

AI-Enabled Predictive Maintenance for Improving the Resilience of Mission-Critical Data Center Infrastructure in Saudi Arabia

Waqas Ahmad^{1*} 

¹Saudi Telecom Company (STC)

*Corresponding Author

Waqas Ahmad

Saudi Telecom Company (STC)

Article History

Received: 19.08.2026

Accepted: 13.10.2026

Published: 14.10.2026

Abstract: Mission critical data centres require continuous power, stable thermal conditions, reliable sensing and disciplined maintenance. As Saudi Arabia ramps up its cloud, AI and digital service infrastructure under Vision 2030, the traditional approach of calendar based maintenance is increasingly inadequate for the operation of such tightly coupled, high availability facilities. This paper reviews the application of AI to enhance predictive maintenance and thus the resilience of mission critical data centre infrastructure in Saudi Arabia. A structured integrative review of the literature was undertaken (2020 to 2025) with a focus on data centre power and cooling assets, batteries, sensors, storage, digital twins, anomaly detection, fault diagnosis, remaining useful life estimation, transfer learning, explainability and maintenance prioritisation. The synthesis indicates that the greatest resilience value is achieved when the predictions are tied to asset criticality, redundancy state, intervention lead time, uncertainty and the existence of controlled operational workflows rather than simply as a standalone measure of prediction accuracy. The strongest evidence base is for battery health assessment, cooling system fault diagnosis, sensor validation, interpretable component failure prediction, transfer learning with limited fault data and digital twin enabled monitoring. The key remaining barriers are sparse failure labels, concept drift, site heterogeneity, imbalanced data, uncertain generalisability and poor integration of the analytical output into engineering governance. A proposed Saudi implementation framework would start with telemetry and asset criticality governance, move through a series of well defined use cases and model assurance and finish with a closed loop maintenance learning process. The paper concludes that AI enabled predictive maintenance can deliver a significant contribution to resilient digital infrastructure when it is deployed as an accountable engineering decision support capability with explicit human oversight, validation, cybersecurity and post intervention verification.

Keywords: Saudi Vision 2030; fuel retail stations; energy efficiency; photovoltaic systems; electric vehicle charging; battery energy storage; digital energy management; decarbonization; climate-responsive HVAC; seasonal energy management.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

1. INTRODUCTION

Mission critical data centres are built around the principle of continuity (power available, thermal

conditions within spec, faults contained so they don't cascade into a service outage). The operating environment is increasingly difficult to manage

Citation: Ahmad, W. (2026). AI-Enabled Predictive Maintenance for Improving the Resilience of Mission-Critical Data Center Infrastructure in Saudi Arabia. *Glob Acad J Econ Buss*, 8(5), 1887-1898. <https://doi.org/10.36348/gajeb.2026.v08i05.073>

though. Higher rack densities, greater compute demands, hybrid cooling, more instrumented power systems and tighter sustainability targets all lead to interactions that are very difficult to capture through fixed preventive maintenance schedules. Saudi Arabia adds a further layer of complexity in that the country is actively expanding its digital infrastructure in parallel with its national ambitions around AI, cloud, smart infrastructure and economic diversification (Vision 2030). Resilience in this context is not just about redundancy, it's about the ability to anticipate degradation, pick up on the faint signs of trouble and act to prevent an outage without having to take any live systems offline.

Predictive maintenance provides that capability by moving maintenance decisions away from calendar-based assumptions and towards decisions based on evidence of asset condition and failure risk. Modern approaches integrate sensor streams, alarms, maintenance histories, equipment logs and environmental data with statistical learning, machine learning and deep learning. Reviews of industrial predictive maintenance have shown significant progress in the area of failure prediction but also highlight a number of continuing problems including data scarcity, class imbalance, concept drift, latency, model transparency and transferability across operating conditions [2]. These issues are particularly acute in the data centre environment where the equipment is maintained proactively to prevent failure and so true failure cases are very rare. Tang *et al.*, have demonstrated the practical value of data-driven prediction in the case of valve regulated lead acid batteries in UPS systems where the early identification of units that are likely to fail is of direct benefit to system continuity [3]. The deep digital twin concept is also shown to be a viable approach for deriving health indicators from operational data without necessarily requiring a large amount of run to failure data [4].

The literature is still very much split across power systems, cooling, servers, storage, sensors and general industrial prognostics. A data centre is a coupled infrastructure system in which an erroneous temperature sensor can lead to incorrect cooling control, a cooling fault can increase server thermal stress, a battery fault can impact ride through capability and a poorly prioritised maintenance action can itself be a source of risk. Resilience therefore requires a system level view that links the prediction to maintenance criticality, change control, redundancy state, human validation and post maintenance verification. This review draws on the published evidence from 2020 to 2025 to consider how AI can support that system level transition in Saudi mission critical data centres. The focus is not on algorithmic innovation per se but on the conditions

under which a prediction can be a defensible operational decision.

2. Aim of the Study and Research Objectives

The objective of this review is to critically assess the potential of AI-enabled predictive maintenance to enhance the resilience of mission critical data centre infrastructure in Saudi Arabia with a specific focus on reliability, operational continuity, maintainability, sustainability and accountable engineering decision making.

The study has four aims: to identify the key failure modes, data sources and maintenance risks across the mission critical electrical, mechanical, environmental and IT support assets; to compare the various AI techniques for anomaly detection, fault diagnosis, RUL estimation and maintenance prioritisation; to look at how the predictions can be combined with resilience controls (i.e. redundancy management, change control, service level requirements and human authorisation); and to establish research and implementation priorities that are appropriate for the high availability and very rapidly expanding Saudi data centre environment.

