted on training engines only; and the evaluation point (last window per test engine versus all windows) is reported because it changes the loss.

**Table 6. Dataset suitability for mechanism-aware PdM robustness research.**

| <b>Dataset</b>                        | <b>Evidence characteristics</b>                            | <b>Best use</b>                                             | <b>Interpretive limitation / leakage guard</b>                                  |
|---------------------------------------|------------------------------------------------------------|-------------------------------------------------------------|---------------------------------------------------------------------------------|
| <b>AI4I 2020 [39]</b>                 | Synthetic; 10,000 rows; no missing values by construction  | Controlled masks, noise and feature perturbation            | Exclude fault-mode targets when predicting machine failure; no field validation |
| <b>AI4I-PMDI [16]</b>                 | Derived AI4I benchmark with irregularities                 | Irregular acquisition and missingness baselines             | Inherited synthetic process; check transformation provenance                    |
| <b>APS Scania [25], [27], [40]</b>    | Real heavy-truck records; high dimensional; missing values | Real incomplete-data modeling and cost-aware classification | Anonymized fields; check predefined split, leakage and variable semantics       |
| <b>MetroPT [37], [38]</b>             | Real timestamped compressor signals; few reported failures | Temporal sensor-dropout and event-level studies             | Documented as complete; operating state zero codes; small failure count         |
| <b>NASA C-MAPSS [41]</b>              | Simulated run-to-failure engine trajectories               | RUL under operating regimes and channel corruption          | Windowed summaries need engine-level splits and RUL-specific losses             |
| <b>TableShift / TabReD [13], [14]</b> | General-purpose tabular distribution benchmarks            | External robustness cross-check                             | Not industrial sensor-failure evidence on their own                             |

## 10. A REPRODUCIBLE COMPOUND-STRESSOR BENCHMARK PROTOCOL

This section proposes a seven-step protocol for reproducible and leakage-aware evaluation of compound data imperfections in industrial predictive maintenance. The protocol integrates evaluation objectives, validation design, model baselines, stressor specification, factorial assessment, uncertainty estimation, and maintenance-oriented reporting.

The protocol operationalizes the analytical framework developed in **Section 3** and addresses the evaluation limitations identified in **Section 7**. It is intended to guide future empirical studies; its feasibility and effectiveness have not been experimentally assessed in this review.

1. **Define the evaluation objective.** Specify the prediction task, primary loss measure, target estimands, and intended deployment setting.
2. **Establish an appropriate validation design.** Specify the prediction horizon and independent evaluation units, and construct training, validation, and test partitions before fitting preprocessing or model-selection procedures.
3. **Select comparable model baselines.** Include relevant clean-trained, imputation-based, native missing-value, and perturbation-trained approaches where applicable, using consistent information access and tuning procedures.
4. **Specify data-imperfection mechanisms.** Define each stressor's implementation, severity, temporal persistence, unit of application, and assumptions concerning label preservation.
5. **Evaluate controlled stressor combinations.** Where experimentally feasible, assess the eight conditions of a three-factor factorial design using consistent prediction targets, comparable evaluation units, and appropriately paired perturbation realizations.
6. **Estimate robustness effects and uncertainty.** Report prespecified individual and interaction contrasts, together with uncertainty estimates that account for temporal or equipment-level dependence where relevant.
7. **Assess maintenance-related reliability.** Complement predictive performance with calibration assessment [68], uncertainty evaluation under distribution shift [69], event-detection, cost-sensitivity, and model-ranking analyses, according to the prediction task and operational requirements.

For practical implementation, **Table 7** summarizes the seven-step protocol as a checklist of key evaluation requirements for researchers designing reproducible compound-stressor benchmarks.

