



Machine Learning-Assisted Fault Diagnosis in Smart Manufacturing Systems

Wojcik A^{1*}, Wojcik B² and Wozniak E³

¹Department of Translational Medicine, Bratislava Institute of Applied Engineering, Czech Republic Scopus Author ID: 57462891032 Email: anna.wojcik@bratislava.edu
ORCID: <https://orcid.org/7656-1041-4446-7126>

²Department of Health Sciences, Bratislava Institute of Applied Engineering, Poland Scopus Author ID: 57582874956
Email: barbara.wojcik@bratislava.edu
ORCID: <https://orcid.org/9615-1746-4037-8876>

³Department of Clinical Research, Bratislava Institute of Applied Engineering, Poland Scopus Author ID: 57886297332
Email: elzbieta.wozniak@bratislava.edu
ORCID: <https://orcid.org/9976-6000-6418-4886>

***Correspondence to:** Anna Wojcik, Department of Translational Medicine, Bratislava Institute of Applied Engineering, Czech Republic Scopus Author ID: 57462891032 Email: anna.wojcik@bratislava.edu

Received: 8 August 2025; **Accepted:** 17 October 2025; **Published:** 8 January 2026

Copyright: © 2026 AW, et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Rotating machinery fault diagnosis has moved, over roughly the past decade, from hand-engineered vibration and current-signal features paired with comparatively simple classifiers toward end-to-end deep learning architectures that learn diagnostic features directly from raw sensor signals, following the same representation-learning shift that transformed image classification. This article synthesizes that literature, tracing the field from convolutional and recurrent architectures for vibration-based diagnosis, through the transfer learning, few-shot learning, and synthetic data augmentation techniques developed to address the chronic labeled-fault-data scarcity industrial settings present, to the integration of fault diagnosis models into the broader smart manufacturing stack of digital twins, industrial IoT prognostics frameworks, and privacy-preserving federated learning across facilities. We give particular attention to a growing strand of work on lightweight and explainable architectures, which pushes back against the assumption that the largest, least constrained model is automatically the right one to deploy. Read together, the evidence supports machine learning-assisted fault diagnosis as an increasingly mature and convergent methodology, while leaving the gap between strong benchmark performance on controlled laboratory datasets and reliable detection of novel fault modes under genuinely heterogeneous, multi-site field conditions as a substantially open research agenda.

Keywords: fault diagnosis; smart manufacturing; deep learning; predictive maintenance; federated learning

1. Introduction

Rotating machinery breaks in predictable ways and unpredictable moments. A bearing spalls, a gear tooth cracks, a rotor bows under thermal load — the failure modes themselves are catalogued in maintenance manuals going back decades, yet the exact machine, the exact hour, and the exact vibration signature that precedes catastrophic failure remain stubbornly hard to anticipate. Smart manufacturing was supposed to change that calculus: instrument every motor, pump, and spindle with accelerometers, current sensors, and acoustic microphones; stream the resulting signals to a data historian; and let a model tell operators what a trained vibration analyst would have taken twenty years to learn to hear. That promise has driven an enormous body of machine learning research aimed specifically at fault diagnosis — the task of mapping a raw or lightly processed sensor signal to a fault type, fault severity, or remaining useful life, ideally before the fault becomes a failure.

The traditional pipeline for this task, still taught in reliability engineering courses and still running on a great many factory floors, is a two-stage affair: extract hand-engineered features from the signal (time-domain statistics like kurtosis and RMS, frequency-domain features from a Fourier or wavelet transform, envelope-spectrum features tuned to bearing characteristic frequencies) and then classify those features with a comparatively simple model — a support vector machine, a shallow neural network, a decision tree [1]. This approach works, and it is the reason condition-based maintenance is not a purely aspirational concept. But it also requires a domain expert to decide, in advance, which features matter for which machine, which is exactly the kind of manual, brittle, non-transferable knowledge engineering that deep learning was built to eliminate elsewhere in computer vision and speech recognition.

The turn toward end-to-end deep learning for fault diagnosis follows almost directly from what happened in image classification roughly a decade and a half ago. Krizhevsky, Sutskever, and Hinton's AlexNet demonstrated that a sufficiently large convolutional network, trained end-to-end on raw pixels, could beat a field of hand-engineered computer vision pipelines by a wide margin on the ImageNet benchmark, and the shock of that result — a machine learning a feature hierarchy

directly from data, outperforming decades of manual feature design — reshaped an entire discipline within a few years [2]. Fault diagnosis researchers noticed. If a convolutional network could learn to recognize a cat without anyone telling it what whiskers are, it seemed plausible that a similar network could learn to recognize a spalled bearing without anyone telling it what a characteristic defect frequency is. That intuition has, over roughly the past decade, developed into a substantial and increasingly mature literature, and it is the subject of this review.

Our aim here is synthesis rather than new experimentation. We trace how the field moved from classical machine learning on hand-crafted features to convolutional and recurrent architectures that consume raw or minimally processed vibration, current, and acoustic signals directly [3,4]; how it has confronted the chronic data-scarcity problem that industrial settings present, through transfer learning, few-shot learning, and synthetic data augmentation; how fault diagnosis research has begun to integrate with the broader smart manufacturing stack of digital twins, industrial IoT, and federated learning; and how a growing subset of the field has started taking explainability and deployment trust seriously rather than treating diagnostic accuracy as the only metric that matters. We close with an honest look at what remains unresolved.