3. REVIEW METHODOLOGY

We used a structured integrative review approach as the research question is quite broad, covering engineering reliability, facility operations, AI, prognostics, cooling, power continuity and maintenance decision support. The evidence window was set at peer reviewed work published between January 2020 and December 2025. The literature search was conducted across the major engineering and multidisciplinary databases (Scopus, Web of Science Core Collection, IEEE Xplore, ScienceDirect). The search concepts were constructed in four broad areas: the infrastructure context ("data centre", "critical infrastructure", "mission critical"), the maintenance function ("predictive maintenance", "condition monitoring", "fault diagnosis", "remaining useful life"), the analytical method ("artificial intelligence", "machine learning", "deep learning", "digital twin", "anomaly detection") and the resilience outcome ("reliability", "availability", "resilience", "downtime", "maintenance prioritisation"). A number of additional component specific searches were also conducted using terms such as UPS, VRLA battery, cooling, HVAC, CRAH, hard disk, thermal anomaly, sensor fault and failure prediction.

Studies were included if they provided peer-reviewed empirical, methodological or review evidence which was relevant to AI for maintenance or health management and if the findings could be usefully interpreted in the context of high availability infrastructure. Preference was given to studies that

used operational or experimentally validated data, had clearly described modelling procedures, had explicit performance measures and had a clear link between the prediction and the asset health. Data centre specific work was given the most contextual weight. Work from manufacturing, automotive and other cyber-physical domains was included where it addressed a problem that was transferable (e.g. rare failure, temporal dependence, cross domain transfer, explainability, maintenance optimisation). Opinion pieces (non peer reviewed), vendor marketing material, papers with no discernible methodological content and papers outside the 2020-2025 period were not part of the evidential core.

Quality appraisal was along six dimensions: relevance to the predictive maintenance application; transparency of data provenance; appropriateness of the model validation; treatment of class imbalance or operating variability; interpretability or engineering plausibility; and evidence that the method would in fact influence a maintenance decision rather than just provide a classification. There is no article count flow as no reproducible export and deduplication log was available for this manuscript (we would be making up screening numbers in that case). The review does however report its search strategy, selection criteria and evidence boundaries. The final evidence base consists of 30 peer-reviewed references with verifiable DOIs.

The process unfolded in three stages. Initially the studies were coded by asset domain, input data, AI task, validation approach and claimed operational benefit. The findings were then grouped into four analytical themes: detection and diagnosis, prognostics, digital twin and transfer learning support and decision integration. The evidence was then interpreted through a resilience lens with resilience defined as the ability to anticipate degradation, absorb or isolate disturbance, maintain critical service, recover safely and learn from the event. This is distinctly different from a narrow accuracy based view of AI. A model which performs extremely well in classification may still be unsuitable for use in a live data centre if it generates unstable alarms, is not able to operate across the seasons, ignores the status of redundancy or provides no lead time at all.

The review also makes a clear distinction between prediction quality and maintenance value. The standard classification metrics (precision, recall, F1, AUC etc) are all important, but for mission critical operations you need to go further and consider things like warning lead time, false alarm burden, the consequence of a missed failure, maintenance window feasibility, the amount of risk avoided, the need for post-intervention verification and the

change in exposure. The RUL estimate is of no use unless you also provide the uncertainty and link that to a threshold for intervention. The anomaly score is only of value if you can interpret it in terms of the consequences for the asset in question. That's the basis of the framework that is developed in the subsequent sections.

4. Failure Domains and Maintenance Risk in Mission-Critical Data Centers

Data centre resilience is a function of a number of inter-related asset families. Electrical continuity is a function of the utility interface, switchgear, transformers, generators, ATS, UPS modules, batteries, distribution boards and protection systems. Mechanical continuity is a function of the chillers, pumps, cooling towers or heat rejection systems, CRAH/CRAC units, valves, fans, heat exchangers and the associated control loops. Environmental assurance is provided by temperature, humidity, differential pressure, water leak, smoke and airflow sensors. IT support resilience is provided by the storage, server telemetry, network elements and the control platforms (building management and data centre infrastructure management systems). A failure in any one of these elements may be tolerated by redundancy, but the failure of more than one element, or the presence of a latent defect which is allowed to degrade, or the performance of maintenance when redundancy is reduced, can turn a manageable defect into a service event.

Empirical work demonstrates why condition awareness is so important. Asgari *et al.*, looked at simultaneous fault detection and diagnosis in a data centre cooling system and found that it's very difficult to tell apart coupled fault conditions using standard monitoring [5]. On the IT side, work on hard drives has shown a similar issue in that you can have a model which is highly predictive yet still be interpretable and allow the operator to see the relationships [6]. More generally the machine learning enabled maintenance literature points out that the predictive value of a model is a function of the data quality, the failure labels, the context and the deployment conditions rather than the algorithm per se [7]. System level reliability modelling has also shown that the dependencies and functional relationships between components have an impact on the value of maintenance, such that isolated health scores are not sufficient for prioritisation [8]. Cloud data centre reliability models go on to show that you need to link the component states to the service level consequences [9].

Power backup is a key area as battery degradation can be totally invisible during normal utility operation. VRLA studies based on data have

demonstrated that voltage, internal resistance, trend characteristics and various engineered metrics can identify a faulty unit well before any of the conventional thresholds are breached [3]. The cooling domain is quite different in that a fault will manifest itself in terms of temperature, pressure, flow and energy signatures as the IT load and

ambient conditions change. Sensor faults add another dimension in that the control system may respond perfectly to a wrong reading. Predictive maintenance therefore has to treat the sensors as maintainable items rather than as perfect data sources. Table 1 provides an overview of the various domains and the associated analytical focus.

