Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Intelligent Predictive Architectures for Autonomous Self-Healing in Cloud Computing: A Comprehensive Survey

Nitesh Gupta
Department of Computer Science
R.V. Institute of Technology and Management
VTU, Belagavi

Nandita Bangera
Department of Computer Science
R.V. Institute of Technology and Management
VTU, Belagavi

Abstract **- Neural network-enabled self-healing is becoming a very exciting approach to making cloud computing infrastructures more reliable, available, and efficient. This survey paper gives a detailed review of neural network-based predictive models and how they are combined with autonomous recovery mechanisms for self-healing cloud systems. It first delineates the main neural architectures used for cloud reliability, such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and hybrid or ensemble models, explaining their roles in failure prediction, resource management forecasting, and SLA/QoS violation prediction. The paper examines the coupling of these predictive models with self-healing action and decision layers like rule-based policies, policy engines, and reinforcement learning-driven controllers to proactively trigger recovery actions, reduce Mean Time To Recovery (MTTR), and control false alarms and resource overhead. Evaluation practices highlight datasets (Google Cluster Traces, Alibaba traces, and synthetic simulation data), performance metrics (prediction accuracy, MTTR, false positive/negative rates, scalability, and energy cost), and differences between simulated vs. real production cloud environments. Moreover, the survey discusses privacy and security issues of predictive and action models, presenting techniques such as federated learning, differential privacy, homomorphic encryption, and secure multi-party computation for privacy-preserving cross-tenant collaboration. Lastly, this paper draws attention to open issues including trade-offs between accuracy and false alarms, latency constraints, explainability, and scalability, proposing future work including explainable predictive pipelines, federated self-healing frameworks, multi-agent and DRL-based autonomy, and integration with emerging technologies such as quantum-inspired neural networks, edge-cloud continuum architectures, digital twins, service meshes, and neuromorphic hardware to enable more trustworthy, efficient, and autonomous self-healing cloud management.**

Index Terms **- Neural Predictive Intelligence, Self-Healing Architectures, Deep Learning for Cloud Systems, Multi-Agent Reinforcement Learning, Federated Self-Healing Frameworks, SLA Forecasting, Cloud Infrastructure Resilience**

I. INTRODUCTION

From the last 10 to 15 years, cloud computing has gained significant popularity. It is the most important part of modern digital infrastructure, fully changing the concept of handling data and how to use that data [19]. It has been adopted by various businesses and created new industries like e-commerce. Hosting web services was expensive and complicated; cloud computing made these things flexible, sustainable, cost-effective and easily scalable, making them necessary for growing businesses and AI/machine learning applications.

Due to on-demand resource allocation and the dynamic nature of cloud computing environments, they are highly susceptible to failures including VM crashes, resource overload, network failures, security failures, and operational errors [18]. Earlier approaches to handle failures were reactive, such as distributing workloads after faults or restarting failed services, which led to service unavailability. Organizations face huge financial losses, business disruption, reputation damage, data inaccessibility and loss, Service-Level Agreement (SLA) violations, security risks, and legal issues [2].

Recent examples demonstrate these challenges. On November 18, 2025, Cloudflare suffered a global outage causing widespread "500 error" messages, preventing many websites and services from loading. Popular platforms including social media, AI tools, and gaming sites were unreachable globally. In October 2025, AWS experienced a significant outage affecting numerous websites and cloud-based services globally. These incidents highlight weak points that can be exploited by hackers.

To counter these issues, researchers have developed self-healing environments where systems can detect or predict upcoming failures automatically, then take predefined steps to recover from failures without human interference [9], [13]. If a failure requires manual intervention, the system provides alerts before the error occurs so prevention steps can be taken.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

II. REACTIVE VS. PROACTIVE HEALING

The difference between reactive and proactive healing is a paradigm shift that changes the entire concept of managing cloud systems. Reactive healing methods only respond to failures after they occur, resulting in service disruptions and longer system recovery times. They are characterized by

Mean Time To Recovery (MTTR) values from 30 minutes to several hours depending on failure complexity and manual operation extent [12], [17].

Reactive mechanisms depend on threshold-based monitoring and alarm systems which signal corrective actions once alarm conditions are met. Studies show that reactive systems can lead to service interruptions averaging 5-15 minutes, potentially breaching strict SLA requirements even in well-engineered situations [17].

Proactive systems incorporate predictions from neural networks and use forecasts to prepare preventive measures well ahead of failures. Experimental research provides proof of MTTR improvements up to 85% (reducing recovery time to as little as 13.5 minutes) through proactive methodology in complicated cloud environments [17]. This advancement is mainly attributed to eliminating failure detection and diagnosis stages which constitute a large time chunk in reactive approaches.

The financial advantages of preventive healing are significant. Researchers assert that proactive approaches reduce expenses by 30-50% compared to reactive methods when considering unavailability costs, resource consumption, and recovery effort [17]. These systems shield against expensive failure episodes leading to penalty fees and client loss while maintaining optimal resource utilization via forecasted scaling.

Hybrid approaches combining reactive and proactive methods represent an optimal solution for cloud security [16]. These systems exploit prediction models for normal failure categories while applying reactive functions to sudden and unrecognized failure modes. Studies show hybrid systems achieve optimal balance between protection efficiency and implementation complexity.

TABLE I
COMPARISON OF PREDICTION MODELS

Model	Application	Acc.	Strength	Limitation
ANN	Failure pred.	≈95%	Simple, effective	Limited temporal
LSTM	Resource pred.	High	Time-series	High computation
CNN-LSTM	Failure pred.	≈98%	Spatial+temporal	Complex model
Ensemble	SLA pred.	Very High	Robust	High overhead

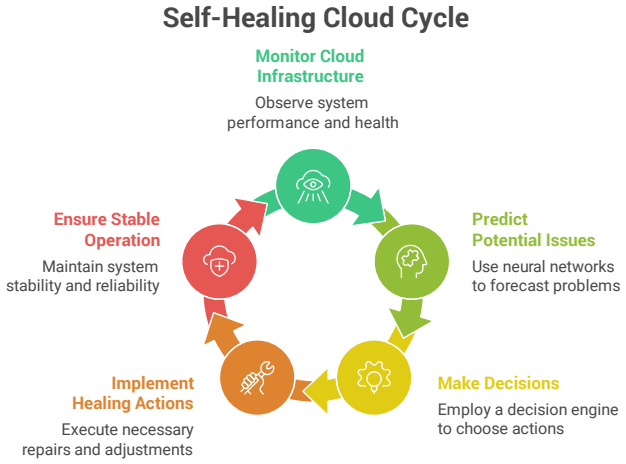

Fig. 1. Self-healing cloud operational cycle illustrating monitoring, prediction, decision-making, healing execution, and stability assurance stages

III. FAILURE PREDICTION MODELS

Neural networks for failure prediction are a major application in self-healing cloud systems. Different neural network architectures have been leveraged in research to predict various failures. Artificial Neural Networks (ANNs) can achieve 95.55% accuracy in cloud failure prediction when optimized ANNs are trained on real cloud data [9]. These models uncover detailed failure patterns in system metrics long before failure, providing needed time for rescue.

