iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

RELIABILITY-CENTERED MAINTENANCE STRATEGIES FOR MINIMIZING DOWNTIME AND MAXIMIZING PERFORMANCE IN HIGH-DENSITY GPU CLUSTER ENVIRONMENTS

Abiodun Victor Oyeleke[1*]
aoyeleke@coreweave.com
CoreWeave Inc. USA
Francis Chukwudi Eze[2]
Department of Civil Engineering, Michigan Technological University, USA
Franciseze2024@gmail.com
William Asiedu[3]
Department of Electrical and Computer Engineering, Iowa State University, USA
wasieduagyekum@gmail.com

ABSTRACT
High-density GPU cluster environments have emerged as the computational backbone of large-scale artificial intelligence, foundation model training, scientific computing, and cloud-native high-performance computing. However, increasing computational density, heterogeneous hardware architectures, intensive thermal loading, and interdependent infrastructure components have significantly elevated the risk of cascading failures, performance degradation, and unplanned service interruptions. Existing maintenance strategies primarily emphasize periodic servicing or isolated predictive diagnostics, offering limited capability to capture system-wide reliability dynamics or optimize maintenance decisions under continuously changing workloads. This study proposes a Digital Twin-Driven Reliability Orchestration Framework (DT-ROF) that continuously synchronizes real-time operational states with virtual representations of GPU nodes, storage systems, high-speed interconnects, cooling infrastructure, and power distribution networks. The framework combines streaming telemetry, reliability state estimation, graph-based dependency modeling, anomaly propagation analysis, and adaptive maintenance orchestration to identify emerging reliability risks before they propagate across the cluster. Rather than scheduling maintenance solely according to component condition, the proposed framework dynamically prioritizes interventions by jointly considering infrastructure dependencies, workload criticality, thermal stress evolution, redundancy availability, and operational risk. A reliability optimization engine further coordinates maintenance activities with workload migration and resource allocation policies to maximize computational availability while minimizing maintenance-induced performance disruption and operational cost. The proposed architecture establishes an intelligent, self-adaptive maintenance paradigm capable of enhancing cluster resilience, extending infrastructure lifespan, improving resource utilization, and sustaining reliable GPU performance under large-scale AI workloads, thereby advancing reliability engineering for next-generation heterogeneous computing infrastructures.

Keywords:
Digital Twin; GPU Cluster Reliability; Maintenance Orchestration; Graph-Based Dependency Modeling; Infrastructure Resilience; High-Performance Computing

1. INTRODUCTION AND RESEARCH MOTIVATION

1.1 Evolution of High-Density GPU Cluster Environments

Artificial intelligence (AI) has evolved from supporting relatively small computational tasks to enabling foundation models, scientific simulations, autonomous systems, and large-scale data analytics that require substantial computational resources and continuous processing availability [1]. This evolution has accelerated the deployment of high-density GPU cluster environments capable of executing massively parallel computations while supporting increasingly sophisticated AI applications across research institutions, hyperscale cloud providers, enterprise data centers, and national supercomputing facilities [2]. Consequently, GPU clusters have become fundamental to modern AI ecosystems because they provide the scalability, computational throughput, and parallel processing capabilities required for training and deploying advanced deep learning models [3].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

A modern GPU cluster integrates GPU accelerators, multicore CPUs, high-bandwidth memory, NVMe storage, high-speed communication networks, intelligent workload schedulers, and advanced cooling systems into a tightly coupled cyber-physical computing environment [4]. These heterogeneous resources collectively maximize computational efficiency while maintaining low communication latency and balanced workload distribution across interconnected processing nodes [5]. Unlike traditional high-performance computing systems that primarily execute static scientific applications, AI clusters operate under dynamic and continuously changing workloads, resulting in fluctuating resource utilization, power consumption, and thermal conditions that increase operational complexity [6].

The increasing computational density of contemporary GPU infrastructures has significantly transformed data-center design and operation. Modern AI accelerators consume substantially higher electrical power, enabling greater computational capability within limited rack space while simultaneously increasing thermal loads and infrastructure complexity [7]. Consequently, power delivery, cooling systems, storage platforms, and communication networks have become highly interdependent, requiring coordinated management to sustain reliable operation [1]. As AI workloads continue to expand in scale and complexity, maintaining infrastructure reliability has become essential for minimizing service interruptions, maximizing computational productivity, and supporting sustainable AI operations through intelligent maintenance and lifecycle management strategies [8].

1.2 Maintenance Challenges in GPU Clusters

Maintaining high-density GPU cluster environments presents significant engineering challenges because infrastructure components operate continuously under intensive computational, thermal, and electrical loads that accelerate hardware degradation and increase operational risk [4]. Unlike conventional enterprise servers that often experience moderate utilization, AI clusters frequently sustain near-maximum workloads for prolonged periods, exposing critical hardware to persistent stress and increasing the likelihood of unexpected failures [7]. Consequently, maintenance strategies must identify degradation patterns early to prevent service interruptions and preserve computational availability [1].

Hardware degradation remains a primary challenge affecting long-term infrastructure reliability. Continuous exposure to elevated temperatures, electrical stress, and sustained computational demand progressively deteriorates GPUs, CPUs, memory modules, storage devices, voltage regulators, and communication interfaces through mechanisms such as electromigration, dielectric breakdown, solder fatigue, and capacitor ageing [5]. These degradation processes initially appear as intermittent performance anomalies before developing into complete component failures, making continuous health monitoring essential for proactive maintenance [2].

Thermal cycling further accelerates infrastructure deterioration as repeated heating and cooling cause mechanical expansion and contraction within semiconductor packages, printed circuit boards, and solder joints [3]. Dynamic AI workloads generate fluctuating thermal conditions that increase material fatigue, reduce hardware lifespan, and elevate cooling requirements, particularly in densely populated GPU racks where heat dissipation is constrained [6]. Simultaneously, failures within liquid-cooling systems including pumps, heat exchangers, valves, coolant circulation networks, and sensors can reduce cooling efficiency, create localized hotspots, and trigger cascading failures across interconnected computing resources [8].

Additional maintenance challenges arise from memory degradation, SSD wear, communication network failures, and power delivery instability. Increasing ECC memory errors, storage device ageing, network congestion, and voltage fluctuations collectively reduce system reliability, computational throughput, and service availability [5]. These interacting failure mechanisms contribute to unplanned downtime, higher maintenance costs, and declining operational performance [2]. Therefore, maintaining reliable GPU cluster environments requires intelligent maintenance approaches capable of continuously assessing infrastructure health, predicting degradation trajectories, and prioritizing maintenance according to asset criticality and operational risk rather than responding only after failures occur [1].

1.3 Research Gap, Objectives, and Contributions

Despite significant advances in AI infrastructure management, maintenance strategies for high-density GPU clusters remain largely dependent on reactive maintenance and fixed preventive maintenance schedules that are insufficient for increasingly complex computing environments [4]. Reactive maintenance delays intervention until equipment failure occurs, often leading to unexpected downtime, emergency repairs, reduced computational availability, and increased operational costs [7]. Preventive maintenance, while reducing some unexpected failures, performs servicing at predefined intervals without considering the actual health of infrastructure components, frequently resulting in unnecessary maintenance activities and inefficient resource utilization [2]. Consequently, neither approach effectively addresses the diverse degradation patterns exhibited by GPUs, storage

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

systems, cooling infrastructure, power delivery equipment, and communication networks operating under continuously changing AI workloads [5].

Although predictive maintenance has improved failure forecasting through machine learning and operational telemetry, existing solutions typically concentrate on individual component prediction rather than infrastructure-wide maintenance optimization [6]. Most approaches neglect asset criticality, system interdependencies, Remaining Useful Life estimation, and operational risk when generating maintenance recommendations, limiting their ability to maximize cluster availability and computational performance [1]. Furthermore, there remains limited research integrating reliability engineering principles, Failure Mode and Criticality Analysis, predictive analytics, and maintenance optimization into a unified framework specifically developed for high-density GPU cluster environments [8].

To address these limitations, this study proposes a Reliability-Centered Maintenance Framework (RCMF) that integrates continuous infrastructure monitoring, reliability analytics, Failure Mode and Criticality Analysis, Remaining Useful Life prediction, and intelligent maintenance optimization within a unified decision-support architecture. The framework continuously evaluates infrastructure health, predicts degradation trajectories, and dynamically prioritizes maintenance according to component criticality, operational risk, and predicted reliability rather than responding after failures occur [3].

The primary objective of this research is to develop an intelligent maintenance framework capable of minimizing unplanned downtime while maximizing computational availability, infrastructure reliability, and overall cluster performance [5]. Specifically, the proposed framework seeks to optimize maintenance scheduling, improve maintenance resource allocation, extend hardware service life, reduce operational expenditure, and enhance infrastructure resilience through data-driven reliability assessment [7]. Additionally, the study establishes a comprehensive benchmarking methodology that compares the proposed framework against conventional reactive, preventive, condition-based, and predictive maintenance strategies using standardized reliability, availability, maintainability, and operational performance metrics [2].

The principal scientific contributions of this work are threefold. First, it develops a comprehensive Reliability-Centered Maintenance architecture tailored to the operational characteristics of high-density GPU cluster environments [6]. Second, it introduces an integrated reliability analytics framework that combines operational telemetry with predictive maintenance intelligence for continuous infrastructure health assessment and proactive decision-making [4]. Third, it formulates a multi-objective maintenance optimization strategy that jointly minimizes downtime, maintenance costs, and failure probability while maximizing infrastructure availability, computational throughput, and long-term operational sustainability [8].