Table 1: Failure domains, data sources, and predictive-maintenance priorities in mission-critical data centers

Domain	Representative assets	High-value signals	AI task	Resilience contribution
Power continuity	UPS modules; batteries; switchgear; generators	Voltage; resistance; temperature; harmonics; load; starts	Trend/anomaly detection; fault classification; remaining-life estimation	Preserve ride-through capability; schedule replacement before redundancy is exposed
Cooling and thermal	Chillers; pumps; CRAH/CRAC units; valves; fans	Supply/return temperatures; pressure; flow; valve position; power	Multivariate fault diagnosis; domain-adapted learning; digital-twin analytics	Prevent hotspots and thermal cascades; protect capacity and efficiency
Sensors and controls	Temperature, humidity, pressure, and flow sensors; controllers	Residuals; cross-sensor consistency; calibration drift; command-response behaviour	Sensor fault detection; correction; probabilistic inference	Maintain trustworthy control loops and prevent false maintenance decisions
IT and storage support	Disk drives; server telemetry; supporting electronics	SMART attributes; errors; temperatures; workload and event histories	Cost-sensitive temporal models; interpretable failure prediction	Reduce component-driven service degradation and improve intervention timing
Operations and maintenance	Alarms; work orders; field reports; MOPs; test records	Event sequences; narratives; duration; recurrence; failure codes	Text classification; risk scoring; maintenance prioritization	Convert operational history into earlier planning and closed-loop learning

The operational implication is that predictive maintenance should be driven by failure consequence rather than by data availability. An asset with a lot of telemetry is not necessarily the most important asset and an asset that fails infrequently may in fact be the one that warrants the most engineering attention. A resilience based programme would therefore start with asset criticality, redundancy topology, failure modes and safe intervention constraints. AI would then be assigned to a very specific decision problem: detect deviation, diagnose a likely cause, estimate deterioration, rank maintenance priority or verify recovery. That sequence of steps avoids the pitfall of a technology led deployment where you build the model before you've actually defined the operational question.

5. AI Methods for Detection, Diagnosis and Prognostics

AI enabled predictive maintenance can be thought of as a hierarchy of increasingly decision relevant tasks. Anomaly detection is the task of

determining whether the current state of affairs is different from some expected baseline. Fault diagnosis is the task of trying to identify the likely fault mode or root cause. Prognostics is the task of estimating future degradation or remaining useful life. Decision support is the task of taking the outputs from the above and converting them into maintenance priorities under the constraints of the operation. Each level adds an additional layer of uncertainty and each requires a different kind of evidence.

Temporal models are key as degradation is almost never a single observation. Attention enhanced recurrent models can learn both feature importance and temporal dependence, which is of benefit in the health estimation of systems where multiple sensor channels are evolving [12]. Temporal convolutional networks offer another way of learning long range dependencies and can perform well on industrial predictive maintenance sequences [13]. Deep transfer learning is of course a very relevant

approach in the data centre context where the target site or failure mode may have very little labelled data available, with potentially a wealth of information in related equipment or historical fleets. Li *et al.*, demonstrated that knowledge can be transferred between failure types to reduce the amount of data required [14]. The broader review of the literature confirms that transfer learning is a practical approach to the problem of domain shift and the scarcity of fault labels [19]. These are all very attractive approaches for the multi-site operators in Saudi Arabia who will have facilities that share equipment families but will have very different climates, loading profiles, control strategies and commissioning histories.

Explainability is just as important. Choi *et al.*, have demonstrated an explainable anomaly detection framework which links the detected anomaly to the variables that are contributing to it rather than just giving you a black box warning [15]. In a critical facility this difference in output is going to have a huge impact on operator trust and response time. A cooling alert that tells you that the abnormality is due to an abnormal suction pressure, an abnormal valve behaviour or an abnormality in the temperature relationships is going to be far more useful than a scalar anomaly score. Automated model selection can reduce the engineering effort required, but the comparison of the supervised and one-class automated machine learning approaches also serves to underline why model automation doesn't get rid of the need for domain validation [16]. The one-class approach is particularly relevant in situations where you have a lot of normal data and very few fault examples, albeit the alarm boundaries for such a system will need to be tuned in terms of operational impact.

Data centre battery work provides very strong evidence for practical prognostics. Li *et al.*, proposed a real time VRLA fault diagnosis method based on historical voltage and internal resistance behaviour which was able to identify faults earlier than conventional threshold based logic [17]. Their subsequent random forest work tackled the problem of extreme class imbalance and the lack of observable variables through feature engineering and was able to predict impending battery failure earlier [23]. These are significant studies as they illustrate a general principle: successful predictive maintenance is often less about using the most complex model and more about creating meaningful features, dealing with class imbalance and having the right prediction horizon in relation to a real maintenance action.

Cooling research comes to the same conclusion via quite different methods. Du *et al.*, combined generative adversarial data augmentation with incremental learning in an attempt to deal with the problem of sparse fault data in data centre HVAC systems [21]. They acknowledge the practical reality that healthy operation generates vast quantities of data whereas significant faults are rare. Data augmentation can be used to improve learning, but the synthetic fault data generated should not be regarded as equivalent to actual field data. A second paper by Du *et al.*, looks at domain adaptation and teacher-student networks to improve fault diagnosis across different control modes and operating conditions [22]. This is very pertinent to the Saudi situation where seasonal ambient variation, partial load operation and changing IT demand will all lead to shifts in the data distribution. A model that only works under the conditions it was trained on is likely to be overconfident.