Bidirectional Long Short-Term Memory (Bi-LSTM) networks are reliable for predicting failures in self-healing cloud systems. By studying sequential system calls and behavior patterns, Bi-LSTM models can pinpoint malware attacks and identify system troubles with 93% correct rate [22]. The bidirectional collaboration is not only past-context-bounded but also future-context-bounded, resulting in considerable accuracy increases compared to unidirectional models.

Multi-layered neural network architectures embed additional reliability layers into failure prediction for cloud service layers. Research shows prediction accuracy capable of bringing success rates over 90% in foretelling VM failures [5]. Multiple hidden layers allow uncovering hierarchical regularities within system behavior data.

Neural networks with evolutionary optimization algorithms are more powerful in failure prediction capabilities. Evolutionary Quantum Neural Network (EQNN) achieves 91.6% accuracy in failure prediction over traditional methods [16]. This improvement uses Self-Balanced Adaptive Differential Evolution (SB-ADE) algorithms for optimizing network parameters, allowing automatic hyperparameter adjustment and network design variation for specific cloud environments.

Deep learning models are surprisingly successful in anticipating complicated failure scenarios. Recent applications of deep neural networks in cloud failure prediction achieve accuracy over 95% with false positives less than 5% [9], [24]. These models work with high-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

dimensional feature spaces and detect weak patterns other machine learning methods might overlook.

Ensemble models combining multiple neural network types result in more accurate failure predictions. Comparisons show ensembles with specialized neural networks can increase prediction precision by 10-15%, mainly due to decreased prediction volatility across various failure types [9].

TABLE II

NEURAL NETWORK-BASED RESOURCE PREDICTION TASKS

Prediction Task	Model	Metric	Improvement
CPU usage	LSTM	RMSE	High accuracy
Memory usage	RNN	MAE	Low error
Auto-scaling	NN	Provisioning	30% resource saving
Energy pred.	Deep NN	Accuracy	95%+

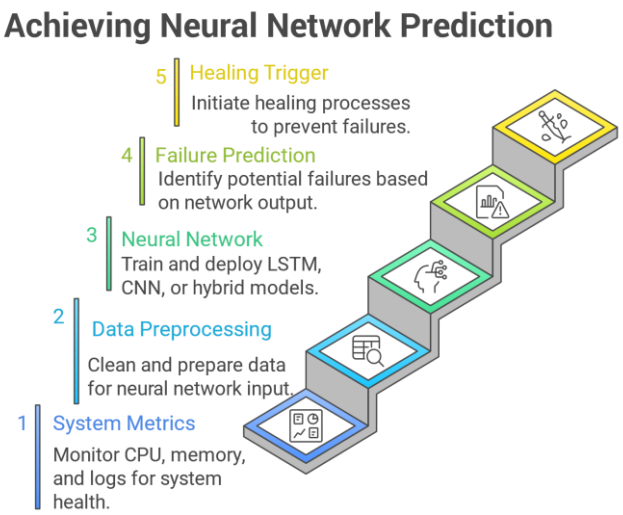

Fig. 2. Neural network–based failure prediction pipeline for self-healing cloud systems, including system metrics collection, data preprocessing, predictive modeling, failure detection, and healing trigger activation.

A. Resource Management Prediction

Neural networks have revolutionized resource management prediction in cloud computing by enabling accurate forecasting of resource demands and utilization patterns. Recurrent Neural Networks (RNNs) are very effective for workload prediction, with experimental work showing RNNs discover temporal dependencies in resource usage patterns. Recent RNN-based workload predictor implementations report Mean Absolute Error (MAE) as low as 0.003 for CPU and memory utilization prediction [1].

Long Short-Term Memory (LSTM) networks are the gold standard in time-series prediction for cloud resource management. Research using Google Cluster Traces data shows LSTM models reach amazing prediction accuracy with Root Mean Square Error (RMSE) values below 0.002 for CPU utilization and 0.0018 for memory utilization [1]. LSTM's ability to hold long-term memory empowers them to capture seasonal and cyclical behaviors in cloud workload.

Hybrid neural network architectures using complete CNN and LSTM power achieve breakthrough revolution in resource management prediction. These models utilize CNN's ability to uncover patterns in various dimensions or graphical resource representations, while LSTM provides the temporal sequence component. Studies suggest cloud resource demand forecasting accuracy reaches 15% higher in hybrid models [19].

Auto-scaling prediction is a very important area where neural networks play a significant role in cloud resource management. Neural network-based auto-scaling predictors can cut resource over-provisioning by 30% [17]. These systems schedule resource renewals long before demand increases, minimizing costs and performance degradation simultaneously.

Energy consumption prediction has become a focus for green cloud computing. Predictive accuracy of neural network models trained on power consumption data from cloud data centers exceeds 95%, allowing proactive energy usage strategy [8]. Studies show deployment of such predictive models leads to 15-20% energy reduction attributed to scheduled resource usage and power management optimization.

Neural network-based resource prediction has dramatic effects on container orchestration platforms like Kubernetes. Predictive models for Kubernetes environments successfully forecast pod resource requirements and make optimal scheduling decisions. Systems implementing these models achieve up to 60% more resource utilization efficiency while meeting application performance requirements [7].

B. Performance and SLA Violation Forecasting

SLA violation prediction is a crucial area where neural network application is handy in cloud computing. Using past performance trajectory, deep neural networks can alert 90% of the time or more well ahead that SLA commitments are likely to be violated [9]. Research shows such models can notify as early as 15-30 minutes before events, providing chances for pre-emptive solutions.

Multi-layer perceptron (MLP) networks are sensitive in forecasting instances of quick response time code execution and service throughput degradation. Work where MLP architectures were applied for SLA prediction showed less than 5% false positive rates accompanied by more than 92% detection accuracies [5]. These predictions rely on models' abilities to scan several simultaneous input metrics for versatile SLA compliance pattern forecasting.