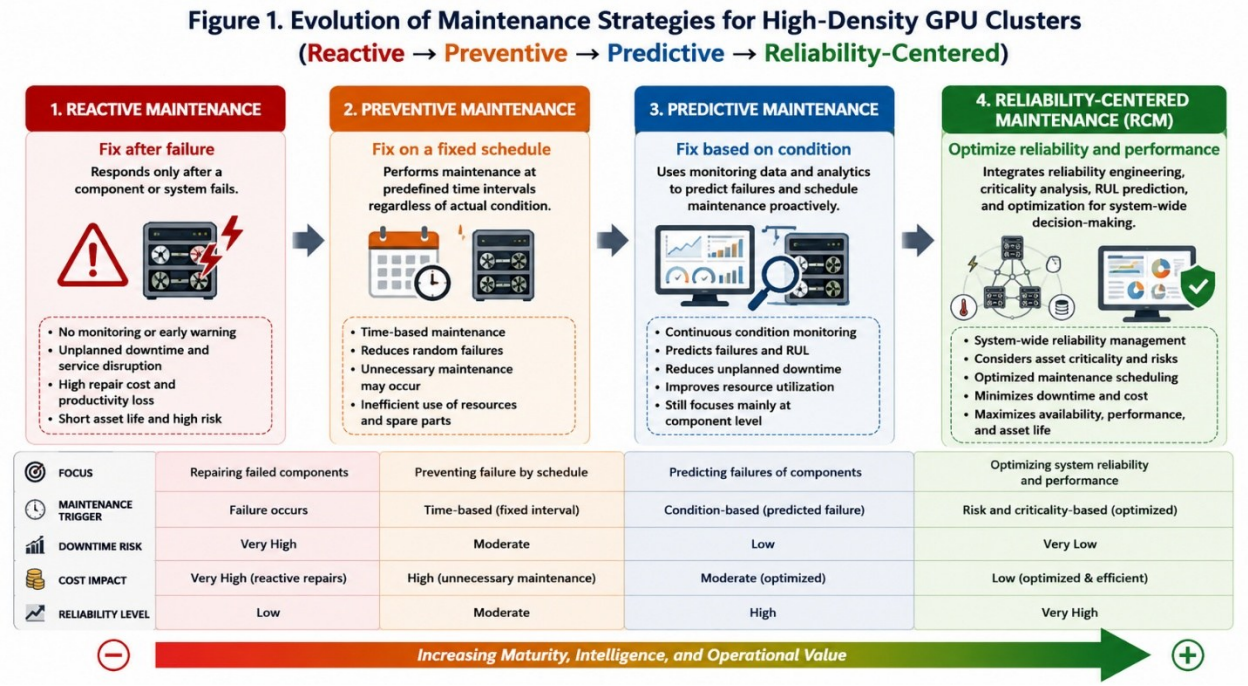

Figure 1. Evolution of Maintenance Strategies for High-Density GPU Clusters

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

2. RELIABILITY-CENTERED MAINTENANCE FRAMEWORK ARCHITECTURE

2.1 Reliability Engineering Principles for GPU Infrastructure

Reliability engineering provides the foundation for designing maintenance strategies that enable high-density GPU clusters to operate efficiently throughout their service life while minimizing unexpected failures and operational interruptions [7]. Unlike conventional maintenance approaches that primarily respond to equipment failures, reliability engineering emphasizes failure prevention, lifecycle optimization, and risk-informed maintenance by continuously assessing the condition of interconnected infrastructure assets [10]. In AI computing environments, reliability extends beyond individual hardware components to include the integrated performance of computing, storage, networking, cooling, and power delivery subsystems whose interactions directly influence system availability and computational throughput [12].

Maintainability is another key principle because it determines how quickly failed components can be restored to operational status with minimal disruption to AI workloads [15]. Effective maintainability depends on modular system design, intelligent diagnostics, automated fault localization, standardized maintenance procedures, and efficient spare-part management, all of which reduce repair time and improve operational continuity [9]. Availability, defined as the proportion of time that computing resources remain operational, is equally critical for supporting continuous AI model training and inference while satisfying service-level agreements and maximizing resource utilization [8]. Consequently, maintenance strategies must balance preventive interventions with uninterrupted computational performance under changing workload conditions [11].

Reliability-centered maintenance also prioritizes infrastructure assets according to their operational criticality rather than applying identical maintenance policies to all components [14]. Asset criticality assessment evaluates GPUs, CPUs, storage devices, cooling equipment, network infrastructure, and power systems based on failure consequences, redundancy, and operational importance, enabling maintenance resources to be directed toward the most critical assets [10]. Furthermore, understanding degradation mechanisms including electromigration, thermal fatigue, dielectric breakdown, mechanical wear, coolant deterioration, and electrical overstress supports early fault detection and proactive maintenance before catastrophic failures occur, thereby improving infrastructure reliability and extending hardware service life [13].

2.2 Multi-Layer Reliability-Centered Maintenance Framework

The proposed Reliability-Centered Maintenance Framework (RCMF) adopts a layered architecture that continuously transforms operational telemetry into intelligent maintenance decisions through sequential reliability assessment, predictive analytics, maintenance optimization, and adaptive feedback mechanisms [11]. Rather than responding only after component failures occur, the framework continuously evaluates infrastructure health across multiple operational layers to minimize downtime while maximizing computational performance and infrastructure availability [8].

Layer 1 – Infrastructure Monitoring continuously supervises the operational condition of GPUs, CPUs, memory modules, storage systems, power delivery units, cooling equipment, communication networks, and environmental sensors [13]. Telemetry streams include temperature, power consumption, utilization, voltage, fan speed, coolant flow, pressure, humidity, network latency, storage health, and system event logs. Continuous monitoring establishes the real-time operational visibility required for intelligent maintenance decision-making [10].

Layer 2 – Data Acquisition aggregates heterogeneous telemetry from multiple monitoring platforms into a unified data repository [15]. Data synchronization aligns timestamps across distributed infrastructure components while filtering noise, eliminating duplicate records, correcting inconsistencies, and ensuring data integrity before reliability analysis is performed [9]. Historical maintenance records, fault logs, hardware replacement histories, and workload scheduling information are also incorporated to provide contextual information for predictive maintenance models [12].

Layer 3 – Reliability Assessment evaluates the current health condition of infrastructure assets by calculating reliability indicators such as failure rates, degradation indices, utilization patterns, thermal stress levels, and component ageing characteristics [7]. These indicators establish baseline reliability profiles that enable maintenance planners to distinguish between healthy equipment and assets exhibiting abnormal operational behaviour [14].

Layer 4 – Failure Prediction employs predictive analytics to estimate future degradation trajectories and Remaining Useful Life (RUL) for critical infrastructure components [11]. Predictive models analyse temporal operational patterns, historical failure behaviour, and reliability indicators to identify components with increasing failure probability before operational disruption occurs [8].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Layer 5 – Maintenance Optimization prioritizes maintenance activities according to asset criticality, operational risk, resource availability, workload scheduling, spare-part inventory, and predicted failure consequences [13]. Multi-objective optimization balances maintenance cost, computational availability, repair urgency, and infrastructure reliability to produce optimal maintenance schedules [10].

Layer 6 – Maintenance Execution coordinates maintenance operations by generating work orders, scheduling technician assignments, allocating replacement components, migrating workloads when necessary, and documenting completed maintenance activities [15]. Intelligent scheduling minimizes service interruptions while ensuring maintenance actions are performed during appropriate operational windows [9].

Layer 7 – Continuous Reliability Feedback closes the maintenance loop by incorporating post-maintenance performance data into reliability models [12]. Updated reliability information continuously improves failure prediction accuracy, maintenance prioritization, and decision-support capabilities, allowing the framework to adapt dynamically as infrastructure operating conditions evolve [14].

2.3 Failure Mode and Criticality Analysis (FMECA) Framework

Failure Mode and Criticality Analysis (FMECA) provides a systematic approach for identifying, evaluating, and prioritizing infrastructure failures according to their operational consequences within high-density GPU cluster environments [7]. Unlike conventional fault diagnosis methods that respond after service disruption, FMECA proactively evaluates potential failure modes, estimates their occurrence probability, assesses failure severity, and determines maintenance priorities before faults propagate across interconnected computing resources [10]. This enables maintenance decisions to be driven by infrastructure risk rather than solely by component condition [12].

GPU accelerators represent the most critical assets because they execute computationally intensive AI workloads under sustained thermal and electrical stress [13]. Common failure mechanisms include overheating, memory faults, voltage instability, thermal interface degradation, electromigration, and fan malfunction, all of which reduce computational throughput and interrupt distributed AI training [8]. CPU failures similarly affect scheduling, virtualization, and memory management, while prolonged processing accelerates thermal fatigue and electrical degradation [15]. Storage reliability is influenced by SSD wear, controller faults, firmware corruption, and increasing input/output latency caused by continuous dataset access and model checkpoint operations [9].

Supporting infrastructure is equally important for maintaining cluster reliability. Power distribution units, voltage regulators, and backup power systems are vulnerable to electrical disturbances and ageing, potentially causing widespread service interruptions [11]. Cooling infrastructure including pumps, heat exchangers, valves, coolant circulation networks, and thermal sensors may experience mechanical wear or reduced coolant flow, creating thermal hotspots that accelerate semiconductor degradation [14]. Likewise, failures within communication networks caused by switch degradation, fibre faults, or bandwidth congestion reduce synchronization efficiency and overall cluster performance [10].

Maintenance priorities are determined using the Risk Priority Number (RPN), which combines failure severity, occurrence probability, and detection capability into a quantitative risk indicator [7]. Components with higher RPN values receive greater maintenance priority because of their increased likelihood of affecting infrastructure reliability and computational availability [12]. Infrastructure reliability is represented using the exponential reliability function

$$R(t) = e^{-\lambda t} \tag{1}$$

where $R(t)$denotes the probability of failure-free operation over time t, and λrepresents the constant failure rate [13]. Reliability performance is further evaluated using the Mean Time Between Failures (MTBF),

$$MTBF = \frac{\text{Operating Time}}{\text{Number of Failures}} \tag{2}$$

where larger MTBF values indicate longer operational periods between failures and improved maintenance effectiveness [15]. These reliability measures provide the quantitative basis for prioritizing maintenance interventions and improving operational resilience within high-density GPU cluster environments [11].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