A word on scope and candor before we begin. Much of the literature synthesized here reports strong, sometimes near-perfect classification accuracy on public benchmark datasets — the Case Western Reserve University bearing dataset chief among them — collected in controlled laboratory conditions with clean labels and comparatively limited operating regimes. That is a genuinely different setting from a noisy factory floor with electrical interference, changing load conditions, and fault labels that arrive, if at all, only after a costly failure has already occurred. We flag this gap explicitly rather than let benchmark accuracy stand in silently for field performance, and we return to it in the discussion of open challenges.

2. Foundations: From Signal Processing to Data-Driven Diagnosis

Before deep learning entered the picture, fault diagnosis was largely a signal-processing and pattern-recognition problem, and it is worth being clear about what that older approach still gets right, because much of the

newer literature builds on top of it rather than discarding it outright. Tayyab, Asghar, Pennacchi, and Chatterton's work on optimal feature extraction and selection for rotating machine elements is representative of this tradition: time- and frequency-domain statistical features are computed from vibration signals, a feature-selection procedure trims the resulting high-dimensional feature set down to the subset most informative for a given fault classification task, and a comparatively lightweight classifier is trained on what remains [1]. The strength of this approach is interpretability — a maintenance engineer can look at which features a model relies on and connect them back to physical phenomena — and its weakness is exactly the brittleness that motivated the shift toward deep learning: features tuned for one bearing geometry, one sampling rate, or one operating speed often transfer poorly to another.

Zhao, Yan, Chen, Mao, Wang, and Gao's survey of deep learning applications to machine health monitoring is one of the field's most-cited attempts to map how the discipline absorbed deep learning methods across the full spectrum of health-monitoring tasks — fault diagnosis, but also anomaly detection and remaining-useful-life estimation — and it is useful precisely because it treats fault diagnosis as one instance of a broader machine health monitoring problem rather than an isolated classification task [3]. Lei, Yang, Jiang, Jia, Li, and Nandi's subsequent review and roadmap goes further, organizing the literature by the specific machine learning methodology used (from classical to deep architectures) and by the specific fault diagnosis subtask being addressed, and it remains, in our reading, the single best entry point for a reader who wants the field's full taxonomy in one place rather than assembled from a dozen narrower papers [4]. A more recent survey by Puntambekar, Vyas, Thakkar, and Patel narrows the focus specifically to vibration-based bearing fault diagnosis and is candid about where the field's needs, challenges, and open research directions actually sit as of the mid-2020s, which makes it a useful complement to the broader, slightly older surveys [5]. Hoang and Kang's survey, focused specifically on deep-learning-based bearing fault diagnosis, sits between these two in scope and gives a particularly clear account of how convolutional, recurrent, and autoencoder-based architectures have each been applied to the bearing diagnosis problem specifically [6].

Schwendemann, Amjad, and Sikora's narrower survey of condition monitoring and predictive maintenance for bearings in grinding machines is a useful reminder that this literature is not monolithic across equipment types — a technique validated on a generic bearing test rig does not automatically transfer to the specific dynamics of a grinding spindle, and application-specific surveys like this one exist precisely to track where the general methodology has and has not yet been validated for a given machine class [7]. Hector and Panjanathan's survey of predictive maintenance planning models and machine learning techniques under the Industry 4.0 umbrella situates fault diagnosis within the larger predictive maintenance decision problem, since correctly classifying a fault is only useful insofar as it feeds a maintenance scheduling and planning process that actually acts on the diagnosis in time [8].

It is worth naming the handful of general deep learning building blocks that essentially every fault diagnosis architecture discussed in this review draws on, because none of them were developed for fault diagnosis in the first place. The convolutional neural network, popularized for vision by Krizhevsky et al., supplies the local, translation-invariant feature extraction that turns out to work almost as well on a windowed vibration waveform as it does on an image patch [2]. The long short-term memory network, introduced by Hochreiter and Schmidhuber specifically to solve the vanishing-gradient problem that had made training recurrent networks on long sequences impractical, supplies the ability to model the temporal dependencies in a vibration or current signal that a convolution's fixed receptive field cannot capture on its own [9]. The generative adversarial network, introduced by Goodfellow and colleagues as a two-network game between a generator and a discriminator, supplies a mechanism for synthesizing realistic-looking fault data when real fault examples are scarce — a use case we return to at length below [10]. And the attention mechanism formalized in Vaswani et al.'s Transformer architecture, originally built for machine translation, has increasingly been adapted to weight the most diagnostically relevant portions of a signal or feature map rather than treating every timestep or channel equally [11]. Fault diagnosis research over the past decade has been, in large part, the project of figuring out how to bend each of these general-purpose tools around the specific structure of vibration, current, and acoustic

emission signals.