Storage device prediction is another good example of the value of cost-sensitive learning. Ahmed and Green developed an LSTM for hard drive failure prediction in the face of extreme imbalance, using loss functions that properly reflect the cost of the rare events [25]. The same principle applies to electrical and mechanical assets where the cost of missing a critical failure is very different from the cost of a false alarm. The training objective should be to manage operational risk rather than simply maximise some generic accuracy metric. Research into deep learning for selective maintenance also shows how prognostic outputs can be linked directly to maintenance optimisation rather than just stop at the point of prediction [18]. That move from 'will it fail?' to 'what should be maintained, when and under what constraints?' is where the AI actually starts to contribute to resilience.

Digital twins can help facilitate that transition by providing a link between the measured data and an explicit model of asset behaviour. Reviews indicate that digital twin based predictive maintenance can be of benefit in terms of condition monitoring, remaining life assessment, scenario analysis and bidirectional feedback, albeit with issues around model fidelity, computational cost, interoperability and uncertainty [10,11]. For a data centre the digital twin need not be a perfect representation of the physical asset. It just needs to capture the relationships that are relevant to the maintenance decision. The effect of cooling degradation on rack temperatures for instance, or the interaction between equipment availability and power path redundancy.

Table 2: Representative evidence and operational implications for AI-enabled data-center predictive maintenance

Study	Asset / context	Approach	Main contribution	Operational caution
Tang <i>et al.</i> , [3]	UPS VRLA batteries	Data-driven classification and prediction	Demonstrates early identification of battery replacement risk from operational records	Fleet-specific data and battery management practices affect transferability
Asgari <i>et al.</i> , [5]	Data-center cooling	Data-driven simultaneous fault detection and diagnosis	Shows that coupled cooling faults can be diagnosed from multivariate facility data	Performance depends on sensor quality and representative operating conditions
Du <i>et al.</i> , [21]	Data-center HVAC	GAN-assisted data generation and fault diagnosis	Addresses limited and imbalanced fault observations through synthetic augmentation	Synthetic data must preserve physically plausible fault relationships
Du <i>et al.</i> , [22]	Data-center HVAC	Domain-adaptation deep learning	Improves diagnosis across control and operating-condition changes	Transferred representations require site-specific acceptance testing
Zhu and Lin [26]	Data-center cooling	Digital twin for monitoring and energy management	Combines abnormal-state detection with operational optimization	Twin fidelity and calibration must be maintained as the plant changes
Wang <i>et al.</i> , [27]	Cooling sensors	Hybrid multi-label random forest and Bayesian correction	Links sensor fault diagnosis with correction to preserve control reliability	Simulation or single-site performance should not be generalized without field validation
Ardebili <i>et al.</i> , [28]	Data-center thermal risk	Temporal deep-learning hazard prediction	Anticipates thermal hazards from spatiotemporal signals	Thermal topology, density, and airflow patterns vary substantially by site
Frederick and Pai [30]	UPS maintenance records	Structured and unstructured maintenance-data analytics	Shows value of combining CRM fields and technician reports for failure prediction	Terminology, record completeness, and governance determine model usefulness

The most recent data centre evidence in this area is quite strong. Zhu and Lin have developed a digital twin approach for real time monitoring and energy management of a data centre heat pipe cooling system, showing that the twin can be used for both anomaly detection and for optimisation [26]. Wang *et al.*, have combined multi-label random forests with Bayesian inference to detect, diagnose and correct sensor faults in a data centre cooling system, thus closing the loop between fault detection and the restoration of measurement quality [27]. HazardNet has used a temporal model for thermal hazard prediction from data centre monitoring signals and this work highlights the importance of spatiotemporal information rather than just individual thresholds [28]. Taken together these papers point to the next stage of predictive maintenance not being a collection of disparate classifiers but rather a coordinated inference layer that has an understanding of time, operating context, measurement credibility and system consequence.

6. From Predictive Maintenance to Infrastructure Resilience

Predictive maintenance only improves resilience if it changes the way you operate early enough to maintain safe options. Resilience is about anticipation, but it's also about absorption, recovery and learning. A warning that comes in after you've already lost your redundancy is of very limited value. A moderate confidence warning issued with sufficient lead time can be very useful if it enables you to do the inspection during a planned window, avoids the need for any preparation, risk review or controlled transfer of load. In that case lead time and intervention feasibility have to be first class model outputs.

Figure 1 shows the layered architecture for this. The sensing layer collects electrical, thermal, mechanical, environmental and IT telemetry with explicit quality flags. The integration layer brings the BMS, DCIM, UPS, generator, maintenance, alarm and event data all onto a common time base. The analytics layer then performs the anomaly detection,

diagnosis, prognostics and uncertainty estimation. The resilience layer applies asset criticality, redundancy state, service impact, maintenance window constraints and risk tolerance. Only after all of these checks have been performed should an

output be a work order, an inspection request, an escalation or a change proposal. The post-maintenance observations are then fed back to confirm that the intervention has in fact corrected the predicted degradation.