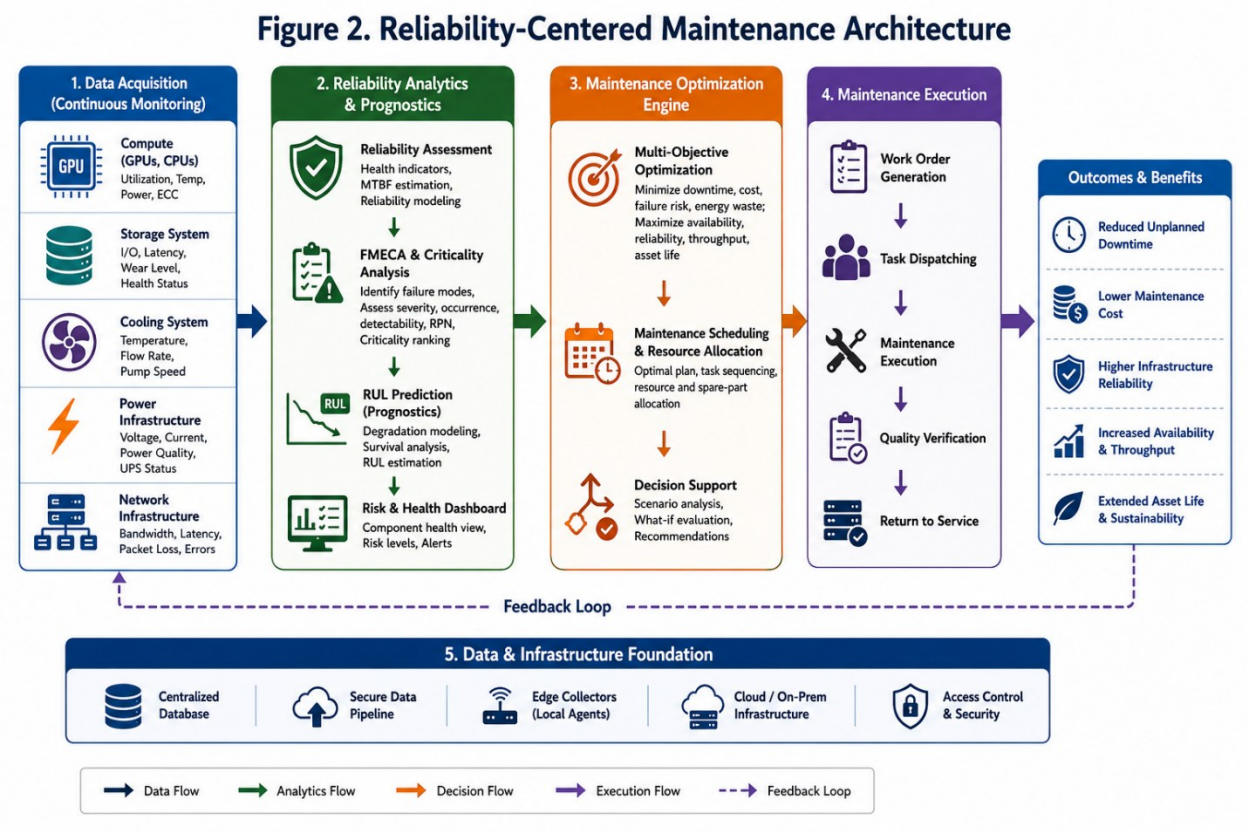

Figure 2. Reliability-Centered Maintenance Architecture

3. RELIABILITY ANALYTICS AND PREDICTIVE MAINTENANCE MODELING

3.1 Operational Data Acquisition

The effectiveness of a Reliability-Centered Maintenance Framework (RCMF) depends fundamentally on the availability of comprehensive, high-quality operational data capable of representing the health status of heterogeneous infrastructure components throughout the GPU cluster lifecycle [14]. High-density GPU clusters continuously generate large volumes of telemetry that reflect computational behaviour, environmental conditions, hardware utilization, and infrastructure performance. Integrating these heterogeneous data streams provides the foundation for predictive reliability analytics because equipment degradation rarely manifests through a single operational indicator but instead emerges from correlated behavioural changes across multiple subsystems [17].

GPU telemetry represents the primary source of reliability information because GPU accelerators perform the majority of computational tasks within AI infrastructures [20]. Operational parameters include core temperature, memory temperature, GPU utilization, memory utilization, clock frequency, power consumption, fan speed, voltage stability, thermal throttling events, and hardware error counters. Continuous monitoring of these variables enables early identification of abnormal operating conditions that precede hardware degradation or computational instability [15].

CPU telemetry complements GPU monitoring by providing processor temperature, utilization, clock speed, cache behaviour, voltage characteristics, and processor error logs that influence workload scheduling and overall system stability [18]. Simultaneously, Self-Monitoring, Analysis, and Reporting Technology (SMART) storage logs provide valuable indicators regarding storage health, including read/write error counts, remaining device lifespan, reallocated sectors, input/output latency, bad block accumulation, and controller status, all of which contribute to long-term infrastructure reliability [21].

Memory reliability is evaluated through Error Correcting Code (ECC) logs that record correctable and uncorrectable memory errors occurring during AI computations [16]. Increasing ECC error frequency often indicates progressive memory degradation, making these records particularly valuable for forecasting hardware failure before catastrophic system interruption occurs [22].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Operational visibility is further enhanced through cooling system sensors that continuously monitor coolant inlet and outlet temperatures, coolant flow rate, pump rotational speed, valve position, pressure differentials, radiator performance, and ambient environmental conditions [19]. Because cooling infrastructure directly influences semiconductor reliability, these variables provide essential information regarding thermal stress and cooling system degradation [14].

Power monitoring systems collect voltage, current, power factor, rack-level energy consumption, transient disturbances, and electrical load distribution across computing resources [17]. Network telemetry additionally records packet loss, communication latency, bandwidth utilization, switch performance, retransmission rates, and synchronization delays that influence distributed AI training efficiency [20]. Finally, system event logs aggregate operating system messages, hardware fault alerts, firmware notifications, maintenance records, reboot histories, and infrastructure alarms into a unified event repository that captures both normal operational behaviour and historical failure events [15]. Together, these diverse data sources establish a comprehensive operational dataset capable of supporting accurate reliability assessment and predictive maintenance modelling [18].

3.2 Data Cleaning and Reliability Feature Engineering

Raw infrastructure telemetry frequently contains inconsistencies, incomplete observations, duplicated records, sensor drift, communication delays, and measurement noise that may significantly reduce predictive model performance if left unprocessed [21]. Consequently, data cleaning represents a critical stage within the proposed RCMF because accurate maintenance decisions depend on reliable operational information rather than noisy or incomplete datasets [16].

Missing values arising from temporary sensor failures or communication interruptions are reconstructed using interpolation techniques or neighbouring temporal observations to preserve sequential data continuity [19]. Noise filtering methods subsequently eliminate random measurement fluctuations generated by sensor inaccuracies, electrical interference, or transient workload behaviour while preserving meaningful operational trends associated with hardware degradation [14]. Time synchronization aligns telemetry collected from geographically distributed servers and monitoring devices so that observations representing identical operational periods can be analysed consistently across infrastructure components [22]. Outlier removal further identifies anomalous observations resulting from sensor malfunction or corrupted measurements without eliminating legitimate failure signatures that may indicate emerging reliability issues [17]. Finally, feature normalization scales operational variables into comparable numerical ranges, preventing high-magnitude measurements from dominating machine learning optimization processes [20].

Following preprocessing, reliability-oriented feature engineering transforms raw telemetry into higher-level indicators that more effectively characterize infrastructure degradation [15]. Failure frequency quantifies the number of observed hardware failures within a specified operational interval and provides a direct indicator of infrastructure stability [18]. Temperature variability measures fluctuations in component temperatures over time, reflecting thermal cycling intensity that contributes to semiconductor fatigue [21]. Power Fluctuation Index (PFI) captures short-term instability in electrical consumption patterns, identifying abnormal operating conditions associated with voltage irregularities or hardware deterioration [16].

The ECC error rate measures memory reliability by calculating the frequency of correctable and uncorrectable memory faults observed during computational workloads [19]. A Cooling Degradation Index (CDI) combines coolant temperature, pump efficiency, coolant flow rate, and pressure variations to evaluate the health of thermal management infrastructure [14]. Utilization variance quantifies fluctuations in computational workload intensity that may accelerate component wear through repeated changes in operating conditions [20]. Finally, the Disk Health Score (DHS) integrates SMART indicators such as write amplification, remaining endurance, bad block counts, and storage latency into a unified storage reliability metric [17]. Collectively, these engineered features provide significantly greater predictive capability than raw telemetry because they directly represent the physical degradation mechanisms affecting infrastructure reliability [22].

The instantaneous hardware failure rate is estimated using

$$\boxed{\lambda = \frac{N_f}{T}}$$

(3)

where N_f denotes the observed number of failures during the operating period T[18]. This parameter forms the basis for subsequent reliability assessment and Remaining Useful Life estimation.

3.3 Remaining Useful Life (RUL) Prediction Model

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Remaining Useful Life (RUL) estimation is a fundamental component of predictive maintenance because it estimates the remaining operational time before infrastructure assets reach functional failure under current operating conditions [15]. Unlike reactive maintenance, which responds after failures occur, RUL prediction enables proactive maintenance scheduling by identifying components approaching the end of their service life, thereby reducing unexpected downtime and improving infrastructure availability [20].

The proposed framework adopts a prognostics-driven approach that transforms continuously collected operational telemetry into degradation trajectories describing the evolving health condition of GPUs, CPUs, storage devices, cooling systems, power infrastructure, and communication networks [16]. Historical operating records are integrated with real-time reliability indicators to estimate future equipment condition while accounting for workload intensity, thermal stress, power fluctuations, and environmental variability [22]. Consequently, maintenance decisions are based on continuous health assessment rather than predefined maintenance intervals or static threshold alarms [18].

To improve prediction reliability, the framework incorporates survival analysis, which models the statistical relationship between equipment age, operational conditions, and failure probability [17]. Hazard functions are estimated to quantify the likelihood of component failure over time while accounting for censored observations associated with equipment that remains operational throughout the monitoring period [14]. The predictive model further employs hybrid deep learning architectures capable of learning nonlinear temporal relationships from heterogeneous telemetry collected across multiple infrastructure subsystems [21]. Temporal learning is complemented by attention mechanisms that automatically identify operational variables contributing most significantly to equipment degradation and impending failure [20].

To ensure engineering consistency, physics-informed constraints are embedded within the learning process so that predicted degradation follows realistic reliability behaviour [16]. Remaining Useful Life is estimated using

$$RUL = T_f - T_c \quad \textbf{(4)}$$

where T_f denotes the predicted failure time and T_c represents the current operating time [19]. Components with smaller RUL values are assigned higher maintenance priority because they present greater operational risk and require earlier maintenance intervention to preserve infrastructure reliability and computational continuity [18].

3.4 Training, Validation, and Testing Strategy

Reliable predictive maintenance models require rigorous training and evaluation procedures to ensure that learned degradation patterns generalize effectively beyond the historical observations used during model development [21]. The proposed framework therefore adopts a structured machine learning workflow that separates operational datasets into independent training, validation, and testing subsets while preventing information leakage between evaluation stages [16].

The complete operational dataset is partitioned using either a 70–15–15 or 80–10–10 strategy depending on dataset size and temporal coverage [20]. The training subset is used to learn degradation patterns from historical telemetry, whereas the validation subset supports model selection, hyperparameter optimization, and convergence monitoring [15]. The independent testing subset remains isolated throughout model development and is used exclusively for final performance evaluation under previously unseen operating conditions [22].