3. Convolutional and Recurrent Architectures for Vibration-Based Fault Diagnosis

The most direct application of the CNN paradigm to fault diagnosis dispenses with hand-crafted features altogether and feeds a convolutional network the raw or lightly normalized signal. Wu, Jiang, Ding, Feng, and Chen's one-dimensional CNN is a clean example: a 1-D convolutional architecture is trained directly on raw vibration signals from the PHM 2009 gearbox challenge dataset and a planetary gearbox test rig, learning features end-to-end rather than consuming a hand-designed feature vector, and the resulting model outperforms three traditional feature-engineering-based diagnostic baselines on both datasets [12]. Xie and Zhang took a hybrid approach a few years earlier, combining empirical mode decomposition — a data-driven signal decomposition technique that splits a signal into a set of intrinsic mode functions without requiring a pre-specified basis, unlike a Fourier or wavelet transform — with a convolutional network trained on the decomposed components, an architecture that sits at the boundary between the classical signal-processing tradition and the end-to-end deep learning one rather than fully choosing a side [13].

A second, closely related strand of work focuses on making these convolutional architectures small enough to matter for deployment rather than only accurate enough to win a benchmark comparison. Yan, Liu, Ye, Jing, and Dai's lightweight convolutional neural network replaces standard convolutions with spatial separable convolutions to sharply reduce the parameter count, explicitly targeting the storage and computational constraints that Industrial Internet of Things deployment imposes; reported diagnosis accuracy exceeds 99.9% once the architecture reaches six lightweight convolutional blocks, and the authors additionally address the interpretability of the resulting model through t-SNE visualization of the learned feature space rather than treating the network purely as a black box [14]. That combination — a design explicitly aimed at the resource envelope a real IIoT gateway or edge device would have to fit inside, evaluated on the same kind of benchmark accuracy metric the field already uses — is characteristic of where a meaningful fraction of the recent architecture literature has moved, and it is a useful corrective to the assumption that deep learning

fault diagnosis inherently demands cloud-scale compute.

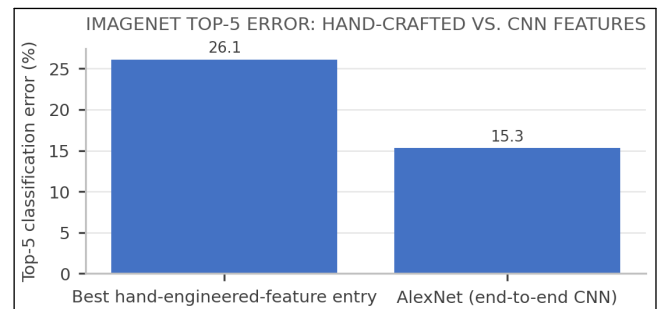


Figure 1: The end-to-end deep feature learning breakthrough that fault diagnosis research inherited from computer vision -- Krizhevsky et al.'s AlexNet reduced top-5 classification error on the ILSVRC-2012 ImageNet benchmark to 15.3%, more than 10 percentage points below the best hand-engineered-feature entry, the same kind of accuracy gap that later motivated replacing hand-crafted vibration features with learned convolutional features in fault diagnosis models [2].

Graph-based architectures represent a further elaboration on the convolutional theme, motivated by the observation that a single sensor channel, or even a single machine, rarely tells the whole diagnostic story on its own — a bearing fault shows up correlated across multiple accelerometer channels, and a fleet of nominally identical machines shares failure statistics that a per-machine model discards. Xiao, Li, Huang, and Yu combine a one-dimensional CNN front end with a modified multi-scale graph convolutional network back end specifically to exploit these relational structures under limited labeled data, treating the diagnosis problem as one where information can propagate across a graph of related signal channels or related machines rather than being processed independently [15]. Li, Cao, Chen, Zhang, Huang, and Qu push the graph-network idea a step further methodologically, using reinforcement learning to automate the architecture search for the graph neural network itself, rather than hand-designing the graph convolution's structure — an acknowledgment that as these architectures accumulate more design choices (which layers, which graph construction, which pooling scheme), manually tuning all of them stops scaling and the search itself becomes worth automating [16]. Deep autoencoder architectures round out this family of representation-learning approaches; Shi, Ding, Wang, Song, Shen, Huang, and Zhu's hypergraph autoencoder embedding extends the standard autoencoder's unsupervised feature-learning objective with a hypergraph structure capable of representing higher-order relationships among samples than a standard pairwise graph can, aimed at the same

fundamental goal as the other architectures in this section — building a feature representation that captures fault-relevant structure in the signal that a hand-designed feature set would miss or that a plain feedforward or convolutional layer would need excessive labeled data to discover on its own [17].

Recurrent and attention-based approaches address a complementary weakness of the purely convolutional architectures: a fixed-size convolutional kernel has a bounded receptive field and no explicit notion of long-range temporal ordering, whereas fault signatures often develop gradually, with early-stage degradation manifesting as subtle drift over many operating cycles well before a discrete fault event. The LSTM architecture, in its original formulation, exists precisely to let a network carry information across long sequences without that information decaying or exploding as the sequence grows [9], and fault diagnosis architectures that incorporate LSTM or attention layers on top of a convolutional feature extractor are, in effect, borrowing this long-range memory capacity to model degradation trends and not just instantaneous fault signatures. The attention mechanism popularized by the Transformer architecture serves an overlapping but distinct purpose here: rather than compressing an entire sequence into a fixed-size hidden state the way a recurrent network does, attention lets the model learn which specific timesteps, frequency bands, or sensor channels are most diagnostically informative for a given input, weighting them accordingly rather than treating the whole signal uniformly [11]. As Figure 1 illustrates using the field's own foundational reference point, the magnitude of the accuracy gain that end-to-end deep feature learning delivered over a hand-engineered-feature baseline in the original ImageNet competition is the same basic pattern — a substantial, not marginal, error reduction from letting the network learn its own features — that motivated fault diagnosis researchers to make the analogous bet on vibration and current signals rather than continuing to refine hand-designed features indefinitely.