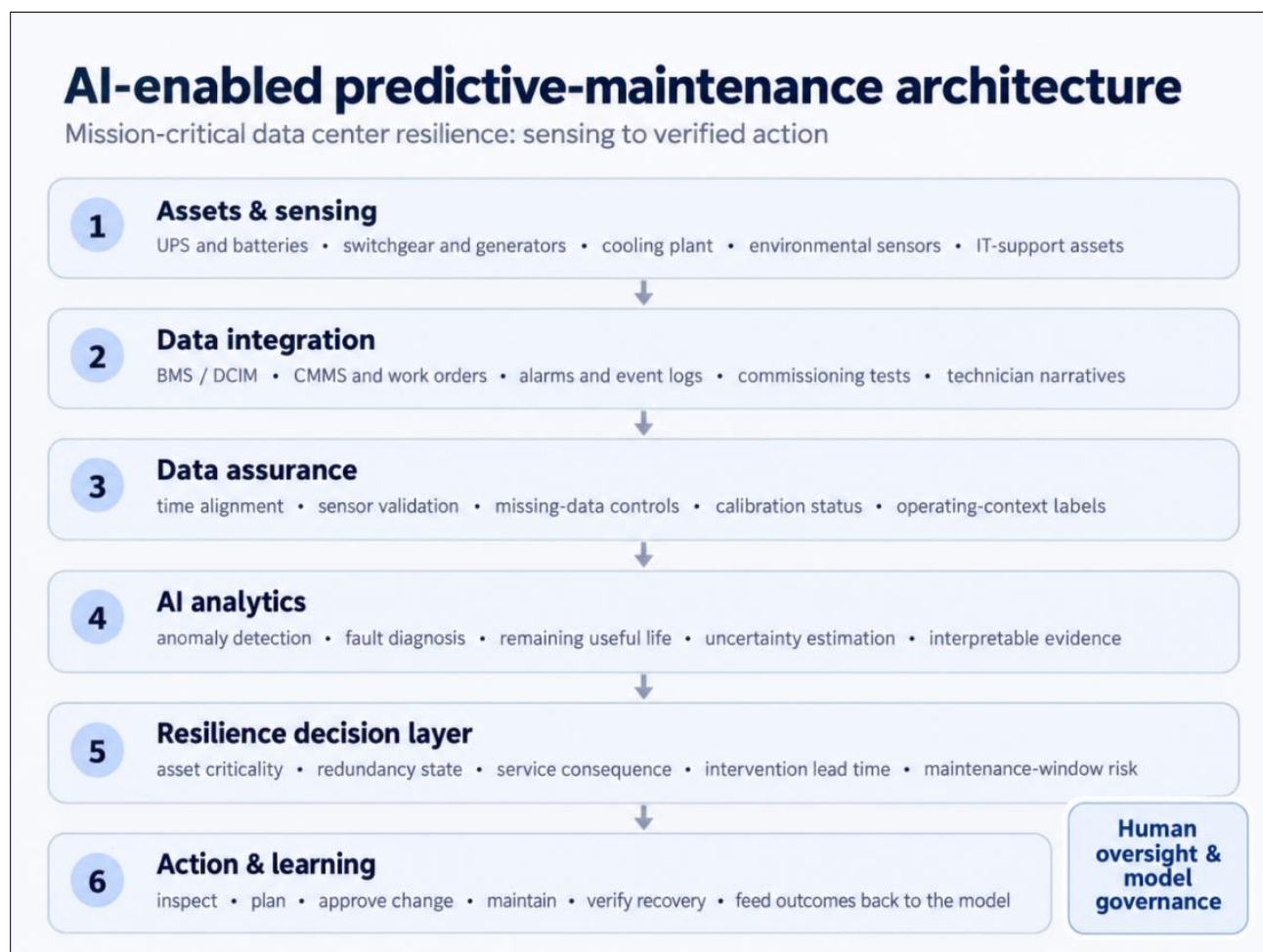


Figure 1: Layered architecture linking data-center sensing, AI analytics, resilience decisions, and verified maintenance action

This is particularly the case for a Saudi mission critical environment where the rate of infrastructure growth can result in very heterogeneous estates. You can have brand new hyperscale or colocation sites sitting alongside older legacy sites which will have very different instrumentation, very different vendor ecosystems and very different levels of maturity in their maintenance. A single universal model is therefore not going to be the answer. You need common data definitions and governance to underpin a single predictive maintenance approach. Model portability is achieved via standardised features, transfer learning, calibration and site specific acceptance testing. Public data reviews confirm that real failure data is still very scarce even in the energy space where you'd expect it to be heavily instrumented [20]. That reinforces the case for federated learning, transfer learning, simulation and carefully governed

synthetic data where you simply don't have enough real failure data.

Sustainability is another resilience dimension. Maintenance which can detect cooling degradation, sensor bias, battery degradation or inefficient control before failure will reduce wasted energy and prevent the system getting into an emergency operating state. Scaife's review of AI in the context of facility predictive maintenance identifies data sufficiency and integration as the key implementation barriers and highlights the potential for improved repair and investment decisions [24]. In the data centre context the sustainability argument should not be divorced from the availability argument. Excessive thermal margins, running equipment that isn't necessary or replacing assets prematurely can all be argued to improve safety (in terms of availability) at the expense of increased

energy or lifecycle cost. The aim is evidence based margin management; to operate efficiently without compromising the level of redundancy and recovery that is required for the critical service.

Maintenance prioritisation is the link between health analytics and the realities of constrained operations. Fadaeefath Abadi *et al.*, developed an availability based prioritisation model for data centres which explicitly takes into account the maintenance options, budget constraints and the resulting availability [29]. It's a piece of work that is not in any way an AI model, but it is important as it demonstrates the decision layer that predictive analytics must ultimately provide for. The outputs of the health model (health probability, remaining useful life, anomaly severity) will feed into a prioritisation function which will also have to take

into account component criticality, current level of redundancy, spare availability, expected duration of the intervention and the constraints around how many items can be maintained at the same time.

Figure 2 shows the closed loop resilience process: detect, diagnose, prognose, prioritise, intervene, verify and learn. There is still a human engineering review throughout. Full automation is not appropriate in situations where a false action could transfer load, disable protection, disturb cooling or reduce redundancy. The appropriate objective is augmented decision making. AI should be used to focus the attention, quantify the risk, surface the evidence and recommend the timing. The engineer is still responsible for any changes that impact the critical infrastructure.