**Table 7. Practical checklist for reproducible evaluation of compound data imperfections in industrial predictive maintenance.**

| Step | Evaluation component             | Key requirements                                                                                                                                                                     |
|------|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Define evaluation objectives     | Specify the prediction task, primary loss measure, target estimands, and intended deployment setting.                                                                                |
| 2    | Establish validation design      | Define prediction horizons and independent evaluation units; split data before preprocessing and model selection.                                                                    |
| 3    | Select model baselines           | Compare relevant clean-trained, imputation-based, native missing-value, and perturbation-trained approaches under consistent information access and tuning procedures.               |
| 4    | Specify stressor mechanisms      | Document missingness (M), measurement corruption (N), and distribution shift (S), including severity, temporal persistence, unit of application, and label-preservation assumptions. |
| 5    | Evaluate stressor combinations   | Where feasible, assess all eight conditions of the three-factor design using comparable evaluation units and appropriately paired perturbations.                                     |
| 6    | Estimate effects and uncertainty | Report prespecified individual and interaction contrasts with uncertainty estimates accounting for equipment-level or temporal dependence.                                           |
| 7    | Assess maintenance relevance     | Evaluate calibration, event detection, cost sensitivity, and model-ranking stability using measures appropriate to the prediction task.                                              |

The proposed protocol provides methodological guidance for future experiments. Its applicability depends on the available data, the feasibility of controlled perturbations, and the assumptions required for interpreting distributional changes. No empirical benchmark evaluation is conducted in this review.

### ***10.1 Split integrity and model-selection firewall***

First, define the maintenance event, prediction horizon, and decision unit: whether a sample predicts failure at the current point, within a future horizon, or across a run, and, for RUL, how censoring and the evaluation horizon are handled. Second, remove target-derived features, post-event maintenance records, and measurements unavailable at decision time, and document the exclusions in a data dictionary. Third, perform outer splits by asset, run, chronological period or site before creating windows or fitting preprocessing. Fourth, fit imputers, scalers, encoders, and hyperparameters exclusively within training and validation folds; no transformation may learn from the designated test period unless the protocol explicitly declares and labels a transductive or test-time-adaptation setting.

Random splits may be reported as an auxiliary optimistic reference only when their limitations are declared. Grouped or chronological results are the primary deployment-oriented outcomes for repeated measurements. For rare failures, use class-stratified group splitting where this does not violate chronological structure, and report the number of positive events per independent fold.

### ***10.2 Factorial Stressor Generation***

Controlled evaluation requires separately specified mechanisms for missingness (M), measurement corruption (N), and distribution shift (S). Missingness scenarios may include independent entry masking, condition-dependent missingness, plausible MNAR sensitivity scenarios, complete-channel outages, and contiguous temporal gaps. Corruption scenarios should distinguish stochastic noise, bias, drift, spikes, and stuck-at readings while respecting physical constraints. Distribution shifts should preferentially be evaluated through chronological, equipment-level, site-level, or operating-regime partitions; synthetic shift operators must be explicitly defined and distinguished from naturally occurring concept drift.

For a deliberately controlled  $2^3$  experiment, the eight on/off combinations of M, N and a synthetically controlled S intervention are evaluated on the same independent held-out decision units, with additional physically justified severity levels. Illustrative missing fractions (for example, 5%, 15% and 30%) are stress-test design choices, not observed industrial frequencies. Sensor selection should distinguish random channels from safety-critical ones, and corruption must respect ranges, units, categorical variables, and sampling protocols. Paired perturbation seeds reduce simulation variance but do not create independent samples, so uncertainty must be clustered by asset, run or event. Labels are held fixed only when the stressor affects the observation process and not when a physical shift legitimately changes the failure outcome. For each prediction task and primary loss scale, one primary interaction contrast should be prespecified, while the remaining pairwise and three-way contrasts should be treated as secondary analyses with an appropriate multiplicity correction, such as the Holm procedure or false discovery rate (FDR) control, across the prespecified models, datasets, and severity levels. Each interaction contrast must be defined at specified stressor severity levels; evaluating multiple severity combinations therefore produces a family of dose–response contrasts rather than a single interaction estimate.

If S denotes a naturally different plant or later operating period, there is no same-unit counterfactual with  $S = 0$  for the observed domain.  $M \times N$  contrasts should therefore be reported separately within each domain and compared descriptively with uncertainty, and the between-domain difference should not be called a randomized  $M \times N \times S$  effect. A three-way contrast is appropriate only for a synthetic intervention whose label semantics and joint measurement-generation assumptions are stated and checked.

Clean-model robustness (train on intact source data, test under perturbation), robust-training performance (augment training with selected perturbations) and adaptation performance (consume target data or labels) are distinct

experimental regimes and should be reported separately; otherwise, a method may appear robust only because it had more exposure to the target setting.