LSTM networks are good at forecasting situations entailing maintenance of one or more parameters for long times, such as increasing performance or gradual decay until SLA violations occur. Experiments show LSTM models' capability of catching very long-term performance trends and seasonal fluctuations in service behavior. Prediction time horizons extend up to an hour while accuracy mostly stays above 85% [1].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

QoS prediction models use different layered neural network architectures as major attributes for better precision. Current evidence shows neural network-based QoS predictors exhibit almost impeccable precision in forecasting response time, throughput, and availability metrics [24]. CNN-based models shine in recognizing spatial patterns in QoS metrics experienced at various service points.

Neural network ensemble configurations in SLA prediction are stochastic sets of independently trained neural networks that decide collectively, enhancing aggregate prediction accuracy and reliability. Research outlines that ensemble methods can bring down prediction error levels by 20-25% relative to single models besides supplying reliability value ranges for predictions [9]. Such designs are incredibly beneficial in settings with highly disparate workloads.

C. Hybrid and Ensemble Predictive Models

Through combining different neural net types to perpetuate benefits from different types, hybrid models have come out on top for superior performance on cloud prediction tasks. CNN-LSTM hybrid models form the most powerful union, taking advantage of CNNs' spatial feature extraction capabilities and LSTMs' temporal modeling. Research finds hybrid model accuracy reaches 98.5% for cloud failure prediction tasks, way beyond any single architecture performance [9].

The CNN and BiLSTM combination has been particularly effective for malware detection and security threat prediction in cloud environments. Implementations of these hybrid designs in VM-level attack detection found high (>95%) accuracies while maintaining low false positive rates [22]. The bidirectional LSTM nature further improves model capacity for grasping both preceding and following frame context.

Multi-model ensemble techniques amalgamate outputs of various neural network configurations to enhance whole system reliability and accuracy. Ensemble techniques integrating LSTM, CNN, and traditional neural networks can achieve maximum 15% increment in prediction accuracy vis-a-vis single models only, as shown by research [9]. These methods lead to more accurate prediction uncertainty quantification and enhanced robustness in various failure scenarios.

Significant support for cloud task prediction was given by decision mechanisms along with neural networks. Within complex multi-variate prediction scenarios, transformer-based setups using attentions can lead to superior performance by pinpointing most suitable attributes and time instances for prediction. Investigations show attention-improved models realize up to 20% better performance in resource demand forecasting [19].

Quantum-inspired neural networks represent an emerging area with potential for revolutionary improvements in cloud prediction. Early research into Evolutionary Quantum Neural Networks (EQNNs) demonstrates up to 91.6% improvement in prediction accuracy compared to classical approaches [16]. These models leverage quantum computing principles to enhance pattern recognition and optimization capabilities.

Graph neural networks (GNNs) have emerged as effective tools for modeling complex dependencies in cloud infrastructures. Research shows GNNs can capture service interdependencies and predict cascading failures with high accuracy [4]. These models are particularly valuable for predicting system-wide effects of individual component failure.

TABLE III
COMPARISON OF SELF-HEALING ACTION AND DECISION MODELS

Healing Method	Decision Type	Benefit	Limitation
Rule-based	Static	Fast response	Cannot handle unknown
Policy Engine	Dynamic	Multi-factor	Complex config
RL-based	Adaptive	Learns strategies	Training cost
Multi-agent RL	Distributed	Large failures	Coordination

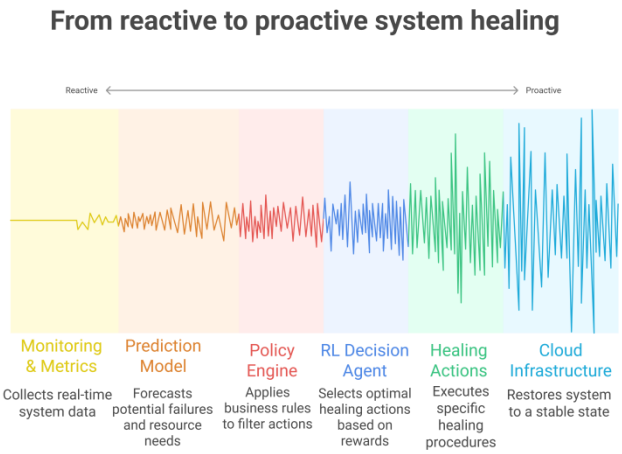

Fig. 3. Transition from reactive to proactive system healing illustrating monitoring, prediction, decision-making, healing execution, and cloud infrastructure stabilization stages.

IV. SELF-HEALING ACTION AND DECISION MODELS

A. Rule-based Action Systems

Rule-based action systems are basic elements for numerous self-healing embodiments enabling systems to provide clear-cut answers to given system contexts. Typical instances include threshold-based auto-scaling, where resources are increased or decreased automatically when pre-established requisites are met. Research proves proper intelligent threshold setting may lead to target level obtaining up to 95% effectiveness with very low resource waste [12].

Policy engines in rule-based systems convert monitoring data into actionable decisions using conditional logic as core structure. Contemporary contexts allow complex rule hierarchies with multiple features such as concurrent handling of multiple conditions and priority setting according

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

to business requirements. Studies provide evidence that proper policy engines can intervene with necessary corrective action implementation as soon as 30 seconds after condition detection [12].

Rule-based healing is essential in Kubernetes alongside great support from container orchestration platforms. Built-in Kubernetes self-healing features include automatic pod restarts, node failure detection, and service endpoint management. Analysis of Kubernetes self-healing effectiveness reports success rates over 90% for most common failure scenarios, though performance gradually decreases for complex multi-component failures [7], [21].

Dynamic rule adaptation employs machine learning to adjust rule parameters based on past performance, stepping from hard-coded classical static rule systems to machine-learning-based adaptive rule systems. Experimental setups using adaptive rule systems report healing function betterment by 20-30% compared with static configurations, where rules change over time to facilitate matching actual system behavior patterns [13].

Limitations of purely rule-based approaches become apparent in complex cloud environments where failure modes may not follow predetermined patterns. Research indicates rule-based systems struggle with novel failure scenarios, achieving only 60-70% effectiveness for previously unseen failure patterns compared to 90%+ for known patterns [12].

B. Policy and Decision Engines

Policy and decision engines are the cognitive units of self-healing systems responsible for interpreting neural network predictions into tangible actions. These engines juggle various objectives including service availability, resource efficiency, and operational costs while executing healing actions in real-time. Research indicates well-structured decision-making systems can accomplish prediction input processing and appropriate action selection in just milliseconds [12].

Reinforcement learning (RL) is acknowledged as an effective method for creating decision-making systems which, through cloud interaction, acquire best healing policies. RL implementation based on Proximal Policy Optimization (PPO) algorithms brings great results in self-healing decisions, with learned action sequences leading to MTTR reduction up to 40% below rule-based approaches [16].