Figure 2: Closed-loop pathway through which predictive maintenance contributes to resilience, with engineering oversight and governance spanning the full cycle

7. Implementation Framework for Saudi Mission-Critical Data Centers

A practical implementation programme should start with data and maintenance governance rather than model procurement. The first step is to build an asset criticality map which links the asset to the power path, the cooling path, the service zone and the redundancy design. The second step is a telemetry quality assessment (sampling frequency, calibration, missingness, timestamp sync, alarm quality, maintenance coding and data ownership). You should not train your predictive model on unverified signals as sensor defects can be learned as the normal behaviour. The sensor correction

evidence from the data centre cooling work is that measurement integrity itself has to be monitored [27].

Step 3 is use case selection. The initial deployments should be focused on failure modes that are consequential, observable and actionable. UPS batteries are a good candidate as degradation is measurable and replacement is something you can do as part of your planned maintenance [3-23]. Cooling equipment and sensors are also good candidates as the condition variables are available all the time and a fault in either will have an impact on both availability and energy performance [5-28]. Hard

drive models can be useful in terms of the techniques they demonstrate for imbalanced failure prediction and interpretability, but facility teams should be wary of simply assuming that a method that works on IT components will work on mechanical or electrical systems without some engineering to validate that assumption [6-25].

Step 4 is model assurance. Each model must have a clearly defined purpose, a defined training population, a defined operating envelope (i.e. what it's expected to do), input quality rules, a method of dealing with uncertainty, an alert threshold, an expected lead time and a retraining trigger. Validation should include temporal holdout and where possible cross season or cross site testing. There is transfer learning evidence to support transfer between related domains, but any transferred model must be recalibrated against the behaviour of the receiving site [14-19]. Model drift should be monitored following any equipment replacement, any change to the control sequence, any firmware update, any capacity increase or any major change to the IT load.

Step 5 is workflow integration. The predictions have to feed into the existing workflows: alarm management, CMMS (computerised maintenance management system), permit to work, method of procedure review, change control, spare parts planning and incident management. Frederick and Pai's work on UPS maintenance data is a good example of the value of combining the structured records with the field service narratives to get failure information that you simply don't get from the sensor data alone [30]. The implication for the Saudi operators is that they should treat the technician notes, vendor reports, commissioning reports, thermography, battery test results and the history of corrective maintenance as analytical assets rather than as discrete documents.

Another Saudi design requirement is for environmental and portfolio variability. The predictive models need to be able to tell the difference between equipment degradation and a legitimate response to high ambient temperature, dust loading, seasonal humidity, load migration and staged capacity expansion. Commissioning data will provide a useful baseline but it should not be treated as a permanent state of normality as the control sequences and thermal profiles will change as white space utilisation changes. Versioned baselines for major operating states should be maintained and engineering changes should be recorded as model context events. Where there are multiple sites using the same equipment, learning can be shared via normalised features or via privacy preserving parameter exchange, with site specific thresholds

remaining under local control. This architecture allows for fleet learning without the assumption that assets that are statistically similar will be in identical operational risk contexts within the Saudi operating environment at this point in time.

Governance should also define accountability. The model owner is accountable for analytical performance; the asset owner is accountable for engineering interpretation; the operations authority is accountable for approving intervention; and the assurance function is accountable for auditing model changes, false alarms, missed events and benefit realisation. Cybersecurity is equally inextricable from predictive maintenance given that the data pipelines will link OT, management platforms, analytics services and potentially cloud resources. Access control, data integrity, model provenance and secure update mechanisms should therefore be part of the maintenance architecture from the outset.

8. DISCUSSION

The evidence is that AI-enabled predictive maintenance can improve data centre resilience, though it's not as simple as 'better prediction = more reliable'. The most compelling work is in solving specific operational problems. Battery models take the sparse failure history and turn it into a signal for earlier replacement [3-23]. Cooling work is about dealing with multiple faults, changing conditions and imbalanced data [5-22]. Sensor analytics work is about restoring the credibility of the data that the control system is relying on [27]. Thermal hazard models are able to pick up on spatio-temporal patterns that a fixed threshold might well miss [28]. In all of these cases the value is in the combination of data engineering, model design, operating context and decision integration.

A persistent problem is the lack of representative failure data. The high availability of the facilities means that failures are deliberately avoided so a good maintenance programme can in fact make supervised learning more difficult. That's not an argument for not using AI at all, that's an argument for preferring methods that are compatible with the data we do have. One class anomaly detection, self supervised learning, transfer learning, digital twins, physics informed features and synthetic augmentation are all techniques that can reduce the reliance on labelled failure data, each of which comes with its own set of assumptions that need to be validated. Reviews of predictive maintenance consistently cite transferability, interpretability and data sufficiency as the key open issues [2-24].

Another issue is metric misalignment. Accuracy can be very misleading when failures are

rare. The model may classify pretty much everything as healthy and still come out as accurate. More useful metrics would be precision, recall, F1 score, calibration and lead time distributions but these are still all quite technical. What the ops manager needs are consequences: avoided downtime, number of false interventions, risk exposure during maintenance, spare consumption, energy impact and the % of alerts that actually led to some form of action. Cost sensitive hard drive prediction [25], and availability based maintenance prioritisation [29], are both examples of more mature objective function designs.

For Saudi Arabia the opportunity is to build up its predictive maintenance capability as the digital infrastructure is still being rolled out. In a greenfield environment you can embed the sensor standards, the asset identifiers, the failure taxonomies, the time sync and the data retention rules at commissioning. In a brownfield environment you can prioritise the critical telemetry gaps and improve your maintenance coding prior to any complex models being deployed. That way you avoid the very expensive pattern of buying in the analytics platform first and only then discovering the data quality issues.