4. Learning Under Data Scarcity: Transfer Learning, Few-Shot Learning, and Synthetic Augmentation

Deep architectures are data-hungry, and industrial fault data is chronically scarce in exactly the way that matters most: healthy-operation data is abundant because

machines spend most of their service life healthy, but labeled fault data — signals collected from a machine known to be in a specific fault state, at a specific severity — is expensive and sometimes dangerous to collect deliberately, and rare to collect incidentally, because plant operators understandably shut a machine down and repair it rather than let it run in a faulted state to generate more training examples. This asymmetry between abundant healthy data and scarce, unevenly distributed fault data is arguably the single most consequential practical obstacle standing between the architectures described above and reliable field deployment, and a substantial fraction of the recent literature is organized specifically around solving it.

Generative adversarial networks offer one route around the scarcity problem: synthesize additional fault examples rather than waiting to collect more real ones. Jiang, Wang, Zou, and Zhang's application of deep learning to rotating machinery fault diagnosis integrates a GAN-based augmentation step with a one-dimensional convolutional classifier specifically to address the class-imbalance problem that arises when healthy-state examples vastly outnumber fault-state examples in a real training set, generating synthetic fault signals that expand the effective size and balance of the training distribution before the classifier is trained [18]. Brito, Susto, Brito, and Duarte take a related but distinct approach, combining transfer learning with augmented synthetic data specifically in service of explainable fault diagnosis for rotating machinery, using the synthetic augmentation to address data scarcity while simultaneously building the model in a way that keeps its predictions interpretable rather than treating augmentation and explainability as separate concerns to be bolted on independently [19].

Transfer learning attacks the same underlying problem from a different angle: rather than synthesizing more data for the target machine or fault type, reuse a model, or a set of learned features, from a related source domain where labeled data is more abundant, and adapt it cheaply to the target domain's smaller labeled dataset. This is a particularly natural fit for fault diagnosis because the underlying physics of, say, an outer-race bearing defect is broadly similar across different bearing geometries and different host machines, even though the exact vibration signature differs in ways that make a model trained on one machine perform poorly,

unadapted, on another. Few-shot learning specifically targets the extreme end of this scarcity spectrum, where only a handful of labeled examples — sometimes single digits — are available for a new fault class or a new operating condition, and the model must generalize from that handful rather than from a conventionally sized training set. Li, Gu, and Wei's few-shot bearing fault diagnosis method combines a Kolmogorov-Arnold network-based classifier with diffusion-based data augmentation specifically to handle small-sample conditions, reporting that the combined approach achieves higher diagnostic accuracy under small-sample conditions than comparison methods evaluated on the same Case Western Reserve University bearing benchmark [20]. Liu, Jiang, Yao, and Zeng's counterfactual-augmented few-shot contrastive learning approach tackles the same limited-sample regime from a contrastive-learning angle, using counterfactual augmentation — generating plausible alternative versions of a labeled example that preserve its fault class while varying other signal characteristics — to expand the effective diversity of a small labeled set without requiring an entirely separate generative model, and pairing that augmentation with a contrastive objective that pulls same-class representations together and pushes different-class representations apart in the learned feature space [21].

It is worth being explicit about what these techniques do and do not solve. Transfer learning, few-shot learning, and synthetic augmentation each reduce how much labeled target-domain data a model needs to reach a given accuracy level; none of them eliminate the need for some labeled examples of each fault type the model is expected to recognize, and none of them straightforwardly handle a genuinely novel fault mode the training data never represented at all, synthetically or otherwise. That distinction — between learning efficiently from scarce examples of a known fault category and detecting a fault category nobody has seen before — remains a meaningfully open problem that we return to in the discussion of open challenges.

5. Fault Diagnosis in the Smart Manufacturing Stack: Digital Twins, IIoT, and Federated Learning

A fault diagnosis model, however accurate on a benchmark dataset, is not by itself a smart manufacturing system; it is one component that has to

be embedded in a larger stack of sensing infrastructure, data pipelines, and — increasingly — cross-facility or cross-organization collaboration before it delivers value on a factory floor. Three developments in this surrounding stack are worth examining specifically because of how directly they interact with the diagnostic modeling choices described above: digital twins, industrial IoT-oriented prognostics and health management frameworks, and federated learning.

Table 1: Representative machine learning approaches to fault diagnosis under data scarcity and distributed deployment constraints, with the effect each cited study reports. [14,28,12,26,20,25].