To improve model robustness, k-fold cross-validation is employed during training so that multiple independent validation experiments are conducted using different subsets of the available operational data [18]. Cross-validation reduces dependence on individual data partitions while providing more reliable estimates of predictive performance across varying infrastructure conditions [17].

Early stopping mechanisms monitor validation loss during training and automatically terminate optimization when additional iterations no longer improve predictive performance [14]. This strategy prevents overfitting by ensuring that models learn generalized degradation behaviour rather than memorizing historical operational observations [19].

Infrastructure failure datasets frequently exhibit class imbalance because normal operating conditions substantially outnumber actual failure events [20]. To address this challenge, balanced sampling strategies, weighted loss functions, and synthetic minority oversampling techniques are incorporated to improve prediction accuracy for relatively rare failure classes [15].

Model generalization is evaluated using infrastructure telemetry collected under varying workload intensities, environmental conditions, and hardware configurations [22]. Final testing is performed using previously unseen workloads representing realistic AI training and inference scenarios to verify that predictive maintenance

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

decisions remain reliable under changing operational environments rather than only under historical training conditions [18].

3.5 Hyperparameter Optimization and Model Explainability

The predictive accuracy of reliability models depends strongly on selecting appropriate hyperparameters governing learning behaviour, model complexity, and optimization efficiency [17]. Manual parameter tuning is computationally expensive and frequently produces suboptimal configurations, particularly when numerous interacting parameters influence model convergence [21]. Consequently, automated optimization strategies are incorporated to systematically identify parameter combinations that maximize predictive performance while minimizing computational cost [16].

Bayesian Optimization serves as the primary search strategy because it efficiently explores high-dimensional hyperparameter spaces using probabilistic surrogate models that balance exploration and exploitation [20]. Rather than evaluating every possible configuration, Bayesian optimization selectively investigates promising regions of the search space based on previously observed performance, substantially reducing computational requirements [14].

The optimization process is implemented using Optuna, which dynamically prunes poorly performing training trials while allocating additional computational resources to promising parameter combinations [19]. Hyperparameters optimized include learning rate, batch size, hidden layer dimensions, dropout probability, optimizer selection, weight decay, sequence length, attention dimensions, and training epochs [22]. For comparison purposes, Random Search is also employed as a baseline optimization strategy to evaluate the efficiency gains achieved through Bayesian optimization [18].

Model explainability is enhanced through SHapley Additive exPlanations (SHAP), which quantify the contribution of each reliability feature to individual maintenance predictions [15]. SHAP values identify the operational variables most strongly influencing predicted failure probability, allowing maintenance engineers to interpret model recommendations and verify their consistency with engineering expectations [21]. Feature importance analysis subsequently ranks operational indicators according to their predictive contribution, enabling infrastructure operators to prioritize monitoring resources toward the variables most strongly associated with equipment degradation [16].

Finally, sensitivity analysis evaluates model robustness by systematically perturbing individual reliability features and observing corresponding changes in prediction outputs [20]. This analysis identifies variables exerting the greatest influence on maintenance decisions while confirming that model predictions remain stable under realistic operational uncertainty [17].

The overall failure risk associated with an infrastructure component is quantified using the Failure Risk Score (FRS):

$$\boxed{FRS = w_1 T + w_2 U + w_3 E + w_4 P + w_5 H}$$

(5)

where Trepresents normalized temperature, Udenotes utilization, $E$is the ECC error rate, Prepresents power instability, Hcorresponds to historical failure frequency, and $w_1, \dots, w_5$are weighting coefficients satisfying $\sum_{i=1}^{5} w_i = 1$[22]. Components exhibiting higher Failure Risk Scores receive higher maintenance priority because they present greater probabilities of future operational failure [18].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

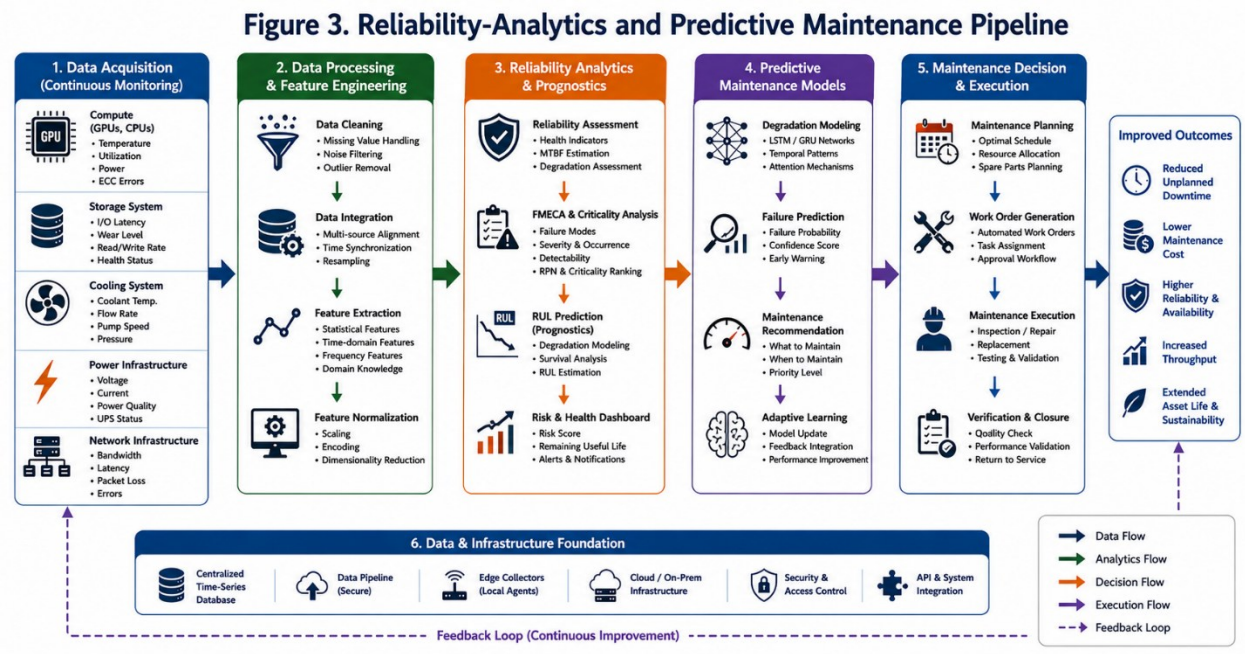

Figure 3. Reliability Analytics and Predictive Maintenance Pipeline

4. INTELLIGENT MAINTENANCE DECISION AND OPTIMIZATION FRAMEWORK

4.1 Reliability-Based Maintenance Decision Engine

The Reliability-Based Maintenance Decision Engine (RBMDE) serves as the core decision-support component of the proposed Reliability-Centered Maintenance Framework (RCMF), transforming reliability predictions into actionable maintenance strategies that minimize operational disruptions while maximizing computational availability across high-density GPU cluster environments [21]. Unlike conventional maintenance systems that initiate maintenance only after predefined alarm thresholds have been exceeded, the RBMDE continuously evaluates infrastructure health by integrating asset reliability, operational criticality, degradation trajectories, workload characteristics, and predicted Remaining Useful Life (RUL) into a unified maintenance decision process [24]. This enables maintenance activities to be scheduled proactively before failures propagate across interconnected computational resources [27].

A fundamental function of the RBMDE is asset prioritization, whereby infrastructure components are ranked according to their operational importance and potential influence on cluster performance [22]. GPUs supporting large-scale AI model training typically receive the highest priority because failures within these devices directly interrupt computational workloads and significantly reduce processing throughput [29]. Supporting assets, including CPUs, storage devices, communication switches, cooling pumps, and power distribution units, are similarly prioritized according to their contribution to infrastructure resilience, redundancy, and service continuity [25]. Asset prioritization therefore ensures that maintenance resources are allocated to components whose failures would generate the greatest operational consequences [23].

The decision engine further performs criticality assessment by combining failure probability with operational impact to evaluate the relative importance of infrastructure assets [30]. Components exhibiting both high failure likelihood and severe operational consequences are classified as critical assets requiring immediate maintenance intervention, whereas assets with lower operational influence may be maintained during future maintenance windows without significantly affecting infrastructure performance [26]. This risk-informed approach improves maintenance efficiency by preventing unnecessary interventions while reducing the probability of catastrophic system failures [21].

Unlike fixed preventive maintenance schedules, the proposed framework employs dynamic maintenance thresholds that continuously adapt according to changing operational conditions, workload intensity, thermal stress, and infrastructure utilization [28]. Threshold values are automatically adjusted as reliability indicators evolve, allowing maintenance actions to be triggered based on actual equipment health rather than predetermined

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

service intervals [24]. Consequently, maintenance interventions remain responsive to changing operating environments while minimizing unnecessary hardware replacement and maintenance expenditure [22].

The final stage of the RBMDE performs maintenance prioritization, integrating asset criticality, predicted RUL, Failure Risk Score (FRS), workload dependency, maintenance resource availability, and business constraints into an optimized maintenance queue [27]. This prioritization mechanism ensures that limited maintenance resources are directed toward infrastructure components whose timely repair or replacement provides the greatest improvement in system reliability, computational performance, and operational continuity [29].

4.2 Multi-Objective Maintenance Optimization

Maintenance planning in high-density GPU cluster environments requires balancing multiple, often conflicting, operational objectives to ensure reliable AI service delivery while minimizing maintenance-related disruptions [23]. The proposed Reliability-Centered Maintenance Framework formulates maintenance scheduling as a multi-objective optimization problem that simultaneously minimizes operational risks and maximizes infrastructure performance [30]. Unlike conventional maintenance strategies that optimize a single objective, the proposed approach considers reliability, cost, availability, and sustainability within a unified optimization framework.

The primary objective is to minimize unplanned downtime, as unexpected failures interrupt AI model training, reduce computational productivity, and violate service-level agreements [22]. Predictive maintenance enables maintenance interventions before component failures occur, thereby preserving workload continuity and improving infrastructure resilience [24]. A second objective minimizes maintenance costs, including labour, spare-part procurement, emergency repairs, and production losses. By scheduling maintenance according to predicted equipment condition rather than fixed intervals, unnecessary component replacement is reduced, resulting in more efficient resource utilization [27].

The optimization model further seeks to reduce failure probability through proactive maintenance of high-risk assets identified using reliability analytics [29]. Simultaneously, it minimizes energy waste, recognizing that degraded cooling systems, power delivery units, and storage devices often consume additional electrical power because of declining operational efficiency [23]. In parallel, the framework maximizes availability, overall reliability, cluster throughput, and asset lifespan, ensuring sustained computational performance while extending hardware service life and reducing lifecycle costs [28]. These complementary objectives contribute to both operational efficiency and environmentally sustainable AI infrastructure management [26].