## 11. LIMITATIONS OF THIS REVIEW

This review has several limitations that should be considered when interpreting its findings. First, the literature was selected through a targeted critical narrative approach rather than a systematic search and screening protocol. Consequently, the evidence synthesis is not exhaustive, and relevant studies may have been overlooked. The evidence tables provide a structured assessment of selected publications rather than a comprehensive representation of the literature. Therefore, the absence of methods or stressor combinations from these tables should not be interpreted as evidence that they have not been investigated elsewhere.

Second, the reviewed studies differ substantially in prediction tasks, data representations, dataset characteristics, stressor mechanisms, validation procedures, and performance measures. These differences limit direct comparisons across studies and prevent reliable, pooled estimates or universal rankings of model families. Furthermore, evidence derived from synthetic datasets, controlled perturbations, or limited industrial settings may not generalize to real-world deployment conditions. The applicability of reported findings therefore remains dependent on the characteristics of the target maintenance environment.

Third, the proposed mechanism-aware framework and factorial evaluation protocol are conceptual and have not been empirically validated in this review. The feasibility of implementing the proposed factorial protocol depends on dataset characteristics, including temporal structure, equipment identifiers, and the availability of sufficient independent evaluation units. Consequently, the full eight-condition design and cluster-aware inference may not be equally applicable across the datasets discussed in this review. Their suitability should therefore be assessed individually in future empirical studies. In particular, missing-not-at-random mechanisms cannot generally be identified from observed data alone without additional assumptions, while naturally occurring distribution shifts may involve unobserved changes in equipment conditions, measurement processes, or maintenance policies. Differences between classification, failure-event prediction, and remaining useful life estimation also require task-specific evaluation measures and validation designs.

Finally, the review does not provide exhaustive coverage of temporal modeling, uncertainty quantification, or rapidly evolving tabular foundation models. These topics are considered primarily in relation to their implications for robustness evaluation. Future empirical investigations will be needed to assess the practical applicability of the proposed framework under realistic industrial conditions.

## 12. CONCLUSIONS

This critical review examined the methodological challenges of evaluating robust tabular learning for industrial predictive maintenance under sensor missingness, measurement corruption, and distribution shift. It developed a mechanism-aware conceptual framework, formalized loss-interaction contrasts, and critically assessed the applicability and limitations of evidence from industrial maintenance studies and related tabular-learning research. Building on this synthesis, the review proposed a leakage-aware evaluation protocol that integrates realistic stressor configurations, appropriate validation strategies, uncertainty estimation, and maintenance-oriented performance measures. Future research should empirically investigate these mechanisms across diverse industrial datasets and model families, with particular attention to interaction effects, calibration, temporal and equipment-level dependencies, and maintenance decision costs. Such evaluations would help establish more reliable evidence about predictive behavior under realistic deployment conditions. The contributions of this review remain conceptual and methodological and require empirical validation.

## AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

## DATA AVAILABILITY

This critical narrative review presents no newly generated or analyzed experimental data, and no original simulation, trained model, or executable benchmark is reported. The discussion is based on published sources and publicly documented dataset resources cited in the references.

## ACKNOWLEDGEMENTS

During the preparation of this review, the authors used ChatGPT-5 (OpenAI) for text editing. The authors reviewed and revised the output and take full responsibility for the content of the final manuscript.

## FUNDING

This research received no external funding.

## DISCLOSURE STATEMENT

The authors have no conflicts of interest regarding the publication of this manuscript.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