Approach	Source	Constraint Addressed	Reported Effect
Lightweight 1-D CNN with spatial separable convolutions	Yan et al. [14]	Model size for IIoT/edge deployment	Diagnosis accuracy exceeds 99.9% at 6 lightweight convolutional blocks
Kolmogorov-Arnold Network classifier with feature attribution	Rigas et al. [28]	Interpretability vs. accuracy tradeoff	Perfect F1-scores for fault detection; 100% F1 in most severity-classification tasks across two benchmarks
End-to-end 1-D CNN on raw vibration signals	Wu et al. [12]	Hand-crafted feature engineering	Outperforms three traditional feature-based baselines on PHM 2009 gearbox and planetary gearbox data
Federated few-shot learning (FedCDAE-MN)	Liang et al. [26]	Cross-client label scarcity in IIoT	Surpasses existing federated and few-shot baselines while keeping raw data on-client
KAN classifier with diffusion-based augmentation	Li et al. [20]	Small-sample bearing fault data	Higher accuracy than comparison methods under small-sample conditions on CWRU benchmark
Federated contrastive prototype learning	Wang et al. [25]	Cross-facility data privacy	Collaborative diagnostic performance approaching centralized training without raw data pooling

A digital twin, in the sense relevant here, is a live, continuously updated virtual model of a physical asset or process that is kept synchronized with real sensor data and can be used to simulate future states, including degraded or faulted ones, without having to observe them on the physical machine first. Yang, Feng, Jiang, and Mei's digital-twin-enabled health prognostics framework for smart manufacturing systems is explicitly designed to operate under uncertain operating conditions — the realistic case where load, speed, and environmental conditions all vary in ways that a model

trained under a single fixed operating regime handles poorly — by using the digital twin to represent and reason about that operating-condition uncertainty directly rather than requiring the diagnostic model to be robust to it implicitly [22]. Sundaram and Zeid's interoperable Smart Prognostics and Health Management framework addresses a related but more systems-level concern: how prognostics and health management components, which in practice are frequently built by different vendors using different data formats and different modeling assumptions, can be made to interoperate within a single smart manufacturing deployment rather than existing as isolated point solutions that cannot share data or models with each other [23]. Both papers are a useful reminder that a substantial fraction of the practical difficulty in deploying fault diagnosis at the smart-manufacturing scale is systems integration rather than model accuracy per se.

Federated learning enters this picture because fault data, precisely because it is scarce at any single facility, becomes considerably more valuable in aggregate — a machine type deployed at a hundred plants generates a hundred times the fault examples that any single plant sees, if only those examples could be pooled — but pooling raw sensor data across organizational or even departmental boundaries runs immediately into confidentiality and competitive-sensitivity concerns that plant operators are, reasonably, unwilling to waive. Federated learning offers a way to share statistical strength across facilities without moving raw data off any individual site: each site trains a model, or a partial model, locally, and only model updates — not raw sensor signals — are aggregated across sites. Bagheri, Rezapoor, and Lee's unified data security framework for federated prognostics and health management lays out what this looks like specifically for the smart manufacturing setting, treating data security as a first-class design constraint on the federated architecture rather than an afterthought layered on top of a system designed without it in mind [24]. Wang, Huang, Zhang, Wang, Ding, and Shen's federated contrastive prototype learning method operationalizes this idea for fault diagnosis specifically, using a contrastive-learning objective to build per-class prototype representations that can be aggregated federally across clients while preserving each client's raw data locally, aiming for collaborative diagnostic

performance that approaches what a centralized model trained on all clients' pooled data would achieve, without requiring that pooling to actually happen [25]. Liang, Zhao, and Wang combine federated learning with few-shot learning directly, proposing a federated few-shot fault diagnosis method explicitly for the Industrial Internet of Things setting where individual edge clients hold only a handful of labeled fault examples each, using the federation to let clients benefit from each other's limited labeled data through shared model updates rather than each client having to solve the few-shot problem in isolation with no cross-client information at all [26]. Xia, Zheng, Li, Tang, and Zhang's vertical federated learning approach addresses a distinct but related data-partitioning scenario: rather than each client holding the same features for different samples (the horizontal case the other federated methods here assume), vertical federated learning handles the case where different clients hold different features — different sensor types, say — for overlapping sets of machines, using a privacy-preserving gradient boosting tree formulation to let those clients collaboratively train a diagnostic model without any client exposing its own feature set to the others [27]. Table 1 summarizes several of the specific reported effects from this section and the preceding one, giving a concrete sense of the kind of performance and efficiency gains this literature has actually documented.

6. Explainability and Trust in Deployed Diagnostic Models

A model that reports "outer-race bearing fault, 97% confidence" is not, by itself, something a maintenance engineer can act on responsibly — not because the number is necessarily wrong, but because acting on a maintenance recommendation that costs real downtime and real parts requires understanding why the model believes what it believes, particularly when the model's training data and the machine currently being diagnosed differ in ways the deployment team may not have anticipated. This concern has produced a growing, still comparatively small subset of the fault diagnosis literature focused explicitly on explainability rather than treating diagnostic accuracy as the only metric worth optimizing.