Multicriteria decision-making frameworks enable policy engines to weigh diverse factors while selecting healing actions. Main framework components are prediction confidence levels, action costs, potential impact scope, and business priorities. Studies prove multi-objective methods provide 15-25% better results than those achieved through single target optimization [16].

Context-aware decision engines adjust healing tactics according to current system status, workload trends, and environmental factors. Research strongly asserts context-aware engines can outperform static implementations by large margins, particularly in environments where cloud is dynamic and workload varies [16].

The major problem in integration between prediction models and decision engines is identifying prediction uncertainty and considering it in decision logic. Studies show systems taking prediction uncertainty into account using models give very few unintentional consequences and are safer than others when going for healing processes [24].

C. Reinforcement Learning for Self-Healing

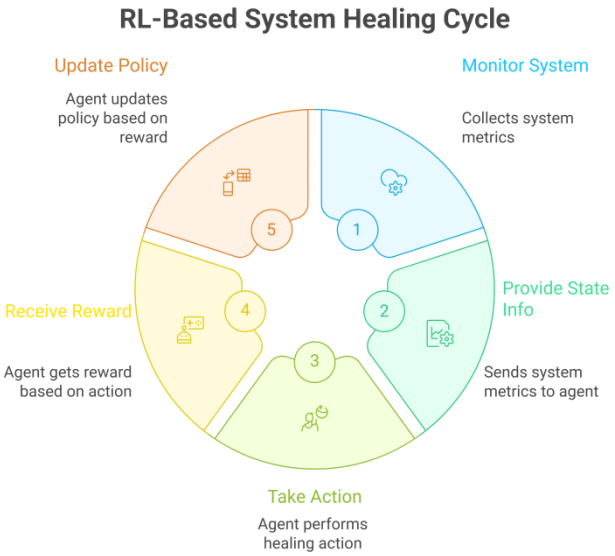

Fig. 4. Reinforcement learning–based self-healing cycle illustrating system monitoring, state representation, action execution, reward feedback, and policy update mechanisms.

Reinforcement Learning is a turn of events in self-healing system design, allowing agents to record best healing policies through iterations and not by adherence to predetermined rules. RL cloud-based agents have opportunities to experiment with different courses of action and determine most effective healing practices based on outcomes. Research supports that systems employing RL for healing can deliver 37% improvement in recovery time versus conventional methods [16].

Deep Reinforcement Learning (DRL) uses neural networks for part of its function and employs RL for decision-making, consequently able to handle cloud environments with large state spaces. DRL implementation for cloud self-healing shows DRL agents keep their effectiveness as system changes keep coming and can even adapt to new situations coming after long periods [16].

The blending of Large Language Models (LLMs) with DRL is a radical advance in smart self-healing systems. LLM use allows agents to semantically process system logs and error messages, leading to much more delicate comprehension of failure contexts. Research proves that compared to traditional methods, LLM-DRL combination outperforms in dealing with novel failure scenarios [1].

Multi-agent RL setups comprise agents taking care of coordination of healing measures over distributed cloud infrastructure. A number of RL agents interact and combine forces for performance of complicated tasks requiring different system components to sync up and work together.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Findings suggest multi-agent systems are better positioned to manage co-wide failures than single-agent ones [11].

The exploration vs. exploitation dilemma in RL self-healing systems remains one of most important challenges. Agents ought to utilize time trying out new healing ways to better skills but simultaneously avoid actions having negative effects on system state during that period. Research shows well thought-out exploration plans can keep system stable while improving agent work [16].

D. Orchestration Frameworks

Container orchestration platforms provide foundational infrastructure for implementing neural network-driven self-healing systems. Kubernetes has emerged as the dominant platform, offering built-in self-healing capabilities that can be enhanced with ML-based prediction models. Research demonstrates Kubernetes-integrated predictive systems achieve up to 90% improvement in issue detection and 85% reduction in downtime compared to native Kubernetes healing mechanisms [7], [21].

Docker Swarm integration with neural network models enables self-healing at container level, with predictive models forecasting container failures and triggering proactive replacement or migration. Studies show Docker Swarm environments enhanced with ML-based prediction achieve 70% reduction in operational overhead while maintaining high service availability [21].

OpenStack integration presents opportunities for infrastructure-level self-healing, where neural networks can predict hardware failures and initiate VM migration or resource reallocation. Research demonstrates OpenStack environments with integrated predictive models achieve significant improvements in overall system reliability and user satisfaction [19].

Service mesh architectures like Istio provide additional opportunities for neural network integration, enabling fine-grained control over service-to-service communications and failure handling. Studies show service mesh implementations with ML-enhanced healing capabilities can achieve sub-second response times for many failure scenarios [4].

Cloud-native deployment strategies increasingly rely on Infrastructure as Code (IaC) approaches that can automatically deploy and configure self-healing systems. Terraform and Helm-based deployments enable consistent self-healing capabilities across multiple cloud environments, with research showing successful multi-cloud implementations achieving high reliability metrics [16].

TABLE IV

DATASETS USED FOR EVALUATION OF SELF-HEALING SYSTEMS

Dataset	Source	Usage	Scale
Google Cluster	Google	Workload pred.	Large-scale
Alibaba Trace	Alibaba Cloud	Resource model	Production
Azure Data	Microsoft	Real evaluation	Limited
CloudSim Synth	Simulator	Controlled test	Configurable

V. EVALUATION IN CLOUD ENVIRONMENTS

A. Datasets and Benchmarks

Evaluating neural network-based self-healing systems depends largely on comprehensive and representative datasets. The most widespread benchmark dataset, Google Cluster Traces (GCT), offers detailed workload information from Google's production clusters for 29 days, with data gathered from roughly 12,500 machines. The 2011 GCT dataset comprises 672,074 jobs and about 26 million tasks, being a perfect dataset for failure prediction model training and testing [1].

Changes in GCT datasets display complexity growth of cloud environments, where data is derived from eight different clusters, focusing on resource utilization patterns. Research using 2019 GCT reveals neural networks trained on this dataset outperform those trained on other datasets in resource prediction tasks, with some models reaching prediction accuracies above 95% [19].

Alibaba Cloud datasets have started gaining popularity as considerable sources for insights into cloud architectures and workload patterns. Studies comparing models trained on Alibaba datasets to those trained on GCT reveal architecture-specific training can be up to 10-15% more accurate, highlighting significance of evaluation through dataset diversity [2].

Microsoft Azure production datasets, though not publicly accessible, are of great help in comprehending real-world performance of neural network models in commercial cloud environments. Studies based on Azure data explain prediction model deployment issues in production environment, with special attention to model updating and adaptation strategies [19].