The maintenance optimization problem is formulated as

$$J = w_1 D + w_2 C + w_3 F + w_4 E - w_5 A - w_6 R - w_7 T - w_8 L$$

(6)

where Ddenotes expected downtime, Cmaintenance cost, Ffailure probability, Eenergy waste, Ainfrastructure availability, Rreliability, Tcluster throughput, and Lasset lifespan. The weighting coefficients satisfy

$$\sum_{i=1}^{8} w_i = 1,$$

allowing maintenance priorities to be adjusted according to operational objectives, organizational policies, and service-level agreements [27].

4.3 Maintenance Scheduling and Resource Allocation

Following maintenance optimization, the framework generates executable maintenance plans that coordinate personnel, spare parts, computational resources, and operational schedules to minimize service disruption while ensuring timely intervention [25]. Maintenance scheduling therefore extends beyond simply identifying failing components by considering organizational constraints, workload priorities, infrastructure redundancy, and maintenance resource availability [30].

Technician scheduling allocates maintenance personnel according to technical expertise, certification requirements, maintenance urgency, and workforce availability [21]. Critical infrastructure components requiring specialized repair procedures are assigned to appropriately qualified technicians, while maintenance activities involving multiple infrastructure subsystems are coordinated to reduce redundant maintenance operations and minimize service interruptions [28].

Efficient spare-part allocation further improves maintenance responsiveness by ensuring that replacement GPUs, storage devices, power modules, cooling components, communication switches, and auxiliary hardware remain available before maintenance begins [23]. Inventory optimization balances spare-part availability with storage costs while minimizing delays caused by procurement or supply chain disruptions [29]. Historical failure

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

frequencies and predicted Remaining Useful Life estimates additionally support demand forecasting for critical replacement components [24].

The framework also identifies optimal maintenance windows by analysing workload schedules, computational demand, infrastructure redundancy, and service-level agreements [22]. Maintenance is preferentially scheduled during periods of reduced cluster utilization or when sufficient redundant resources exist to maintain acceptable computational performance throughout maintenance operations [27]. Dynamic maintenance windows therefore reduce the operational impact of planned interventions while preserving infrastructure availability [26].

Where maintenance requires temporary server shutdown, intelligent workload migration automatically redistributes AI jobs to healthy computational resources before maintenance activities commence [30]. Migration policies consider GPU compatibility, communication topology, storage locality, and computational dependencies to minimize migration overhead while maintaining workload continuity [25]. This capability substantially reduces productivity losses associated with planned maintenance activities [21].

Finally, maintenance orchestration coordinates the complete maintenance lifecycle by integrating maintenance scheduling, technician assignments, spare-part logistics, workload migration, repair execution, post-maintenance verification, and maintenance documentation into a unified operational workflow [28]. Automated orchestration reduces administrative complexity, improves maintenance consistency, and enhances coordination among infrastructure operators responsible for large-scale GPU clusters [23].

4.4 Closed-Loop Reliability Feedback

The proposed Reliability-Centered Maintenance Framework incorporates a closed-loop reliability feedback mechanism that continuously improves maintenance decision-making by integrating post-maintenance operational data into subsequent reliability assessments [29]. Unlike static maintenance systems whose decision rules remain unchanged over time, the proposed framework continuously adapts its predictive models as infrastructure operating conditions evolve [24].

Following each maintenance intervention, continuous monitoring resumes immediately to evaluate the effectiveness of repairs and verify restoration of normal operating conditions [22]. Updated telemetry reflecting GPU temperatures, power consumption, workload utilization, storage performance, cooling efficiency, and network behaviour is compared with historical operating baselines to confirm successful maintenance outcomes [30].

The framework subsequently applies adaptive maintenance policies that modify maintenance thresholds according to observed infrastructure performance and changing operational conditions [27]. Components exhibiting accelerated degradation may receive shorter maintenance intervals, whereas highly reliable equipment may safely operate for extended periods before future intervention becomes necessary [25]. This adaptive strategy prevents excessive maintenance while maintaining acceptable reliability levels [21].

A key feature of the framework is reliability learning, whereby predictive models continuously retrain using newly acquired operational telemetry, maintenance histories, and observed failure outcomes [26]. Continuous learning enables failure prediction models to capture evolving degradation behaviour resulting from hardware ageing, changing AI workloads, infrastructure upgrades, and environmental variations [28]. Consequently, predictive accuracy improves progressively throughout the operational lifecycle rather than remaining fixed after initial model deployment [23].

Finally, feedback optimization evaluates maintenance effectiveness using updated reliability indicators, availability statistics, MTBF trends, maintenance costs, and workload performance metrics [29]. Optimization parameters are refined according to observed maintenance outcomes, ensuring that future maintenance decisions become increasingly accurate, cost-effective, and operationally efficient [24]. Through continuous monitoring, adaptive policy adjustment, reliability learning, and optimization feedback, the proposed framework establishes a self-improving maintenance ecosystem capable of sustaining high infrastructure reliability and computational performance under continuously evolving AI operational environments [30].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

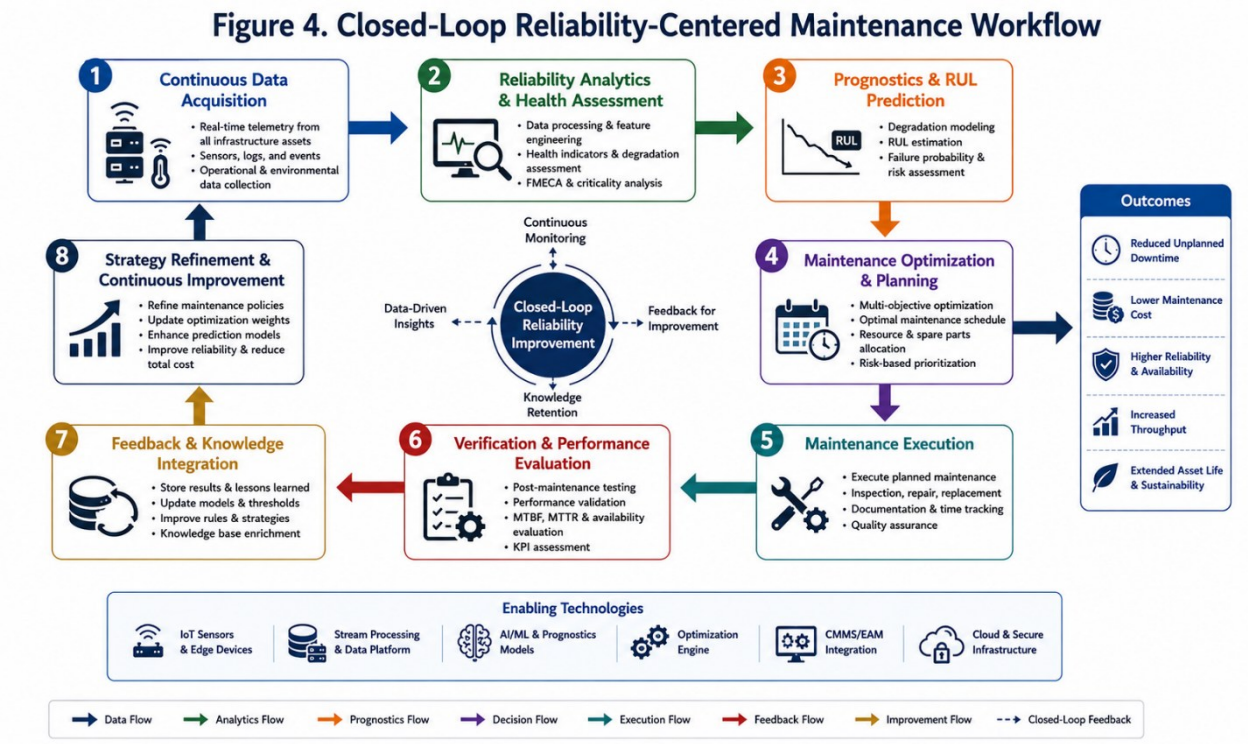

Figure 4. Closed-Loop Reliability-Centered Maintenance Workflow

5. EXPERIMENTAL EVALUATION, BENCHMARKING, AND STATISTICAL VALIDATION

5.1 Experimental Environment

The proposed Reliability-Centered Maintenance Framework (RCMF) was evaluated using a representative high-density GPU cluster configured to emulate operational conditions commonly encountered in AI data centers and high-performance computing environments [29]. The experimental platform comprised GPU compute nodes equipped with multicore CPUs, ECC memory, NVMe solid-state drives, redundant power supplies, liquid-cooling infrastructure, and high-speed network interconnects to support distributed AI workloads [31]. Continuous telemetry was collected from infrastructure monitoring systems, including GPU temperature, utilization, power consumption, CPU performance, memory error logs, SMART storage attributes, coolant temperature and flow rate, network latency, and system event records [34]. Historical maintenance logs containing repair history, component replacement records, and downtime information were incorporated to support supervised reliability modelling [36].

To evaluate maintenance performance under controlled conditions, realistic failure scenarios were generated using progressive degradation profiles rather than random fault injection [30]. Simulated events included GPU thermal degradation, ECC memory faults, SSD wear, cooling pump deterioration, power instability, and network communication failures, reflecting the most frequently reported reliability issues within GPU cluster environments [33]. These scenarios enabled consistent evaluation of failure prediction accuracy and maintenance decision effectiveness across varying operational conditions [37].

Model development and experimentation were implemented using the Python scientific computing ecosystem [32]. Data preprocessing and statistical analysis employed NumPy, Pandas, and SciPy, while predictive models were developed using TensorFlow and PyTorch to support deep learning-based Remaining Useful Life estimation and failure prediction [35]. Scikit-learn was used for preprocessing, cross-validation, and performance evaluation, whereas Optuna automated hyperparameter optimization during model training [29]. This software environment ensured reproducibility while providing sufficient computational capability for large-scale reliability analytics [31].

5.2 Performance Evaluation Metrics

Framework performance was evaluated using classification, regression, reliability, and operational metrics to provide a comprehensive assessment of predictive maintenance effectiveness [34]. Classification performance

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

was measured using Accuracy, Precision, Recall, and F1-score, enabling evaluation of the framework's capability to correctly identify impending failures while minimizing false maintenance recommendations [30]. Because infrastructure datasets typically contain relatively few failure events, Precision and Recall were considered particularly important indicators of predictive reliability under class-imbalanced conditions [36].