Brito, Susto, Brito, and Duarte's explainable AI approach, discussed above in the context of data augmentation, is explicit that the two goals — handling

data scarcity through synthetic augmentation and producing a model whose predictions can be interrogated and understood — are meant to be pursued jointly rather than as a sequence of an accuracy-focused model followed by a post-hoc explanation layer bolted on afterward [19]. Rigas, Papachristou, Sotiropoulos, and Alexandridis take a more architecturally direct route to the same goal, replacing the standard multilayer perceptron classifier at the end of a fault diagnosis pipeline with a Kolmogorov-Arnold network, whose learnable activation functions on the network's edges rather than fixed activation functions on its nodes yield symbolic representations that a human can inspect directly, combined with a feature-attribution step that identifies which input features or signal types drove a given classification decision; validated on two established bearing fault diagnosis benchmarks, the resulting framework achieves perfect F1-scores for fault detection and near-perfect F1-scores across most severity classification tasks, while remaining lightweight enough — the authors emphasize shallow architectures and reduced feature counts — for real-time deployment [28]. That combination of competitive-to-perfect benchmark accuracy with genuine model interpretability, rather than a tradeoff between the two, is a meaningful data point against the common assumption that explainability necessarily comes at an accuracy cost, though it is worth noting this result, like most in this literature, comes from controlled benchmark datasets rather than field deployment.

It is worth situating the explainability question alongside the earlier discussion of lightweight architectures, because the two concerns are not unrelated: Yan et al.'s lightweight CNN pursued t-SNE visualization of its learned feature space as an interpretability check specifically because a smaller, more constrained architecture is easier to inspect than a very large one, even without a KAN-style symbolic representation built in from the start [14]. Taken together, the resource-constraint literature and the explainability literature are converging, in our reading, on a shared conclusion: the largest, least constrained model that achieves the best benchmark number is frequently not the model a plant is actually willing to deploy, and a meaningful fraction of the field's recent energy has gone into architectures that trade a small amount of headline accuracy, or none at all, for interpretability and a deployable resource footprint.

7. Open Challenges

Several gaps recur across the literature synthesized here and deserve to be stated plainly rather than left implicit in the discussion above. First, and most fundamentally, the benchmark-to-field gap: the overwhelming majority of the accuracy figures reported in this literature, including the near-perfect and perfect classification results described in the previous two sections, come from a small number of public, laboratory-collected bearing and gearbox datasets under controlled operating conditions, and the field has been considerably less thorough about validating these same architectures on messy, real-world plant data with genuinely non-stationary operating conditions, electrical and mechanical noise sources the lab datasets do not include, and label uncertainty that a curated benchmark simply does not have [4,6,5]. A model achieving 99.9% accuracy on a clean bearing dataset collected at a fixed rotating speed says less than it might first appear about that same model's reliability on a variable-speed industrial fan whose vibration signature drifts with ambient temperature and load.

Second, novel fault detection remains substantially harder than the few-shot and transfer learning literature's framing sometimes suggests. Few-shot methods reduce how many labeled examples of a known fault category are needed, but a fault mode entirely absent from training — a failure mechanism nobody has yet observed and labeled — is a genuinely different, open-set recognition problem that most of the architectures reviewed here are not designed to solve, and are rarely evaluated against [20,21]. Third, the interaction between federated learning's privacy guarantees and its diagnostic performance under genuinely heterogeneous, non-identically-distributed fault data across client sites — different machine ages, different maintenance histories, different fault-label quality — is still comparatively under-explored relative to how well each half of that combination is understood in isolation; most of the federated fault diagnosis results reviewed here are demonstrated on data partitioned from a single benchmark dataset to simulate multiple clients, rather than on genuinely independent client populations with the kind of heterogeneity a real multi-plant deployment would present [24,25,26,27].

Fourth, explainability techniques, while a genuine and welcome addition to the field, have mostly been

validated in terms of whether they produce something a human can inspect, rather than whether the resulting explanations actually improve a maintenance engineer's decision-making or trust calibration in a real deployment setting — a distinction the explainability literature more broadly has increasingly flagged as important, and one that fault diagnosis-specific explainability work has not yet addressed empirically to the same degree [19,28]. Fifth, and more mundane but no less consequential: reporting standards across this literature are inconsistent enough — some papers report accuracy, some F1-score, some both, rarely with confidence intervals or statistical significance testing against baselines — that comparing reported numbers across papers, as we have done in places in this review, requires caution about exactly what is being compared. We do not think any of these gaps undermine the basic case that machine learning-assisted fault diagnosis is a productive, increasingly mature research direction; they are, read together, an honest accounting of the distance remaining between strong benchmark results and dependable factory-floor deployment.

8. Conclusion

The literature surveyed here traces a reasonably coherent arc: fault diagnosis moved from hand-engineered features paired with comparatively simple classifiers [1] toward end-to-end convolutional and recurrent architectures that learn directly from raw vibration, current, and acoustic signals, borrowing the same basic representation-learning bet that transformed image classification a decade earlier [2,12,13]. That architectural shift, in turn, exposed a chronic industrial data-scarcity problem that a substantial and still-growing literature on transfer learning, few-shot learning, and synthetic data augmentation has worked to address, with meaningful reported gains even under genuinely limited labeled data [20,21,18]. More recently, the field has started reaching outward from the individual diagnostic model toward the broader smart manufacturing systems it has to operate inside — digital twins that model operating-condition uncertainty explicitly [22], interoperable prognostics and health management frameworks [23], and federated learning architectures that let multiple facilities share statistical strength without sharing raw data [24,25,26,27] — while a parallel and increasingly important strand of work has pushed back against the assumption that the largest, most opaque model is automatically the right one to

deploy, favoring lightweight and explainable architectures instead [14,19,28].