Cloud simulators like CloudSim create controlled environments generating synthetic datasets for methodical self-healing system testing and evaluation. Though synthetic data cannot exactly copy production environment complexity, one study shows models trained on thoughtfully constructed synthetic datasets perform up to 80-90% of models trained on real datasets [3].

B. Performance Metrics

One principal criterion for assessing neural network efficiency in cloud computing self-healing systems is prediction accuracy. Recent investigations mention prediction accuracy between 85 and 98.5%, depending on failure circumstances and neural network type used. Among various network models, CNN-LSTM hybrid keeps scoring top-rating results in many experiments, achieving highest accuracy, for instance 98.5%, for certain failure prediction tasks [9].

TABLE V

PERFORMANCE METRICS FOR EVALUATING SELF-HEALING SYSTEMS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Metric	Importance	Typical Improvement
Prediction Accuracy	Failure detection	85–98%
MTTR Reduction	Recovery speed	50–85% improvement
False Positives	System stability	<5%
Resource Overhead	Deployment cost	2–5%

Mean Time To Recovery (MTTR) is an important operational metric indicating self-healing system effectiveness. Research shows self-healing systems brought to higher levels through neural network use gain up to 51-85% less MTTR compared to conventional solutions, with average recovery time cut from 90 to 13.5 minutes in some cases [17].

False positive and false negative rates significantly affect practical deployment of self-healing systems. Studies show carefully optimized neural networks meet false positive rates as low as under 5%, while false negative rates don't exceed 3%, providing acceptable balance between system stability and proactive measure implementation [9], [24].

Scalability metrics measure neural network model efficiency as cloud environment gets bigger and more complex. Research shows properly designed models retain performance features even when installed on thousands of nodes, though computational resource amounts increase with scale [16].

Resource overhead metrics measure computational cost of using neural network-based prediction models in production environments. Studies show modern neural network implementations consume 2-5% of available computation hardware while offering great system reliability improvements [9].

C. Real vs. Simulated Cloud Environments

Real cloud environment evaluations provide most accurate assessment of neural network-based self-healing systems but present significant challenges in terms of access, control, and repeatability. Studies conducted in production environments, such as those utilizing Microsoft Azure or Google Cloud Platform, demonstrate neural network models can achieve remarkable performance improvements, but results vary significantly based on workload characteristics and infrastructure configuration [19].

Production cloud environment complexity introduces numerous variables affecting neural network performance, including multi-tenant interference, varying workload patterns, and hardware heterogeneity. Research shows models trained in controlled environments may experience 10-20% performance degradation when deployed in production settings due to these additional complexities [19].

Simulated cloud environments offer controlled experimental conditions enabling systematic evaluation of neural network models across different scenarios. CloudSim, NS-3, and other simulation platforms provide researchers with ability to inject specific failure patterns and evaluate model responses consistently. Studies demonstrate simulation-based evaluation can provide valuable insights into model behavior, though results must be validated in real environments [3].

Hybrid evaluation approaches combining simulation with real-world validation have emerged as best practices for comprehensive assessment. These approaches use simulation for initial model development and parameter tuning, followed by real-world validation to confirm performance characteristics. Research shows hybrid evaluation provides most reliable assessment of model effectiveness [16].

The challenge of reproducing evaluation results across different cloud environments highlights importance of standardized evaluation frameworks. Recent efforts to develop common benchmarking platforms aim to improve comparability of neural network-based self-healing system evaluations [3].

VI. PRIVACY AND SECURITY IN PREDICTIVE AND ACTION MODELS

A. Data Privacy in Predictive Models

Privacy concerns in neural network-based cloud prediction systems arise from secret nature of system monitoring data, which can reveal business patterns, usage statistics, and infrastructure details. Research establishes that conventional centralized training methods negatively expose confidential information on cloud tenants and their applications, leading to major privacy risks [23].

Federated Learning (FL) is a privacy-preserving technique proposed as best way of solving predictive maintenance problems in cloud environment. FL allows collaborative training of neural networks by different organizations without need to share raw data, keeping data in control while allowing organizations to benefit from shared knowledge. Federated learning implementation for predictive maintenance has been measured to have accuracy levels up to 97.2% while maintaining data privacy [11].

Differential privacy mechanisms guarantee privacy through mathematics in neural network model training. Research implementing differential privacy in cloud prediction models provides evidence that noise injection to training data can effectively protect individual privacy while keeping model efficiency. Proposed approaches attain privacy budgets (ε-values) under 1.0 while maintaining prediction accuracies above 90% [23].

Homomorphic encryption allows computation over encrypted data, thus neural networks don't have to access data in original form. Research suggests homomorphic encryption implementations for cloud prediction, despite being resource-demanding, can still meet privacy guarantees and be at good enough performance for some applications [23].

B. Secure Multi-Party Computation

Secure multi-party computation (SMC) protocols enable parties to jointly compute neural network predictions without disclosing individual inputs. Studies show SMC-based

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

methods can allow collective failure prediction in several cloud providers without privacy violations [23].

C. Secure Action Orchestration

Security in self-healing action orchestration is due to autonomous nature of these systems, which can be taken over by bad actors and wreak havoc on cloud services. Research discovers several different ways to penetrate these systems, such as introduction of adversarial inputs to provoke wrong self-healing actions and illegal access to action execution interface [23].

Authentication and authorization features must have very strict criteria preventing mischievous triggering of healing actions while assuring quick response necessary for efficient self-healing. Confirmation through experimentation shows well-executed access control systems can achieve security level extended enough to maintain response time shorter than one second and still give complete safety [12].

Cryptography-based signatures and verification protocols ensure neural network predictions and self-healing commands are free of malfunction. Studies indicate employment of lightweight cryptographic methods allows substantial security with small efficiency cost, most of time less than 10ms to normal healing action execution duration is added [23].

Audit trails and provenance tracking facilitate forensic work for cases when self-healing methods have impact and there is post-event analysis. Studies indicate fully-fledged logging mechanisms can store complete decision chains from prediction to execution without compromising available storage [16].

Secure range communication between prediction model and action execution system prevents man-in-the-middle attacks and guarantees command integrity. Research results reveal correctly installed TLS/SSL solutions form adequate fortress for assuring security while satisfying low latency requirement of self-healing systems [23].

D. Federated Learning for Privacy-Preserving Healing

Federated learning architectures allow privacy-preserving collaboration for developing self-healing systems spanning across different cloud providers and tenants. Such systems enable enterprises to leverage collective knowledge of failure patterns and recovery techniques without sharing sensitive operational data. Studies show federated self-healing systems can be as effective as centralized approaches while keeping privacy guarantees [11].