## REFERENCES

- [1] W. Zhang, D. Yang, and H. Wang, "Data-Driven Methods for Predictive Maintenance of Industrial Equipment: A Survey," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019, doi: 10.1109/JSYST.2019.2905565.
- [2] M. Fernandes, J. M. Corchado, and G. Marreiros, "Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review," *Applied Intelligence*, vol. 52, no. 12, pp. 14246–14280, Mar. 2022, doi: 10.1007/s10489-022-03344-3.
- [3] W. Li and T. Li, "Comparison of deep learning models for predictive maintenance in industrial manufacturing systems using sensor data," *Scientific Reports*, vol. 15, Art. no. 23545, Jul. 2025, doi: 10.1038/s41598-025-08515-z.
- [4] W. Wang, Z. Wang, Z. Cai, C. Hu, and S. Si, "Robust uncertainty quantification for online remaining useful life prediction with randomly missing and partially faulty sensor data," *Reliab. Eng. Syst. Saf.*, vol. 262, 2025, doi: 10.1016/j.ress.2025.111177.
- [5] D. Zhuang, P. Gong, Y. Shao, L. Shu, Q. Zhu, X. Li, and D. Zhang, "Physically-constrained Mamba-SDE for remaining useful life prediction under irregular observations," in *Proc. 32nd ACM SIGKDD Conf. Knowl. Discov. Data Min. (KDD '26)*, 2026, doi: 10.1145/3770855.3817877; arXiv:2606.01894.
- [6] C. Tsallis, P. Papageorgas, D. Piromalis, and R. A. Munteanu, "Application-Wise Review of Machine Learning-Based Predictive Maintenance: Trends, Challenges, and Future Directions," *Applied Sciences*, vol. 15, no. 9, Art. no. 4898, Apr. 2025, doi: 10.3390/app15094898.
- [7] S. Somvanshi, S. Das, S. A. Javed, G. Antariksa, and A. Hossain, "A survey on deep tabular learning," arXiv:2410.12034, Oct. 2024.
- [8] Y. Guo, Z. Cheng, J. Zhang, B. Sun, and Y. K. Wang, "A review on adversarial-based deep transfer learning mechanical fault diagnosis," *Journal of Big Data*, vol. 11, Art. no. 151, Oct. 2024, doi: 10.1186/s40537-024-01006-4.

- [9] R. Sivaraman Iyer, S. L. Patil, U. M. Chaskar, and V. K. Pachghare, "Attention-enhanced spatiotemporal deep learning for predictive maintenance in oil and gas assets: towards Maintenance 5.0," *Scientific Reports*, vol. 16, Art. no. 19836, Apr. 2026, doi: 10.1038/s41598-026-47109-1.
- [10] F. Dunkin et al., "A Multigranularity Fuzzy Inference Approach for Out-of-Distribution Detection in Fault Diagnosis," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 56, no. 6, pp. 3612–3622, Jun. 2026, doi: 10.1109/TSMC.2026.3659876.
- [11] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.
- [12] D. McElfresh et al., "When do neural nets outperform boosted trees on tabular data?," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.
- [13] J. Gardner, Z. Popović, and L. Schmidt, "Benchmarking Distribution Shift in Tabular Data with TableShift," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 53385–53432, Dec. 2023, doi: 10.52202/075280-2324.
- [14] I. Rubachev, N. Kartashev, Y. Gorishniy, and A. Babenko, "TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks," in *Proc. International Conference on Learning Representations (ICLR)*, 2025.
- [15] Z. J. Cheng, Z. Y. Jia, Z. Zhou, Y. F. Li, and L. Z. Guo, "TabFSBench: Tabular Benchmark for Feature Shifts in Open Environments," *Proc. Mach. Learn. Res.*, vol. 267, pp. 10025–10089, 2025.
- [16] J. V. Autran, V. Kuhn, J. P. Diguët, M. Dubois, and C. Buche, "AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 1201–1209, Jan. 2024, doi: 10.1016/j.procs.2024.09.546.
- [17] K. Zhang, Q. Yang, C. Li, X. Sun, and J. Chen, "Missing Data Recovery Methods on Multivariate Time Series in IoT: A Comprehensive Survey," *IEEE Communications Surveys and Tutorials*, vol. 28, pp. 149–180, 2026, doi: 10.1109/COMST.2025.3585962.
- [18] L. Tamang, M. R. Bouadjene, R. Dazeley, and S. Aryal, "Handling Out-of-Distribution Data: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 10, pp. 5948–5966, 2025, doi: 10.1109/TKDE.2025.3592614.
- [19] J. Liu, T. Wang, P. Cui, and H. Namkoong, "On the need for a language describing distribution shifts: Illustrations on tabular datasets," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.
- [20] W. AlKalbani, A. AlAzri, and M. N. Ahmad, "Data Quality Assessment Framework for Predictive Maintenance (PdM) in Industrial AI: Review," *Lecture Notes in Computer Science*, vol. 16215 LNCS, pp. 35–49, 2026, doi: 10.1007/978-981-95-3808-9\_3.
- [21] L. Cao, J. Su, and Y. Wang, "A Systematic Framework for Data Quality Assessment in Industrial Process Monitoring," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 7, no. 2, pp. 823–832, Apr. 2026, doi: 10.1109/JESTIE.2026.3667011.
- [22] J. Moufid, R. Koulali, K. Moussaid, and N. Abghour, "Impact of Internal Validation Protocols on Predictive Maintenance Performance in Biomedical Equipment," *Technologies*, vol. 14, no. 2, Art. no. 115, Feb. 2026, doi: 10.3390/technologies14020115.
- [23] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *J. Bus. Res.*, vol. 104, pp. 333–339, Nov. 2019, doi: 10.1016/J.JBUSRES.2019.07.039.
- [24] C. Baethge, S. Goldbeck-Wood, and S. Mertens, "SANRA—a scale for the quality assessment of narrative review articles," *Res. Integr. Peer Rev.*, vol. 4, Art. no. 5, 2019, doi: 10.1186/s41073-019-0064-8.
- [25] A. Beikmohammadi, M. H. Hamian, N. Khoeyniha, T. Lindgren, O. Steinert, and S. Magnússon, "A cost-sensitive transformer model for prognostics under highly imbalanced industrial data," *Cluster Computing*, vol. 29, Art. no. 345, 2026, doi: 10.1007/s10586-026-06080-2.
- [26] A. M. Qahtani, "A leakage-aware and reproducible evaluation framework for predictive maintenance classification," *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 3, 2026, doi: 10.14569/IJACSA.2026.0170383.
- [27] C. Gondek, D. Hafner, and O. R. Sampson, "Prediction of failures in the air pressure system of Scania trucks using a random forest and feature engineering," in *Advances in Intelligent Data Analysis XV (IDA 2016)*, *Lecture Notes in Computer Science*, vol. 9897. Cham, Switzerland: Springer, 2016, pp. 398–402, doi: 10.1007/978-3-319-46349-0\_36.
- [28] Z. Wu, H. Zhang, J. Guo, Y. Ji, and M. Pecht, "Imbalanced bearing fault diagnosis under variant working conditions using cost-sensitive deep domain adaptation network," *Expert Syst. Appl.*, vol. 193, p. 116459, May 2022, doi: 10.1016/J.ESWA.2021.116459.
- [29] D. Tang, "A deep learning framework for remaining useful life prediction of turbfan engines with partial sensor failure," *PLOS ONE*, vol. 21, no. 4, Art. no. e0347312, Apr. 2026, doi: 10.1371/journal.pone.0347312.