What this synthesis does not support is a claim that machine learning-assisted fault diagnosis is a solved, off-the-shelf technology ready for unsupervised deployment across arbitrary industrial equipment. The gap between controlled benchmark performance and field reliability under genuinely novel operating conditions, unseen fault modes, and heterogeneous multi-site data remains real and, in our reading, under-addressed relative to how much attention benchmark accuracy itself continues to receive. What the evidence does support is a more specific and, we think, defensible claim: the field has converged, across many independent research groups working on convolutional architectures, transfer and few-shot learning, federated and privacy-preserving training, and explainable model design, on a broadly consistent set of techniques for extracting reliable diagnostic signal from scarce, distributed, and increasingly privacy-sensitive industrial sensor data. That convergence is the strongest evidence available that machine learning-assisted fault diagnosis is a durable architectural pattern for smart manufacturing rather than a collection of disconnected benchmark results, and the honest gaps catalogued above are best read as a research and deployment roadmap rather than as reasons for skepticism about the pattern itself.

9. References

1. Tayyab, S.M., Asghar, E., Pennacchi, P., Chatterton, S. (2020) Intelligent Fault Diagnosis of Rotating Machine Elements Using Machine Learning Through Optimal Features Extraction and Selection. *Procedia Manufacturing*, 51: 266-273. DOI: [10.1016/j.promfg.2020.10.038](https://doi.org/10.1016/j.promfg.2020.10.038)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 60(6): 84-90. DOI: [10.1145/3065386](https://doi.org/10.1145/3065386)
3. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., Gao, R.X. (2019) Deep Learning and Its Applications to Machine Health Monitoring. *Mechanical Systems and Signal Processing*, 115: 213-237. DOI: [10.1016/j.ymssp.2018.05.050](https://doi.org/10.1016/j.ymssp.2018.05.050)
4. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K. (2020) Applications of Machine Learning to Machine Fault Diagnosis: A Review and Roadmap. *Mechanical Systems and Signal Processing*, 138: 106587. DOI: [10.1016/j.ymssp.2019.106587](https://doi.org/10.1016/j.ymssp.2019.106587)