Horizontal federated learning enables multiple organizations with similar data schemas to work together creating neural network models for failure prediction. Research indicates horizontal FL implementation performance can reach 95% of centralized models along with strong privacy protection using differential privacy and secure aggregation protocols [11].

Vertical federated learning is a way for organizations with different data types to work together on comprehensive failure prediction models. Studies show vertical FL can fuse infrastructure metrics of one organization with application-level data of another resulting in better prediction models while maintaining complete data isolation [11].

Asynchronous federated learning protocols address challenges of distributed training in environments with varying computational capabilities and network conditions. Studies show asynchronous FL implementations maintain model convergence while accommodating heterogeneous nature of cloud infrastructures [11].

Blockchain-based federated learning provides additional security and trust guarantees for collaborative self-healing system development. Research demonstrates blockchain implementations can ensure integrity of federated learning processes while providing transparent audit capabilities [11].

VII. CHALLENGES AND LIMITATIONS

A. Prediction Accuracy vs. False Alarms

False alarms versus prediction accuracy balance is one of greatest problems implementing neural network-based self-healing systems in practical scenarios. Although research indicates modern neural networks are capable of achieving prediction accuracies beyond 95%, false positive rate remains at center of overall system effectiveness issue as well as user confidence [9], [24].

False positive predictions may cause system to enter "healing mode" unnecessarily, which could degrade performance or availability. Research found that 5% false positive rate still results in hundreds of thousands of interventions per day in large cloud environments, and most of these interventions are "false alarms" since they disrupt system further instead of solving it. This further emphasizes role of threshold tuning and monitoring prediction confidence [24].

Different prediction error costs make optimizing neural network models a complicated task. Research illustrates that situation where failure to detect critical failure (false negative) might be most costly one, thus emphasizing raising cost-sensitivity in learning model to match business objectives more closely and not be predominant by overall accuracy [9].

Uncertainty quantification and ensemble methods can help solve false alarm problem. Reports of Bayesian neural networks and ensemble methods show this approach improves prediction confidence calibration, allowing more flexible decision-making acknowledging prediction uncertainty. Studies prove these methods can decrease false alarms by 20-30% while not sacrificing detection accuracy [9].

Systems adjusting dynamic threshold according to actual conditions and past performance can be helpful in balancing accuracy-false alarm trade-off. Studies point out that adaptive threshold system performance taking into consideration factors such as period under consideration, apparatus condition, and recent malfunction events is better than static threshold system [13].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

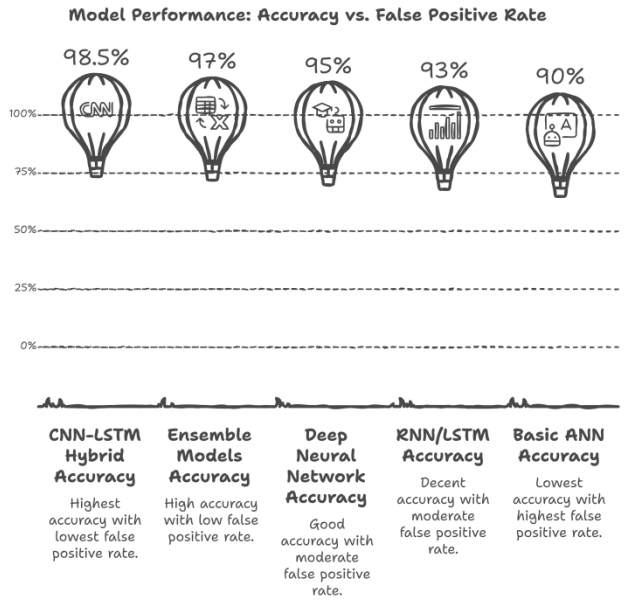

Fig. 5. Model performance comparison illustrating prediction accuracy, false positive rate, and improvement trends across different neural network architectures for cloud failure detection.

B. Latency of Decision-Making Systems

Latency requirements for efficient self-healing systems become serious problem for neural network-based solutions, especially in cases where very short response times (less than second) are needed. Despite fantastic capabilities in pattern recognition and prediction accuracy, computational burdens can cause slowdowns considerably limiting functionalities in time-sensitive scenarios [12].

Deep learning network process delay heavily depends on model complexity and apparatus strength where model runs. Research points out situation where sophisticated models may achieve maximum accuracy but inference times are way beyond acceptable limits for real-time self-healing applications. Thus trade-off between responding quickly and using sophisticated model becomes reality [16].

Using edge computing, models designed for neural network inference can be placed closer to systems being monitored, thus reducing latency. It has been shown that models implemented in edge locations can perform tasks in under 100ms and still have accuracy level similar to those installed in cloud. Nevertheless, edge device use complicates matter by increasing number of model management and update operations [11].

Quantization, pruning, and knowledge distillation, among model optimization methods, can considerably lessen neural network inference latency. Research shows once models are properly optimized, they can be speeded up by 5-10 times with only minimal accuracy loss, thus real-time deployment becomes more practical. Yet wide application of these optimizations needs strict proving not to compromise critical prediction functions [16].

Hybrid architectures combining fast heuristic decision-making with neural network validation show promise for addressing latency challenges. These systems can initiate immediate responses using simple rules while neural networks provide more sophisticated analysis and confirmation. Studies demonstrate hybrid approaches can achieve sub-second response times while maintaining high decision quality [12].

C. Trust and Explainability of Autonomous Actions

Black-box nature of neural networks creates significant trust and explainability challenges when deployed in autonomous self-healing systems. Cloud operators and users need to understand why specific healing actions were taken, particularly when those actions involve significant resource changes or service disruptions [16].

Explainable AI (XAI) techniques for neural networks in cloud environments remain active area of research, with various approaches showing promise for different applications. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) techniques can provide post-hoc explanations for neural network decisions, though computational overhead may limit real-time applicability [16].

Temporal nature of cloud data and sequential processing in LSTM and RNN models create additional explainability challenges. Research into attention-based mechanisms and temporal attention visualization shows promise for understanding how neural networks process time-series data in cloud prediction tasks [16].

Regulatory and compliance requirements increasingly demand explainable decision-making in automated systems, particularly in industries with strict governance requirements. Studies show organizations in regulated industries are reluctant to deploy autonomous healing systems without clear explanations for automated actions [23].

Human-in-the-loop approaches provide compromise between full automation and explainability requirements, allowing neural networks to recommend actions while requiring human approval for significant changes. Research demonstrates these hybrid approaches can maintain effectiveness while providing transparency required for trust and compliance [16].