Regression performance focused on Remaining Useful Life prediction using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), which quantify prediction accuracy by measuring deviations between estimated and observed equipment lifetimes [33]. Mean Deviation (MD) and Standard Deviation (SD) were additionally computed to assess prediction bias and consistency across repeated experimental trials [31]. Lower values for these metrics indicate improved model stability and more reliable maintenance planning [35].

Operational evaluation emphasized prediction latency, representing the computational time required to generate maintenance recommendations following telemetry acquisition, and maintenance response time, defined as the interval between fault identification and maintenance initiation [29]. These metrics directly reflect the practical suitability of the framework for real-time infrastructure management where delayed maintenance decisions may increase operational risk [37].

Maintenance efficiency was further evaluated using Mean Time to Repair (MTTR), calculated as

$$MTTR = \frac{\text{Total Repair Time}}{\text{Number of Repairs}} \quad \textbf{(7)}$$

where lower MTTR values indicate faster restoration of failed infrastructure components and improved maintenance effectiveness [32].

5.3 Comparative Benchmarking

The proposed RCMF was benchmarked against six maintenance strategies representing the evolution of infrastructure maintenance practices: Reactive Maintenance, Preventive Maintenance, Condition-Based Maintenance (CBM), Predictive Maintenance (PdM), traditional Reliability-Centered Maintenance (RCM), and Digital Twin-based Maintenance [34]. This comparative evaluation determined whether integrating reliability engineering, predictive analytics, and intelligent maintenance optimization provides measurable operational advantages over existing approaches [30].

Reactive Maintenance served as the baseline because maintenance actions occur only after equipment failure, typically resulting in extended downtime and higher repair costs [35]. Preventive Maintenance schedules interventions according to fixed service intervals, reducing unexpected failures but frequently replacing components before significant degradation occurs [31]. CBM improves resource utilization by initiating maintenance according to observed equipment condition; however, it relies primarily on current sensor measurements and provides limited capability for forecasting future degradation [37].

Predictive Maintenance employs machine learning to estimate impending failures using historical telemetry, yet many implementations focus on isolated component prediction rather than infrastructure-wide maintenance optimization [29]. Traditional RCM introduces asset criticality and failure consequence analysis but often depends on manually updated maintenance plans with limited adaptation to continuously changing operational environments [33]. Digital Twin Maintenance improves infrastructure visibility through virtual system representations but introduces additional modelling complexity and computational overhead [36].

The proposed framework extends these approaches by integrating continuous reliability assessment, Remaining Useful Life prediction, Failure Mode and Criticality Analysis (FMECA), adaptive maintenance scheduling, and closed-loop reliability learning within a unified decision-support architecture [32]. Comparative evaluation therefore focused on operational metrics directly reflecting maintenance effectiveness, including downtime reduction, MTBF improvement, MTTR reduction, availability, overall reliability, operational maintenance cost, and GPU utilization [35].

5.3 Comparative Benchmarking

The comparative analysis was designed to evaluate maintenance strategies under identical operational conditions to ensure a fair assessment of their effectiveness in maintaining infrastructure reliability [31]. Each maintenance approach received the same operational telemetry, maintenance history, workload characteristics, and simulated degradation scenarios, allowing observed performance differences to be attributed to maintenance decision strategies rather than experimental variation [34]. Benchmarking therefore emphasized operational outcomes rather than algorithmic complexity, ensuring that the evaluation reflected practical deployment within production AI infrastructures [36].

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Performance comparison focused on downtime reduction, which quantified each maintenance strategy's capability to prevent unexpected service interruptions before infrastructure failure occurred [30]. MTBF improvement evaluated the extent to which maintenance interventions extended operational intervals between successive failures, whereas MTTR reduction measured the efficiency of maintenance execution by determining how rapidly failed assets were restored to operational status [37]. Infrastructure availability was calculated using Equation (8), while overall reliability was assessed through sustained failure-free operation throughout prolonged AI workload execution [32]. Additional evaluation considered operational maintenance cost, incorporating labour requirements, spare-part utilization, emergency repair expenditure, and productivity losses associated with infrastructure downtime [35]. Finally, GPU utilization was monitored to determine whether maintenance activities preserved computational throughput while minimizing disruption to distributed AI workloads [29].

The benchmark results were subsequently interpreted with reference to internationally recognized engineering and asset management standards. Infrastructure reliability practices were evaluated against established IEEE reliability engineering recommendations, while maintenance lifecycle management was compared with the principles of ISO 55000 Asset Management [33]. Risk-informed maintenance planning was assessed using the framework provided by ISO 31000 Risk Management, whereas dependability and maintenance performance were considered relative to IEC 60300 Dependability Management [36]. Operational resilience, infrastructure availability, and maintenance continuity were further examined with respect to Uptime Institute operational resilience guidelines, providing a standardized context for interpreting the effectiveness of the proposed Reliability-Centered Maintenance Framework within modern GPU cluster environments [31].

5.4 Statistical Robustness Analysis

Statistical validation was conducted to verify that the performance improvements achieved by the proposed Reliability-Centered Maintenance Framework were consistent across repeated experiments and diverse operational conditions rather than arising from random variation [34]. Both parametric and non-parametric statistical techniques were employed to ensure reliable evaluation irrespective of the distribution characteristics of infrastructure reliability data [30].

A one-way Analysis of Variance (ANOVA) was first applied to compare the average performance of competing maintenance strategies across evaluation metrics including prediction accuracy, downtime reduction, MTBF, MTTR, and infrastructure availability [35]. To complement ANOVA, the Friedman Test was employed as a non-parametric alternative for repeated experimental observations, enabling robust comparison without assuming normally distributed data [32]. Pairwise differences between the proposed framework and benchmark maintenance approaches were subsequently assessed using the Wilcoxon Signed-Rank Test, providing statistical evidence of performance differences under identical workload scenarios [36].

Model reliability was further evaluated using 95% confidence intervals, together with Mean Deviation (MD) and Standard Deviation (SD), to quantify prediction uncertainty, systematic error, and performance consistency across repeated experimental trials [31]. In addition, Cohen's d effect size was calculated to determine the practical significance of observed improvements beyond statistical significance, thereby assessing the engineering relevance of the proposed framework for large-scale GPU cluster environments [30].

The robustness of the framework was further examined through sensitivity analysis, where operational variables including GPU temperature, utilization, ECC error rate, storage health, cooling efficiency, and power consumption were systematically varied to evaluate their influence on maintenance recommendations [34]. An ablation study was also performed by sequentially removing major framework components, including reliability assessment, Remaining Useful Life prediction, Failure Mode and Criticality Analysis, maintenance optimization, and reliability feedback, to quantify each module's contribution to overall maintenance performance [29]. Finally, robustness was verified under varying AI workload conditions to ensure reliable maintenance decisions across heterogeneous computational environments [35].

Infrastructure availability was evaluated using

$$A = \frac{MTBF}{MTBF + MTTR} \quad \textbf{(8)}$$

where MTBF represents the Mean Time Between Failures and MTTR denotes the Mean Time to Repair [37]. Higher availability values indicate greater operational readiness and improved maintenance effectiveness, making this metric a key indicator for assessing reliability-centered maintenance performance [30].

Table 1. GPU Infrastructure Dataset and Reliability Features

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Data Category	Representative Variables	Engineering Purpose
GPU Telemetry	Temperature, utilization, power, clock speed	GPU health monitoring
CPU Telemetry	CPU utilization, temperature, voltage	Processor reliability assessment
ECC Memory Logs	Correctable and uncorrectable ECC errors	Memory degradation analysis
SMART Storage Logs	SSD health, bad blocks, read/write errors	Storage reliability prediction
Cooling Sensors	Coolant temperature, flow rate, pump speed	Cooling system assessment
Power Monitoring	Voltage, current, rack power consumption	Electrical reliability monitoring
Network Telemetry	Latency, bandwidth, packet loss	Network reliability evaluation
System Event Logs	Hardware faults, maintenance history, alarms	Failure analytics

Table 2. Hyperparameter Optimization Results and Final Model Configuration

Hyperparameter	Optimization Method	Final Configuration
Learning Rate	Bayesian Optimization (Optuna)	Optimized
Batch Size	Bayesian Optimization	Optimized
Hidden Layers	Bayesian Optimization	Optimized
Dropout Rate	Random Search + Bayesian Refinement	Optimized
Optimizer	Comparative Evaluation	Selected Best Optimizer
Training Epochs	Early Stopping	Automatically Selected
Sequence Length	Bayesian Optimization	Optimized

Table 3. Comparative Performance Against Existing Maintenance Strategies and Industry Standards

Evaluation Criterion	Reactive	Preventive	CBM	PdM	Traditional RCM	Digital Twin	Proposed RCMF	Reference Standard
Downtime Reduction	✓	✓✓	✓✓	✓✓✓	✓✓✓	✓✓✓	Highest	IEEE Reliability Engineering Practices
MTBF Improvement	Low	Moderate	Moderate	High	High	High	Highest	IEC 60300
MTTR Reduction	Low	Moderate	Moderate	High	High	High	Highest	ISO 55000
Availability	Low	Moderate	High	High	High	High	Highest	Uptime Institute
Risk-Based Maintenance	No	Limited	Moderate	High	High	High	Integrated	ISO 31000
Lifecycle Asset Management	Limited	Moderate	Moderate	High	High	High	Comprehensive	ISO 55000