5. Puntambekar, R., Vyas, P., Thakkar, A., Patel, D. (2026) A Survey of Machine Learning and Deep Learning Methods for Vibration-Based Bearing Fault Diagnosis: The Need, Challenges, and Potential Future Research Directions. *Neurocomputing*, 659: 131628. DOI: [10.1016/j.neucom.2025.131628](https://doi.org/10.1016/j.neucom.2025.131628)
6. Hoang, D.-T., Kang, H.-J. (2019) A Survey on Deep Learning Based Bearing Fault Diagnosis. *Neurocomputing*, 335: 327-335. DOI: [10.1016/j.neucom.2018.06.078](https://doi.org/10.1016/j.neucom.2018.06.078)
7. Schwendemann, S., Amjad, Z., Sikora, A. (2021) A Survey of Machine-Learning Techniques for Condition Monitoring and Predictive Maintenance of Bearings in Grinding Machines. *Computers in Industry*, 125: 103380. DOI: [10.1016/j.compind.2020.103380](https://doi.org/10.1016/j.compind.2020.103380)
8. Hector, I., Panjanathan, R. (2024) Predictive Maintenance in Industry 4.0: A Survey of Planning Models and Machine Learning Techniques. *PeerJ Computer Science*, 10: e2016. DOI: [10.7717/peerj-cs.2016](https://doi.org/10.7717/peerj-cs.2016)
9. Hochreiter, S., Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, 9(8): 1735-1780. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014) Generative Adversarial Networks. *arXiv preprint*. DOI: [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017) Attention Is All You Need. *arXiv preprint*. DOI: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)
12. Wu, C., Jiang, P., Ding, C., Feng, F., Chen, T. (2019) Intelligent Fault Diagnosis of Rotating Machinery Based on One-Dimensional Convolutional Neural Network. *Computers in Industry*, 108: 53-61. DOI: [10.1016/j.compind.2018.12.001](https://doi.org/10.1016/j.compind.2018.12.001)
13. Xie, Y., Zhang, T. (2017) Fault Diagnosis for Rotating Machinery Based on Convolutional Neural Network and Empirical Mode Decomposition. *Shock and Vibration*, 2017: 1-12. DOI: [10.1155/2017/3084197](https://doi.org/10.1155/2017/3084197)
14. Yan, J., Liu, T., Ye, X., Jing, Q., Dai, Y. (2021) Rotating Machinery Fault Diagnosis Based on a Novel Lightweight Convolutional Neural Network. *PLOS ONE*, 16(8): e0256287. DOI: [10.1371/journal.pone.0256287](https://doi.org/10.1371/journal.pone.0256287)
15. Xiao, X., Li, C., Huang, J., Yu, T. (2024) Rotating Machinery Fault Diagnosis Based on One-Dimensional Convolutional Neural Network and Modified Multi-Scale Graph Convolutional Network. *Engineering Applications of Artificial Intelligence*, 137: 109129. DOI: [10.1016/j.engappai.2024.109129](https://doi.org/10.1016/j.engappai.2024.109129)
16. Li, J., Cao, X., Chen, R., Zhang, X., Huang, X., Qu, Y. (2023) Graph Neural Network Architecture Search for Rotating Machinery Fault Diagnosis Based on Reinforcement Learning. *Mechanical Systems and Signal Processing*, 202: 110701. DOI: [10.1016/j.ymssp.2023.110701](https://doi.org/10.1016/j.ymssp.2023.110701)
17. Shi, M., Ding, C., Wang, R., Song, Q., Shen, C., Huang, W., Zhu, Z. (2023) Deep Hypergraph Autoencoder Embedding for Machinery Fault Diagnosis. *Knowledge-Based Systems*, 260: 110172. DOI: [10.1016/j.knosys.2022.110172](https://doi.org/10.1016/j.knosys.2022.110172)
18. Jiang, W., Wang, C., Zou, J., Zhang, S. (2021) Application of Deep Learning in Fault Diagnosis of Rotating Machinery. *Processes*, 9(6): 919. DOI: [10.3390/pr9060919](https://doi.org/10.3390/pr9060919)
19. Brito, L.C., Susto, G.A., Brito, J.N., Duarte, M.A.V. (2023) Fault Diagnosis Using eXplainable AI: A Transfer Learning-Based Approach for Rotating Machinery Exploiting Augmented Synthetic Data. *Expert Systems with Applications*, 232: 120860. DOI: [10.1016/j.eswa.2023.120860](https://doi.org/10.1016/j.eswa.2023.120860)
20. Li, Y., Gu, X., Wei, Y. (2024) Deep Learning-Based Method for Bearing Fault Diagnosis with Few-Shot Learning. *Sensors*, 24(23): 7516. DOI: [10.3390/s24237516](https://doi.org/10.3390/s24237516)
21. Liu, Y., Jiang, H., Yao, R., Zeng, T. (2024) Counterfactual-Augmented Few-Shot Contrastive Learning for Machinery Intelligent Fault Diagnosis with Limited Samples. *Mechanical Systems and Signal Processing*, 216: 111507. DOI: [10.1016/j.ymssp.2024.111507](https://doi.org/10.1016/j.ymssp.2024.111507)
22. Yang, H., Feng, C., Jiang, G., Mei, X. (2024) Digital Twin-Enabled Health Prognostics for Smart Manufacturing Systems Under Uncertain Operating Conditions. *IEEE Transactions on Industrial Informatics*, 20: 14072-14082. DOI: [10.1109/tii.2024.3441633](https://doi.org/10.1109/tii.2024.3441633)
23. Sundaram, S., Zeid, A. (2021) Smart Prognostics and Health Management (SPHM) in Smart Manufacturing: An Interoperable Framework. *Sensors*, 21(18): 5994. DOI: [10.3390/s21185994](https://doi.org/10.3390/s21185994)
24. Bagheri, B., Rezapoor, M., Lee, J. (2020) A Unified Data Security Framework for Federated Prognostics and Health Management in Smart Manufacturing. *Manufacturing Letters*, 24: 136-139. DOI: [10.1016/j.mfglet.2020.04.011](https://doi.org/10.1016/j.mfglet.2020.04.011)
25. Wang, R., Huang, W., Zhang, X., Wang, J., Ding, C., Shen, C. (2023) Federated Contrastive Prototype Learning: An Efficient Collaborative Fault Diagnosis Method With Data Privacy. *Knowledge-Based Systems*, 281: 111093. DOI: [10.1016/j.knosys.2023.111093](https://doi.org/10.1016/j.knosys.2023.111093)
26. Liang, Y., Zhao, P., Wang, Y. (2023) Federated Few-Shot Learning-Based Machinery Fault Diagnosis in the Industrial Internet of Things. *Applied Sciences*, 13(18): 10458. DOI: [10.3390/app131810458](https://doi.org/10.3390/app131810458)
27. Xia, L., Zheng, P., Li, J., Tang, W., Zhang, X. (2022) Privacy-Preserving Gradient Boosting Tree: Vertical

Federated Learning for Collaborative Bearing Fault Diagnosis. IET Collaborative Intelligent Manufacturing, 4(3): 208-219. DOI: [10.1049/cim2.12057](https://doi.org/10.1049/cim2.12057)

28. Rigas, S., Papachristou, M., Sotiropoulos, I., Alexandridis, G. (2025) Explainable Fault Classification and Severity Diagnosis in Rotating Machinery Using Kolmogorov-Arnold Networks. Entropy, 27(4): 403. DOI: [10.3390/e27040403](https://doi.org/10.3390/e27040403)

10. Declarations

Funding

This literature synthesis received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial or personal interests that could have influenced the work reported in this article.

Data Availability Statement

This article is a literature synthesis and did not generate or analyse any new primary data. All studies discussed are cited in the reference list; readers should consult the original publications for their underlying datasets and code availability statements.

Ethical Approval

Not applicable. As a literature synthesis, this article did not involve human participants, animal subjects, or the collection of primary experimental data.

Citation: Wojcik A, Wojcik B, Wozniak E. "Machine Learning-Assisted Fault Diagnosis in Smart Manufacturing Systems." J Eng Sci Future Technol (2026): 001-011 doi: [10.5281/zenodo.22849421](https://doi.org/10.5281/zenodo.22849421)