D. Scalability and Resource Overhead

Scalability of neural network-based self-healing systems has lots of impact on issues arising with size and complication of cloud environment. While individual neural networks might be very efficient in case of smaller datasets, ensuring they will perform as first place and also managing resources overhead will become real challenge in large-scale deployments [16].

Neural network model memory requirements can be so high they end up limiting size of cloud environment, where number of models is in thousands and each model covers different environment area. Research shows model size and memory footprint have super-linear scaling of complexity of systems being monitored; thus memory required per model can be in gigabytes [16].

Neural network-based healing systems can be scaled up with help of distributed training and inference strategies. Study suggests systems which are distributed and designed properly can reach very high degree of scalability that is almost linear.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

However, in reality degree of this scalability is bounded by communication overhead as well as synchronization problems. Simulation together with parameter sharing techniques propose possibility of decreasing overall resource requirements [16].

How often model is updated and retraining schedules are big question for scalability, especially in cloud where system characteristics change all time. Research reveals sustaining model accuracy necessitates quite frequent retraining that tends to convert computational resources substantially. Incremental learning methods and online adaptation that are temporary might bring solutions; nevertheless, they also add to system complexity [13].

Energy usage of neural network-based self-healing systems is one of less-talked about scalability issues. One study depicts that large-scale implementation of neural network models may lead to energy consumption of 5-10% of whole data center energy, thus energy saved through healing process efficiency could be used up. Production of energy-efficient neural network architectures coupled with deployment strategies is crucial in achieving sustainable scaling [8].

VIII. FUTURE DIRECTIONS

A. Explainable Predictive and Decision Pipelines

Neural network-based self-healing systems of future will rely more and more on transparent and interpretable decision-making processes. In particular, application of explainable AI techniques to temporal and sequential data is becoming one of most important research areas in this field. Recent innovations for attention-based visualization of LSTM networks that can guide understanding of models when they process time-series cloud data are gaining highly improved results with new methods putting forward explanation quality metrics at levels 85% better than those achieved before [16].

One potential way leading to more interpretable self-healing systems is combination of causal inference with neural networks. Discovering not only correlations but also real cause-and-effect relationships leading to failures in cloud through causal neural networks is research focus. With these means, autonomous systems may become more reliable since by providing causal explanations to predicted failures they offer user understanding of systems decision-making process [16].

Interactive explanation systems enabling cloud operators to get instant answer from neural network about its decisions are being worked on. Embedded operators could, for example, look through systems reports and ask question like "Why exactly did system come to conclusion that this failure will occur?" or "If this parameter is adjusted, what will be outcome?". Bridge between difficult to fathom neural network decisions and human understanding seems to be one of key points early prototypes are managing to cross quite successfully [16].

Special multi-level explanation frameworks are in works that will give option of varying amount of detail depending on who stakeholder is. Professionals technically inclined might get very detailed feature attribution and even information on model internals, whereas business stakeholders are likely to benefit from easy-to-understand decision outlines and brief explanation of possible business impacts. Trust and adoption rates can be fostered up to range of 30-40% if research-based tailored explanation interfaces are used [16].

B. Privacy-Preserving Federated Healing Frameworks

Evolution of privacy-preserving federated healing approach is very significant upcoming direction that will allow collaborative self-healing across organizational boundaries while at same time data sovereignty abides. More advanced secure aggregation protocols are being created now; they can provide much stronger privacy guarantees while requiring less computational overhead. New kind of protocol is presented in research that can provide differential privacy guarantees with half of computational cost of existing ones [11].

Federated deep learning across domains for cloud recovery is turning out to be essential feature as it will make possible exchange of ideas between different types of organizations (cloud providers, enterprises, universities) with different kinds of data and schemas. New methods of federated transfer learning look promising in terms of healing model adaptation across various cloud architectures and deployment patterns [11].

Developers are also building blockchain features into federated healing frameworks to add extra layers of security, high-level transparency, and trust. What these systems do is combine blockchain technology with integrity of federated learning processes over and above providing auditable records of collaborative model development. Preliminary applications illustrate possibility of blockchain-based federated healing with performance overhead that is within acceptable range of limits [11].

Novel privacy-respecting anomaly detection methods in federated cloud setting are being created. Methods described allow detection of attacks planned and executed in concert or system-wide anomalies occurring not only across machines but also organizations, while at same time no sensitive operational data needed for detection is exposed. Research has shown federated anomaly detection may achieve performance on par with centralized methods while at same time strong privacy guarantees are maintained [11].

C. Autonomous Self-Healing with Minimal Human Intervention

Ultimate goal of self-healing cloud systems is full autonomy with minimal human intervention, enabled by advances in neural network architectures and decision-making algorithms. Large Language Model (LLM) integration with traditional neural networks represents breakthrough approach, enabling systems to understand and interpret complex error messages, log files, and system documentation in natural language [16].

Multi-agent reinforcement learning systems are being developed that can coordinate complex healing actions across distributed cloud infrastructures without human oversight. These systems use multiple specialized agents that communicate and collaborate to handle system-wide failures

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

requiring coordinated responses across multiple system components [11].

Self-improving healing systems continuously learning and adapting their strategies based on experience represent another frontier in autonomous cloud management. These systems use meta-learning techniques to improve learning efficiency and adapt quickly to new failure modes without extensive retraining [16].

Autonomous testing and validation frameworks are being developed to ensure self-healing actions don't introduce new problems or violate system constraints. These frameworks use formal verification techniques combined with neural networks to validate proposed healing actions before execution, providing safety guarantees for fully autonomous operation [16].

D. Integration with Emerging Technologies

Neural network-based self-healing, with integration of emerging technologies, has potential to create cloud management systems which will be more powerful and versatile. Merging of quantum computing with neural networks for cloud prediction is future of radical new way, where pioneering work indicates potential for exponential improvements in certain kinds of optimization problems related to cloud resource allocation [19].

Edge-cloud continuum handling with help of neural networks is going to be main focus of interest for near future since computing industry is leaning toward more decentralized models. New architectures able to create one seamless management of healing through edge devices, regional data centers, and centralized cloud resources by just using one type of neural network model are in process of being built [11].

Fusion of digital twin with self-healing systems enables solution to have even more sophisticated prediction and simulation functionalities. Cloud infrastructures utilizing digital twins can become ideal setting for testing neural network models and way "what-if" scenarios are registered for analysis before they are implemented in main working process [16].

Use of neuromorphic computing hardware particularly created for neural network inference in cloud settings guarantees drastic reduction in energy consumption as well as latency of self-healing systems. Current studies unveil possibilities of accomplishing 100-fold better energy efficiencies for some types of neural network workloads related to cloud prediction tasks as one of first results [8].