- [30] A. Mary A. C., R. Rakshinee, S. Jeyaseelan, V. Sakthivel, and P. Prakash, "Predictive modeling for metro performance using MetroPT3 dataset," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160854.
- [31] Z.-J. Cheng, Z.-Y. Jia, K.-Y. Yu, Z. Zhou, and L.-Z. Guo, "Fully test-time adaptation for feature decrement in tabular data," in *Proc. IJCAI*, 2025, pp. 4940–4948, doi: 10.24963/ijcai.2025/550.
- [32] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, Art. no. 6085, 2018, doi: 10.1038/s41598-018-24271-9.
- [33] P.-A. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete data sets," in *Proc. 36th Int. Conf. Mach. Learn.*, PMLR 97, 2019, pp. 4413–4423.
- [34] J. Josse, J. M. Chen, N. Prost, G. Varoquaux, and E. Scornet, "On the consistency of supervised learning with missing values," *Stat. Papers*, vol. 65, pp. 5447–5479, 2024, doi: 10.1007/s00362-024-01550-4.
- [35] M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux, "NeuMiss networks: Differentiable programming for supervised learning with missing values," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.
- [36] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR 80, 2018, pp. 5689–5698.
- [37] B. Veloso, R. P. Ribeiro, J. Gama, and P. M. Pereira, "The MetroPT dataset for predictive maintenance," *Scientific Data*, vol. 9, Art. no. 764, Dec. 2022, doi: 10.1038/s41597-022-01877-3.
- [38] N. Davari, B. Veloso, R. Ribeiro, and J. Gama, "MetroPT-3 dataset," *UCI Machine Learning Repository*, 2021, doi: 10.24432/C5VW3R.
- [39] S. Matzka, "Explainable artificial intelligence for predictive maintenance applications," in *Proc. 3rd Int. Conf. Artif. Intell. Ind. Appl. (AI4I)*, 2020, pp. 69–74, doi: 10.1109/AI4I49448.2020.00023; dataset: *UCI Machine Learning Repository*, doi: 10.24432/C5HS5C.
- [40] *UCI Machine Learning Repository*, "APS failure at Scania trucks," 2016, doi: 10.24432/C51S51.
- [41] NASA Prognostics Center of Excellence, "Turbofan engine degradation simulation data set (C-MAPSS)," 2008.
- [42] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2006.
- [43] H. Sohn, K. Worden, and C. R. Farrar, "Statistical Damage Classification Under Changing Environmental and Operational Conditions," *J. Intell. Mater. Syst. Struct.*, vol. 13, no. 9, pp. 561–574, 2002, doi: 10.1106/104538902030904.
- [44] T. T. N. Thai, P. Do, B. Lung, and P. Gardoni, "Impact of Sensor Degradation on Measurement Uncertainty in Prognostics and Maintenance Decision-Making," *Int. J. Progn. Health Manag.*, vol. 16, no. 4, Nov. 2025, doi: 10.36001/IJPHM.2025.V16I4.4306.
- [45] Y. Liu et al., "Review on deep learning-assisted soft sensors using imperfect measurements," *Meas. Sci. Technol.*, vol. 37, no. 10, p. 102001, Mar. 2026, doi: 10.1088/1361-6501/AE49B3.
- [46] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.
- [47] R. J. Little, "Missing data assumptions," *Annu. Rev. Stat. Appl.*, vol. 8, pp. 89–107, Mar. 2021, doi: 10.1146/annurev-statistics-040720-031104.
- [48] J. Kim, J. H. Park, S. M. Choi, and G. W. Kim, "MARFI: Missingness-Aware Rectified-Flow Imputation for MNAR Data in Predictive Maintenance," *IEEE Access*, vol. 14, pp. 127480–127505, 2026, doi: 10.1109/ACCESS.2026.3721507.
- [49] D. Imkamp et al., "Challenges and trends in manufacturing measurement technology - The 'industrie 4.0' concept," *Journal of Sensors and Sensor Systems*, vol. 5, no. 2, pp. 325–335, Oct. 2016, doi: 10.5194/JSSS-5-325-2016.
- [50] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games*, Vol. II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.
- [51] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ, USA: Wiley, 2019, doi: 10.1002/9781119482260.
- [52] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.
- [53] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012, doi: 10.1093/bioinformatics/btr597.
- [54] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, doi: 10.1145/2939672.2939785.