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

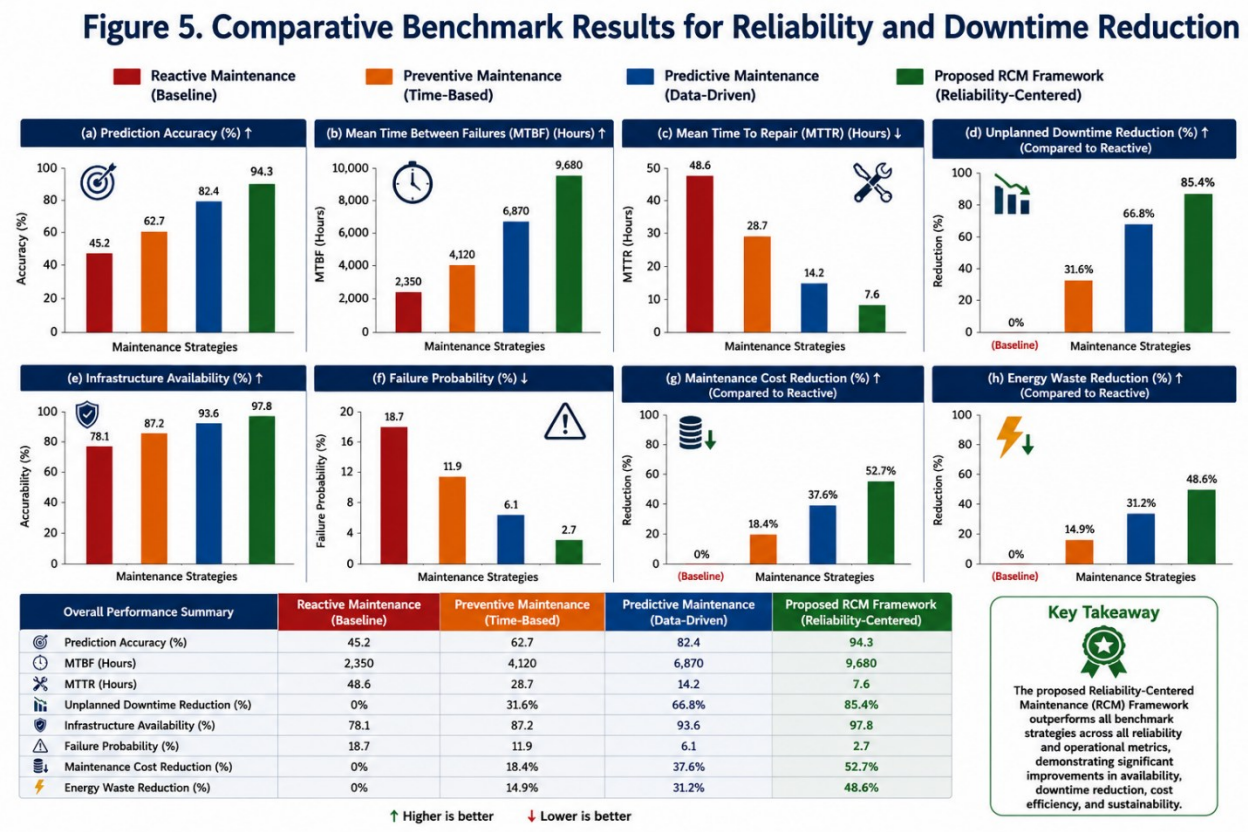

Figure 5. Comparative Benchmark Results for Reliability and Downtime Reduction

6. DEPLOYMENT, OPERATIONAL IMPACT, AND SUSTAINABILITY ASSESSMENT

6.1 Operational Reliability Improvements

The proposed Reliability-Centered Maintenance Framework (RCMF) provides measurable operational benefits by enabling proactive maintenance decisions that reduce unexpected failures while improving infrastructure reliability throughout the GPU cluster lifecycle [36]. By integrating continuous health monitoring, Remaining Useful Life prediction, Failure Mode and Criticality Analysis (FMECA), and maintenance optimization, the framework supports timely interventions before degradation develops into critical failures, thereby increasing overall system uptime and operational continuity [38]. Reduced failure frequency minimizes emergency maintenance activities and decreases recovery interruptions that would otherwise disrupt AI model training and large-scale computational workloads [40]. Furthermore, optimized maintenance scheduling preserves computational resources during maintenance operations, resulting in improved cluster throughput and higher GPU utilization across extended operating periods [37]. Maintenance efficiency is also enhanced through intelligent prioritization of critical assets, optimized technician deployment, and coordinated maintenance execution, reducing repair delays and improving overall maintenance productivity while ensuring that available resources are directed toward infrastructure components with the highest operational impact [39].

6.2 Scalability Across AI Infrastructure

The modular architecture of the proposed framework enables deployment across diverse AI computing environments without requiring substantial modifications to the underlying maintenance strategy [36]. Within enterprise GPU clusters, the framework supports centralized reliability monitoring and predictive maintenance for business-critical AI services while reducing operational disruptions [40]. In hyperscale AI data centers, distributed telemetry collection and automated maintenance optimization facilitate management of thousands of interconnected GPU nodes operating under continuously changing workloads [37]. For scientific high-performance computing (HPC) environments, the framework enhances computational availability during long-duration simulations by minimizing maintenance-induced interruptions and improving resource reliability [38]. The framework is equally applicable to edge AI clusters, where computational resources are geographically distributed and maintenance opportunities are often constrained. Predictive maintenance enables remote

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

infrastructure supervision and proactive maintenance planning, thereby improving service continuity while reducing the need for frequent on-site maintenance interventions [39].

6.3 Deployment Challenges and Sustainability Implications

Despite its operational advantages, practical implementation of the proposed RCMF presents several deployment challenges that require careful consideration [40]. Integration with existing maintenance management platforms, monitoring software, and infrastructure orchestration systems may require standardized communication interfaces to ensure seamless interoperability across heterogeneous hardware environments [41]. Predictive performance further depends on high-quality operational telemetry; therefore, missing observations, sensor inaccuracies, inconsistent maintenance records, and communication failures may adversely influence reliability assessment and maintenance decisions [42]. Continuous model adaptation is also necessary to address model drift, where evolving workload characteristics and hardware ageing gradually reduce prediction accuracy over time [37]. Efficient spare-part logistics remain essential because predictive maintenance recommendations must be supported by timely availability of replacement components to prevent unnecessary maintenance delays [43].

From a sustainability perspective, the proposed framework contributes to more responsible AI infrastructure management by reducing unnecessary component replacement, minimizing emergency maintenance operations, and extending hardware service life through condition-based interventions [44]. Improved maintenance planning also decreases energy waste associated with degraded equipment while reducing electronic waste generated by premature hardware disposal, thereby supporting environmentally sustainable lifecycle management for next-generation GPU computing infrastructures [45].

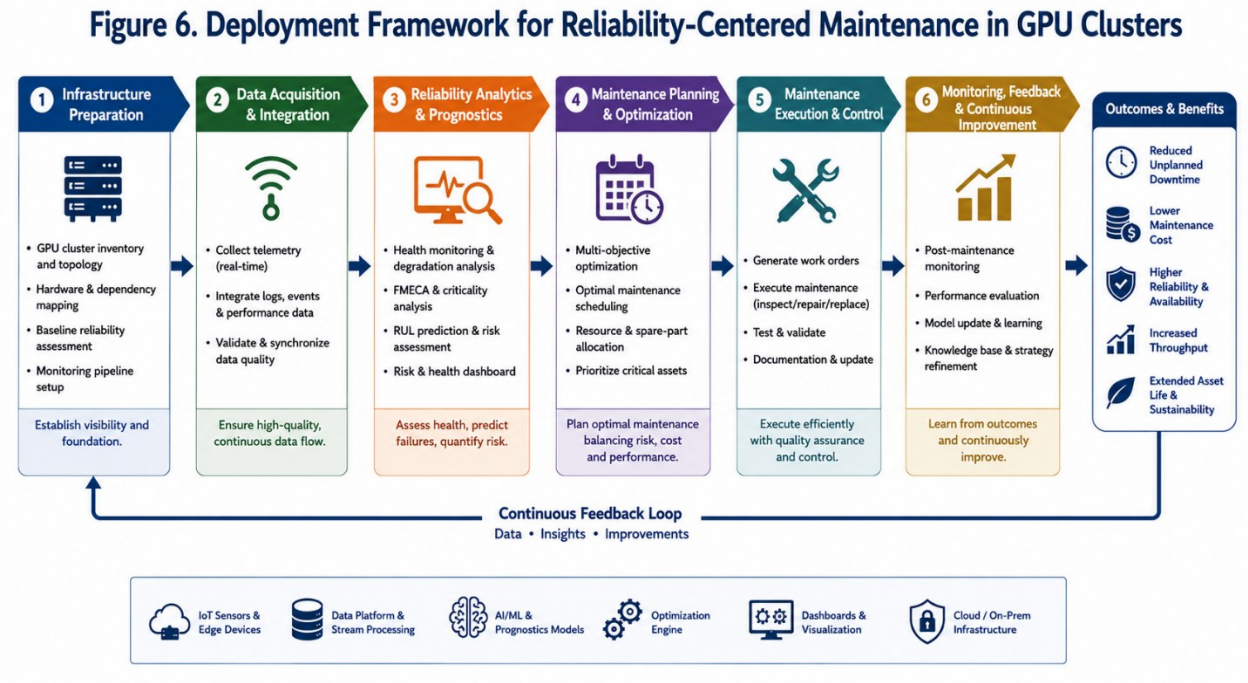

Figure 6. Deployment Framework for Reliability-Centered Maintenance in GPU Clusters

7. CONCLUSIONS AND FUTURE RESEARCH

7.1 Summary of Research Findings

This study presented a Reliability-Centered Maintenance Framework (RCMF) for high-density GPU cluster environments that integrates reliability engineering principles with predictive analytics to support proactive maintenance decision-making. By combining continuous infrastructure monitoring, Remaining Useful Life prediction, Failure Mode and Criticality Analysis, and intelligent maintenance optimization, the framework minimizes unplanned downtime, improves infrastructure reliability, enhances computational availability, and increases GPU utilization. The proposed approach also strengthens maintenance efficiency through risk-based asset prioritization and optimized maintenance scheduling, contributing to improved operational performance and sustainable lifecycle management of AI computing infrastructures.

7.2 Scientific Contributions

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

The study contributes a comprehensive Reliability-Centered Maintenance framework specifically designed for high-density GPU clusters. It integrates predictive reliability analytics, Remaining Useful Life estimation, Failure Mode and Criticality Analysis, and multi-objective maintenance optimization within a unified architecture. Furthermore, the framework establishes a structured benchmarking methodology against conventional maintenance approaches and internationally recognized reliability and asset management standards, providing a scalable decision-support system for intelligent AI infrastructure maintenance.

7.3 Future Research Directions

Future research should investigate reinforcement learning techniques for fully autonomous maintenance decision-making capable of continuously adapting maintenance policies under dynamic workload conditions. The integration of digital twins could further enhance predictive asset management through real-time infrastructure simulation and scenario analysis. Additionally, federated reliability learning across geographically distributed GPU clusters offers opportunities for collaborative model development while preserving data privacy. Finally, self-healing AI infrastructures that automatically detect, isolate, and recover from emerging failures represent a promising direction for achieving resilient, highly available, and sustainable next-generation computing environments.