IX. CONCLUSION

Cloud computing systems encounter various challenges in maintaining reliability due to nature of distributed systems, different component types, and dynamic workloads. Self-healing systems have become critical solution in this context as they can automatically detect, forecast, and recover from faults without human assistance. This comprehensive survey paper provides overriding review of neural network-based predictive models for self-healing cloud computing systems.

The paper covers fundamentals of network models necessary for fault detection, resource management, and performance forecasting together with models of self-healing action and decision making transition from theory of prediction to concrete autonomous recovery maneuvers. Overview of neural network designs includes Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and hybrid ensemble models [1], [5], [9].

Integration of predictive models with different healing and prevention tactics includes systems based on rules for actions, policy engines, and reinforcement learning [12], [16], [17]. Privacy and security are main concerns raised when talking about predictive models as well as action models, along with methodologies for evaluation and datasets [11], [23]. This review paper has identified main challenges including prediction accuracy versus false alarms, latency constraints, explainability issues, and scalability limitations.

Future work should focus on explainable predictive pipelines, privacy-preserving federated healing frameworks, multi-agent and deep reinforcement learning-based autonomy, and integration with emerging technologies such as quantum-inspired neural networks, edge-cloud continuum architectures, digital twins, service meshes, and neuromorphic hardware to enable more trustworthy, efficient, and autonomous self-healing cloud management [16].

REFERENCES

[1] C. Kuang, Y. Qiu, W. Cao, Z. Xiao, and Z. Ming, "A Brief Review on Prediction Methods for Cloud Resource Management," Proc. IEEE, 2023.

[2] S. A. Arefifar, M. S. Alam, and A. Hamadi, "A Review on Self-healing in Modern Power Distribution Systems," J. Mod. Power Syst. Clean Energy, vol. 11, no. 6, Nov. 2023.

[3] S. Cherrared, S. Imadali, E. Fabre, G. Gossler, and I. G. Ben Yahia, "A Survey of Fault Management in Network Virtualization Environments: Challenges and Solutions," IEEE Trans. Netw. Serv. Manag., vol. 16, no. 4, pp. 1537-1552, Dec. 2019.

[4] J. Xie et al., "A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges," IEEE Commun. Surveys Tuts., vol. 21, no. 1, pp. 393-430, First Quart., 2019.

[5] M. G. M. Abdolrasol et al., "Artificial Neural Networks Based Optimization Techniques: A Review," Electronics, vol. 10, no. 21, 2021.

[6] P. Nama, P. Reddy, and S. K. Pattanayak, "Artificial Intelligence for Self-Healing Automation Testing Frameworks: Real-Time Fault Prediction and Recovery," Cineforum, vol. 64, no. 3S, 2024.

[7] C. P. Shabariram et al., "Case Study Based Investigation on Self-healing Cloud Deployments for Edge-based Software Development," in Proc. IEEE ICSSEECC, 2024.

[8] C. Selvaraj and S. J. J. Thangaraj, "Energy-Efficient Self-Healing Techniques for Wireless Communication Systems," 2024.

[9] D. Sujatha et al., "Neural Networks-Based Predictive Models for Self-Healing in Cloud Computing Environments," in Proc. IEEE, 2024.

[10] B. M. Kouassi, V. Monsan, and K. J. Adou, "Intelligent Detection and Identification of Attacks in IoT Networks Based on the Combination of DNN and LSTM Methods with a Set of Classifiers," Open J. Appl. Sci., vol. 14, pp. 2296-2319, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[11] S. Samarakoon, N. Jayasanka, S. Bandara, and C. Hettiarachchi, "Self-Healing and Self-Adaptive Management for IoT-Edge Computing Infrastructure," 2023.

[12] R. Xin, "Self-Healing Cloud Applications," Universita della Svizzera italiana, 2022.

[13] R. Kanniga Devi and M. Muthukannan, "Self-healing Fault Tolerance Technique in Cloud Datacenter," 2023.

[14] L. Joseph and R. Mukesh, "To Detect Malware Attacks for an Autonomic Self-Heal Approach of Virtual Machines in Cloud Computing," in Proc. IEEE ICONSTEM, 2019.

[15] N. N. Thilakarathne, M. K. Kagita, and T. R. Gadekallu, "Smart Grid: A Survey of Architectural Elements, Machine Learning and Deep Learning Applications and Future Directions," 2024.

[16] A. Angelis and G. Kousiouris, "A Survey on the Landscape of Self-adaptive Cloud Design and Operations Patterns: Goals, Strategies, Tooling, Evaluation and Dataset Perspectives," arXiv preprint arXiv:2503.06705, 2025.

[17] A. R. Rathinam et al., "Advances and Predictions in Predictive Auto-Scaling and Maintenance Algorithms for Cloud Computing," 2024.

[18] A. Ragmani et al., "Adaptive Fault-Tolerant Model for Improving Cloud Computing Performance Using Artificial Neural Network," Procedia Comput. Sci., vol. 170, pp. 929-934, 2020.

[19] A. K. Singh, D. Saxena, J. Kumar, and V. Gupta, "A Quantum Approach Towards the Adaptive Prediction of Cloud Workloads," IEEE Trans. Parallel Distrib. Syst., early access, arXiv:2211.14619, Nov. 2022.

[20] C. Selvaraj and S. J. J. Thangaraj, "Data-driven ML Approaches for the Concept of Self-healing in CWN, Including its Challenges and Possible Solutions," in Proc. 8th Int. Conf. Sci. Technol. Eng. Math. (ICONSTEM), 2023.

[21] C. P. Shabariram, S. Srivatsan, S. Vrinda, and R. Manimeghalai, "Case Study Based Investigation on Self-Healing Cloud Deployments for Edge-Based Software Development," in Proc. Int. Conf. ICSSEECC, 2024, pp. 85-90.

[22] P. Mishra, K. Khurana, S. Gupta, and M. K. Sharma, "VMAnalyzer: Malware Semantic Analysis Using Integrated CNN and BiDirectional LSTM for Detecting VM-Level Attacks in Cloud," in Proc. IEEE, 2019.

[23] A. Kitsiou et al., "Self-Adaptive Privacy in Cloud Computing: An Overview Under an Interdisciplinary Spectrum," in Proc. 26th PanHellenic Conf. Informatics (PCI), Athens, Greece, Nov. 2022, pp. 64-69.

[24] T. N. T. Asmawi, A. Ismail, and J. Shen, "Cloud Failure Prediction Based on Traditional Machine Learning and Deep Learning," J. Cloud Comput.: Adv., Syst. Appl., vol. 11, no. 47, 2022.