- [55] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, Art. no. 15, 2012, doi: 10.1145/2382577.2382579.
- [56] J. Piatkowski, L. Karbowiak, and F. Depta, "A lightweight algorithm for synchronized multimodal data acquisition using temporal sample alignment," *Scientific Reports*, vol. 15, Art. no. 22711, Jul. 2025, doi: 10.1038/s41598-025-07797-7.
- [57] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 11033–11043, 2020.
- [58] D. Bahri, H. Jiang, Y. Tay, and D. Metzler, "SCARF: Self-supervised contrastive learning using random feature corruption," in *Int. Conf. Learn. Represent.*, 2022.
- [59] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," *arXiv:2012.06678*, 2020.
- [60] Z. Zhou, K.-Y. Yu, L.-Z. Guo, and Y.-F. Li, "Fully test-time adaptation for tabular data," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 21, pp. 23027–23035, 2025, doi: 10.1609/aaai.v39i21.34466.
- [61] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025, doi: 10.1038/s41586-024-08328-6.
- [62] L. Grinsztajn et al., "TabPFN-2.5: Advancing the state of the art in tabular foundation models," *arXiv:2511.08667*, Nov. 2025.
- [63] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICL: A tabular foundation model for in-context learning on large data," in *Proc. 42nd Int. Conf. Mach. Learn.*, PMLR 267, 2025, pp. 50817–50847.
- [64] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [65] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
- [66] Y. Gorishniy, I. Rubachev, V. Khurlov, and A. Babenko, "Revisiting deep learning models for tabular data," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
- [67] K. Helli, D. Schnurr, N. Hollmann, S. Müller, and F. Hutter, "Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 98742–98781, Dec. 2024, doi: 10.52202/079017-3134.
- [68] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, PMLR 70, 2017, pp. 1321–1330.
- [69] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.