9. Research Gaps and Future Research

Future research needs to shift from component level prediction to multi-system resilience. Firstly there's a need for more work on the interactions between the electrical, cooling, control and IT load states (particularly in the case of degraded redundancy). Secondly the models need to be able to quantify uncertainty and decision lead time rather than just provide a point prediction. Thirdly cross site generalisation is an area that deserves specific investigation in hot climate environments where the ambient temperature, the dust, the cooling architecture and the load profiles are all likely to result in persistent domain shifts. Fourthly predictive maintenance should be assessed in terms of its impact on resilience outcomes (service continuity, safe recovery etc) rather than simply on diagnostic accuracy.

Another gap is around human-AI collaboration. Explainable models are clearly of value [6-15], but future work should look at how operators actually interpret the recommendations under alarm pressure and how confidence, the quality of the evidence and the use of counterfactuals influence the decisions they make. Digital twin work also needs to move from the current stage of visualisation to that of validated what-if analysis. If a component is isolated what redundancy is left, what are the changes to thermal and electrical margins and what is the optimal intervention sequence to reduce risk? Lastly shared anonymised failure datasets and benchmark

definitions for data centre facility equipment would go a long way to improving reproducibility and the ability to compare models.

10. Practical, Managerial and Policy Implications

For the operator the priority is to build a governed predictive maintenance pipeline around the key assets that have measurable condition signals. That has to be done in a staged fashion, starting with decision support pilots and only progressing further once you've proven out the alert quality, the lead time and the maintenance benefit. For the manager the investment decision has to take into account the data engineering, the sensor calibration, the integration, the cyber security and the workforce capability as much as it does the model development. The performance dashboards need to be a combination of model performance and availability, maintenance backlog, energy and risk.

At the sector level common taxonomies for asset classes, failure modes, alarm severity and maintenance outcomes would enable the Saudi data centre operators to learn across their facilities without necessarily exposing any of their sensitive operational data. Workforce development should be focused on the combination of facility engineering and data literacy so that the predictive outputs can be interrogated and validated. That's the way to align AI adoption with resilient infrastructure development rather than just treating the analytics as a separate technology programme.

11. Limitations

The review is constrained by the uneven level of maturity of the evidence. The data centre specific work is very much skewed towards batteries, cooling, thermal monitoring, storage and a few reliability models. Generators, switchgear, transformers and the integrated electrical path are all under-represented in the more recent open literature. The review also pulls together disparate datasets and metrics which precludes any kind of quantitative meta-analysis. There are also very few field studies from the Saudi environment so many of the implementation conclusions are in fact reasoned extrapolations from the international mission critical and industrial evidence base.

12. CONCLUSION

AI enabled predictive maintenance can go a long way to improving the resilience of mission critical data centres if the prediction is embedded in engineering governance. The evidence from 2020 to 2025 demonstrates the practicality of temporal learning, anomaly detection, explainable models, transfer learning, digital twins, fault diagnosis and cost aware prediction across batteries, cooling, sensors, storage and other industrial assets. The key

requirement is to link the prediction to criticality, redundancy, intervention timing, uncertainty and post maintenance verification.

For Saudi Arabia the most defensible route is a staged, asset based programme that starts with reliable telemetry and clear maintenance decisions and then progresses to cross system analytics and digital twins. The aim shouldn't be autonomous maintenance for its own sake. The aim should be earlier awareness, safer intervention, preserved redundancy, reduced avoidable energy consumption and better operational learning. If AI is treated as an accountable decision support layer rather than a black box alarm generator then predictive maintenance can become a genuine capability for resilient digital infrastructure under Vision 2030.

REFERENCES

- Hassan, O. (2020). Artificial Intelligence, Neom and Saudi Arabia's Economic Diversification from Oil and Gas. *The Political Quarterly*, 91(1), 222-227. <https://doi.org/10.1111/1467-923X.12794>
- Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Computers in Industry*, 123, 103298. <https://doi.org/10.1016/j.compind.2020.103298>
- Tang, J. X., Du, J. H., Lin, Y., & Jia, Q. S. (2020). Predictive Maintenance of VRLA Batteries in UPS towards Reliable Data Centers. *IFAC-PapersOnLine*, 53(2), 13607-13612. <https://doi.org/10.1016/j.ifacol.2020.12.854>
- Booyse, W., Wilke, D. N., & Heyns, S. (2020). Deep digital twins for detection, diagnostics and prognostics. *Mechanical Systems and Signal Processing*, 140, 106612. <https://doi.org/10.1016/j.ymssp.2019.106612>
- Asgari, S., Gupta, R., Puri, I. K., & Zheng, R. (2021). A data-driven approach to simultaneous fault detection and diagnosis in data centers. *Applied Soft Computing*, 110, 107638. <https://doi.org/10.1016/j.asoc.2021.107638>
- Amram, M., Dunn, J., Toledano, J. J., & Zhuo, Y. D. (2021). Interpretable predictive maintenance for hard drives. *Machine Learning with Applications*, 5, 100042. <https://doi.org/10.1016/j.mlwa.2021.100042>
- Theissler, A., Perez-Velazquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, 215, 107864. <https://doi.org/10.1016/j.ress.2021.107864>
- Han, X., Wang, Z., Xie, M., He, Y., Li, Y., & Wang, W. (2021). Remaining useful life prediction and predictive maintenance strategies for multi-state manufacturing systems considering functional dependence. *Reliability Engineering & System Safety*, 210, 107560. <https://doi.org/10.1016/j.ress.2021.107560>
- Li, X.-Y., Liu, Y., Lin, Y.-H., Xiao, L.-H., Zio, E., & Kang, R. (2021). A generalized Petri net-based modeling framework for service reliability evaluation and management of cloud data centers. *Reliability Engineering & System Safety*, 207, 107381. <https://doi.org/10.1016/j.ress.2020.107381>
- van Dinter, R., Tekinerdogan, B., & Catal, C. (2022). Predictive maintenance using digital twins: A systematic literature review. *Information and Software Technology*, 151, 107008. <https://doi.org/10.1016/j.infsof.2022.107008>
- Liu, H., Xia, M., Williams, D., Sun, J., & Yan, H. (2022). Digital Twin-Driven Machine Condition Monitoring: A Literature Review. *Journal of Sensors*, 2022, 6129995. <https://doi.org/10.1155/2022/6129995>
- Jiang, Y., Dai, P., Fang, P., Zhong, R. Y., Zhao, X., & Cao, X. (2022). A2-LSTM for predictive maintenance of industrial equipment based on machine learning. *Computers & Industrial Engineering*, 172, 108560. <https://doi.org/10.1016/j.cie.2022.108560>
- Sharma, D. K., Brahmachari, S., Singhal, K., & Gupta, D. (2022). Data driven predictive maintenance applications for industrial systems with temporal convolutional networks. *Computers & Industrial Engineering*, 169, 108213. <https://doi.org/10.1016/j.cie.2022.108213>
- Li, Z., Kristoffersen, E., & Li, J. (2022). Deep transfer learning for failure prediction across failure types. *Computers & Industrial Engineering*, 172, 108521. <https://doi.org/10.1016/j.cie.2022.108521>
- Choi, H., Kim, D., Kim, J., Kim, J., & Kang, P. (2022). Explainable anomaly detection framework for predictive maintenance in manufacturing systems. *Applied Soft Computing*, 125, 109147. <https://doi.org/10.1016/j.asoc.2022.109147>
- Ferreira, L., Pilastri, A., Romano, F., & Cortez, P. (2022). Using supervised and one-class automated machine learning for predictive maintenance. *Applied Soft Computing*, 131, 109820. <https://doi.org/10.1016/j.asoc.2022.109820>
- Li, X., Yang, W., Pang, A., Jiang, C., Zhao, Q., & Haider, S. N. (2022). A fault diagnosis method for VRLA battery in data center. *Energy Reports*, 8, 14220-14235. <https://doi.org/10.1016/j.egyr.2022.10.380>
- Hesabi, H., Nourelfath, M., & Hajji, A. (2022). A deep learning predictive model for selective