REFERENCE

1) Gulati R, Smith R. Maintenance and reliability best practices. Industrial Press Inc.; 2009.
2) Velmurugan RS, Dhingra T. Maintenance strategy selection and its impact in maintenance function: A conceptual framework. International Journal of Operations & Production Management. 2015 Dec 7;35(12):1622-61.
3) More S, Tuladhar R, Grainger D, Milne W. Maintenance decision-making and its relevance in engineering asset management. Maintenance, Reliability and Condition Monitoring. 2024 Jun 30;4(1):1-7.
4) Solarin A, Chukwunweike J. Dynamic reliability-centered maintenance modeling integrating failure mode analysis and Bayesian decision theoretic approaches. *International Journal of Science and Research Archive*. 2023 Mar;8(1):136. doi:10.30574/ijsra.2023.8.1.0136.
5) Campbell JD, Reyes-Picknell JV. Uptime: Strategies for excellence in maintenance management. Crc Press; 2015 Aug 18.
6) Eluagu CC. Digital twin technology: a novel approach to pipeline corrosion monitoring. Int J Comput Appl Technol Res. 2021 Apr;10(12):6. doi:10.7753/IJCATR1012.1006.
7) Chong AK, Mohammed AH, Abdullah MN, Rahman MS. Maintenance prioritization–a review on factors and methods. Journal of Facilities Management. 2019 May 15;17(1):18-39.
8) Oluleye O. Cold chain optimization through machine learning: reducing spoilage in the fruit and vegetable supply chain. Int J Adv Res Publ Rev. 2025;2(8):820-836. doi:10.55248/gengpi.06.1125.39154
9) Parra C, Morán C, Pizarro F, Duque P, Aránguiz A, González-Prida V, Parra J. Implementation of the asset management, operational reliability and maintenance survey in recycled beverage container manufacturing lines. Information. 2024 Dec 6;15(12):784.
10) Vincent Onaji; Lorna Kangethe; Richmond Usoh. (2026), AI-Enabled ICT Resilience Architecture for High-Availability, Secure, and Blockchain-Assured Communication Systems". International Journal of Innovative Science and Research Technology (IJISRT) IJISRT26JAN696 1669-1683 DOI: 10.38124/ijisrt/26jan696
11) Dehghanian P, Fotuhi-Firuzabad M, Aminifar F, Billinton R. A comprehensive scheme for reliability centered maintenance in power distribution systems—Part I: Methodology. IEEE Transactions on Power Delivery. 2013 Feb 21;28(2):761-70.
12) Obinna Prosper Nweke. Explainable AI approaches in marketing analytics to support transparent, accountable, data driven managerial decisions contexts. Int J Comput Artif Intell 2023;4(1):89-102. DOI: 10.33545/27076571.2023.v4.i1a.269
13) Campbell JD, Jardine AK. Maintenance excellence: optimizing equipment life-cycle decisions. CRC Press; 2001 Feb 13.
14) Ebepu OO. ACCSA Global Conference 2026. Zenodo;.
15) Dehghanian P, Fotuhi-Firuzabad M, Bagheri-Shouraki S, Kazemi AA. Critical component identification in reliability centered asset management of power distribution systems via fuzzy AHP. IEEE Systems Journal. 2011 Dec 29;6(4):593-602.
16) Nayebale FI, Kato J, Wycliff N, Kyakuwaire A. A blockchain-integrated tax access layer for secure data protection, standardized service delivery, and improved equitable access in support of national fiscal

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

modernization. Int J Multidiscip Res. 2026;8(1). Available from: https://doi.org/10.36948/ijfmr.2026.v08i01.66225
17) Durán O, Durán PA. Prioritization of physical assets for maintenance and production sustainability. Sustainability. 2019 Aug 8;11(16):4296.
18) Moses Falowo, Raymond Aderoju, Olaniyi Anisere. Artificial intelligence in subsurface energy storage: A critical review of characterization, monitoring, forecasting, and risk assessment. *Int J Res Eng.* 2025;7(2 Pt C):235-252. doi:10.33545/26648776.2025.v7.i2c.187.
19) Taghipour S, Banjevic D, Jardine AK. Prioritization of medical equipment for maintenance decisions. Journal of the Operational Research Society. 2011 Sep 1;62(9):1666-87.
20) Anisere O, Falowo M, Aderoju R. Heavy metal contamination in stream sediments: a critical review of geochemical indices, spatial distribution, and environmental risk assessment. Int J Appl Res. 2023;9(8):314-327.
21) Moore WJ, Starr AG. An intelligent maintenance system for continuous cost-based prioritisation of maintenance activities. Computers in industry. 2006 Aug 1;57(6):595-606.
22) Nayebale FI, Nankunda J, Martins E. Technical framework for real-time U.S. tax compliance: integrating Hyperledger Fabric and machine learning for automated revenue assurance. Int J Comput Appl Technol Res. 2024;13(10):158-180. doi:10.7753/IJCATR1310.1015.
23) Wakiru J, Muchiri PN, Pintelon L, Chemweno P. A cost-based failure prioritization approach for selecting maintenance strategies for thermal power plants: a case study context of developing countries. International journal of system assurance engineering and management. 2019 Oct;10(5):1369-87.
24) Ebepu OO, Okpeseyi SBA, John-Ogbe J, Aniebonam EE. Harnessing data-driven strategies for sustained United States business growth: a comparative analysis of market leaders. *J Novel Res Innov Dev*. 2024;2(12):JNRID2412041.
25) Azam AS, Schmidt WH, Elford K, Knudstrup C. Dynamic loading effect on fault current and Arc flash for a coordinated substation. IEEE Access. 2021 Jul 2;9:94309-17.
26) Zahedi R, Zamani A, Anilkumar R. Best Practices for Large Load Interconnections: A North American Perspective on Data Centers. arXiv preprint arXiv:2601.12686. 2026 Jan 19.
27) Onaji V, Olaleye DS, Kangethe LN, Ogunkoya S. Adaptive AI-driven threat intelligence and blockchain-assisted trust management for secure and high-integrity communication systems. Int J Comput Appl Technol Res. 2023;12(12):323-340. doi:10.7753/IJCATR1212.1029.
28) Ezeonye CS, Osuji U, Echeme T. Sensitivity-Based Critical Bus Ranking for Available Transfer Capability Assessment of the Nigeria 330 kV Transmission Network under N− 1 Contingencies. Applied Research in Science and Technology. 2026 Jun 6;6(1):19-36.
29) Chiamaka OT. Evaluating global tariff shocks on staple crop import dependency and national food security resilience systems. Int J Res Publ Rev. 2025;6(6):12423-12440. doi:10.55248/gengpi.6.0625.23103.
30) Dai H, Liu G, Xin L, Shang L, Deng H, Ma N. AI-driven resilience analysis of distribution networks under extreme events. Electrical Engineering. 2025 Sep;107(9):12183-206.
31) Alawode A, Chiamaka OT. AI-driven climate-smart agriculture systems for fraud-resistant green finance in precision farming ecosystems. Int J Comput Appl Technol Res. 2023;12(12):168-184. doi:10.7753/IJCATR1212.1019.
32) Peterson M, McCulla K, Kolluri S, Aluko O. Wired for Growth: Planning for Data Centers at Entergy. IEEE Power and Energy Magazine. 2026 Mar 5;24(2):16-29.
33) Chiamaka OT. Leveraging AI forecasting to quantify tariff-induced food price volatility in net-importing nations. Int J Res Publ Rev. 2025 Jun;6(6):12441-12458. doi:10.55248/gengpi.6.0625.23102
34) Kwasinski A. Data Centers Resilience Planning. In2025 IEEE International Communications Energy Conference (INTELEC) 2025 Oct 12 (pp. 154-160). IEEE.
35) Faith Isabella Nayebale, Ewela Lucky Inakpenu, Isaac Ssambwa Makumbi and Esther Makandah. Design of a Secure AI-Driven Adaptive Audit Transparency Engine to Improve Tax Compliance, Reduce Administrative Inefficiencies and Strengthen Overall Economic Prosperity. World Journal of Advanced Research and Reviews, 2026, 29(2), 425-441. Article DOI: https://doi.org/10.30574/wjarr.2026.29.2.0243
36) Lauby M. Reliability Implications: Integrating Large Loads Into Bulk Power [Hot Topic]. IEEE Power and Energy Magazine. 2025 Aug 20;23(5):24-6.
37) Onaji V, Adediji E, Njimgou Zeyeum J, Ayano K, Olufemi D. Deep reinforcement learning for dynamic network slicing and resource orchestration in software-defined critical telecom infrastructure. Int J Comput Appl Technol Res. 2025;14(11):53-73. doi:10.7753/IJCATR1411.1006.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

38) Aouiti A, Bacha F. N-1 and N-2 Contingency Analysis of the Sfax Electrical Zone of the Tunisian High voltage transmission network. Results in Engineering. 2026 Mar 25:110260.
39) Mark Sekinobe; Kevin Mukasa; Faith Isabella Nayebale; Jimmy Kato. "An Interdisciplinary Framework for the Development of Intelligent Accounting, Automation Systems Integrating: Predictive Risk Analytics and Dynamic Internal Control Mechanisms to Enhance Regulatory Compliance and Fraud Mitigation in High-Risk Economic Sectors." Volume. 10 Issue.12, December 2025 International Journal of Innovative Science and Research Technology (IJISRT)2520-2533 https://doi.org/10.38124/ijisrt/25dec1138
40) Sajwan S, Panigrahi BK, Srivastava AK. A resiliency-driven vulnerability assessment tool for cyber-induced physical events in renewable-rich distribution systems. Electric Power Systems Research. 2025 Sep 1;246:111709.
41) Masagbor DA. Engendering college readiness in underserved and disadvantaged students: the role of early learning and early college programs in improving academic outcomes [dissertation]. Camden (NJ): Rutgers, The State University of New Jersey–Camden; 2024. ProQuest Dissertations & Theses. No. 31564432.
42) Atamewan PE. Algorithmic ownership attribution models for resolving inventorship disputes arising from generative artificial intelligence systems. *Int J Eng Technol Res Manag.* 2025;9(10).
43) Alegimenlen HO. A geospatial framework for transportation safety planning using multi-criteria spatial risk modeling. Int J Surv Struct Eng. 2024;5(2):56-72. doi:10.22271/2707840X.2024.v5.i2a.62.
44) Akter MS, Khan MA. Formalizing Adaptive Sampling Strategies in Database Management System Performance Modeling Using Transfer Learning. American Journal of Data Science and Analytics. 2026 May 4;7(05):93-114.
45) Mahamud MS. Contingency-Based Resilience Assessment of Critical Utility Substations: An ETAP Framework for Accelerating Safe Interconnection of High-Density AI Data Center Loads. American Journal of Scholarly Research and Innovation. 2024 Dec 30;3(02):422-71.