- maintenance optimization. Reliability Engineering & System Safety, 219, 108191. <https://doi.org/10.1016/j.res.2021.108191>
19. Azari, M. S., Flammini, F., Santini, S., & Caporuscio, M. (2023). A Systematic Literature Review on Transfer Learning for Predictive Maintenance in Industry 4.0. IEEE Access, 11, 12887-12910. <https://doi.org/10.1109/ACCESS.2023.3239784>
20. Jovicic, E., Primorac, D., Cupic, M., & Jovic, A. (2023). Publicly Available Datasets for Predictive Maintenance in the Energy Sector: A Review. IEEE Access, 11, 73505-73520. <https://doi.org/10.1109/ACCESS.2023.3295113>
21. Du, Z., Chen, K., Chen, S., He, J., Zhu, X., & Jin, X. (2023). Deep learning GAN-based data generation and fault diagnosis in the data center HVAC system. Energy and Buildings, 289, 113072. <https://doi.org/10.1016/j.enbuild.2023.113072>
22. Du, Z., Liang, X., Chen, S., Li, P., Zhu, X., Chen, K., & Jin, X. (2023). Domain adaptation deep learning and its T-S diagnosis networks for the cross-control and cross-condition scenarios in data center HVAC systems. Energy, 280, 128084. <https://doi.org/10.1016/j.energy.2023.128084>
23. Li, X., Pang, A., Yang, W., Zhao, Q., et al. (2023). VRLA battery fault prediction for data center based on random forest model and feature enhancement method. Journal of Energy Storage, 72, 108666. <https://doi.org/10.1016/j.est.2023.108666>
24. Scaife, A. D. (2024). Improve predictive maintenance through the application of artificial intelligence: A systematic review. Results in Engineering, 21, 101645. <https://doi.org/10.1016/j.rineng.2023.101645>
25. Ahmed, J., & Green, R. C. (2024). Cost aware LSTM model for predicting hard disk drive failures based on extremely imbalanced S.M.A.R.T. sensors data. Engineering Applications of Artificial Intelligence, 127, 107339. <https://doi.org/10.1016/j.engappai.2023.107339>
26. Zhu, H., & Lin, B. (2024). Digital twin-driven energy consumption management of integrated heat pipe cooling system for a data center. Applied Energy, 373, 123840. <https://doi.org/10.1016/j.apenergy.2024.123840>
27. Wang, J., Tian, Y., Qi, Z., Zeng, L., Wang, P., & Yoon, S. (2024). Sensor fault diagnosis and correction for data center cooling system using hybrid multi-label random forest and Bayesian inference. Building and Environment, 249, 111124. <https://doi.org/10.1016/j.buildenv.2023.111124>
28. Seyedkazemi Ardebili, M., Acquaviva, A., Benini, L., & Bartolini, A. (2024). HazardNet: A thermal hazard prediction framework for datacenters. Future Generation Computer Systems, 155, 340-353. <https://doi.org/10.1016/j.future.2024.01.031>
29. Fadaeefath Abadi, M., Haghighat, F., & Nasiri, F. (2025). Availability-based maintenance prioritization for data centres: a dynamic programming approach. Safety and Reliability, 44(3), 180-215. <https://doi.org/10.1080/09617353.2024.2441545>
30. Frederick, L., & Pai, M. M. M. (2025). Generation of Real World Maintenance Data of Data Center Uninterruptible Power Supply Systems and Failure Prediction. IEEE Access, 13, 155096-155109. <https://doi.org/10.1109/ACCESS.2025.3605758>