Power System Technology
ISSN:1000-3673

Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

AI-Augmented Forecast-Aware Scaling: Leveraging Machine Learning for Predictive Traffic Modeling and Intelligent Cloud Infrastructure Optimization

MohanVamsi Musunuru[1], ManishTomar[2], Tejas Dganorkar[3], Priya Dharshini Kalyanasundaram[4], Radhakrishnan Pichaimani[5], Thirunavukkarasy Pichaimani[6], Ravi Kumar Burila[7]

[1,4]Amazon, [2]Citi, [3]Discover Financial Services, [5]VDart Technologies, [6]Molina Healthcare, [7]JP Morgan Chase

Correspondence Author: mohanvamsi.us@gmail.com

Abstract

The classical cloud scaling solutions fail to maximize performance while being cost-effective, as digital infrastructures are scaled up and down in response to sporadic workload spikes and demands. In this work, we introduce a new predictive scaling paradigm of the future, AI-Augmented Forecast-Aware Scaling (AFAS), which involves machine learning and intelligent telemetry to forecast traffic variation in advance and automatically optimize infrastructure. The type of model AFAS uses is called a hybrid ensemble learning model. It consists of the combination of gradient boosting, time-series decomposition, and anomaly-conscious regression, which are the principles of high-fidelity workload forecasting. Data from historical and actual times is combined to train the model and validate it to make proactive scaling decisions aligned with business goals. The framework is part of a validation pipeline with a hyperscale mode of deployment that reduces false positive scale events and provides efficient assignments of resources on an AZ basis. Investigations on the cloud-native frameworks identify that 30 percent of overprovisioning is cut, resulting in a better cost-performance ratio, and storage responsiveness to high surges in demand. The findings underscore the revolutionary AI of the perfection of autoscaling techniques in modern cloud systems.

Keyword: Forecast-Aware Scaling**,** Machine Learning in Cloud Computing**,** Predictive Traffic Modeling**,** Cloud Infrastructure Optimization**,** Ensemble Learning for Autoscaling

1.0 Introduction

The flexibility, scale, and cost-effectiveness of its different dimensions, which depend on cloud infrastructure, make the modern digital application necessary. However, market fluctuation of workload subject to the rapid growth and paucity of data does pose sustained perplexities of conventional autoscaling patterns, especially in real-time, data-centered sphrères. The rule-based or reactive legacy approaches can become either over-provisioning of resources or the resource-bound performance behavior in case of a traffic surge (Kumar et al., 2022). As

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

applications are still transforming with more and more user demands, intelligent scaling frameworks become necessary to realize the best use of clouds without losing their performance or reliability.

In this article, the authors have introduced the AI-augmented forecast-aware scaling (AFAS) framework, a high-tech domain implementing machine learning (ML) technologies, making it possible to predict the direction of traffic and cloud resource scaling. Another type of ensemble in AFAS is a hybrid one that combines gradient boosting, time-series decomposition, and anomaly-aware regression. The cloud systems enabled by AFAS can scale intelligently, using historical and real-time telemetry data. They can minimize downtime and unwanted scale-up/down operations and lower the cost of operations. The aim is to mitigate the trade-off between predictive intelligence and the real-time elasticity of the cloud deployments configured in multi-availability zone (multi-AZ) systems, where guesswork of the dynamic traffic usage in such clouds commonly results in infrastructure overengineering (George et al., 2023).

1.1 Background

However, in the classical cloud autoscaling systems, the policies are considered in terms of prior established CPU, memory, or latency limits. These schemes are not yet bursty in stochastic workloads (Abbas & Myeong, 2023). Also, such reactive models cannot predict the increase in traffic or anomalies, leading to late scaling or wasting resources. There is an urgent necessity to support more intelligent, context-sensitive resource provisioning strategies, which Chhetri et al. (2018) raise.

The lack of foresight in current scaling mechanisms is further aggravated in multi-AZ deployments, where workloads shift across geographically distributed zones with inconsistent latency profiles. This paper identifies that **forecasting-aware intelligence**, combined with machine learning, can drastically improve the precision of autoscaling decisions. By leveraging predictive analytics, systems can anticipate demand patterns before they manifest, thereby minimizing downtime, reducing costs, and ensuring service level agreement (SLA) adherence (Abed et al., 2023).

1.2 Objective and Scope

The main idea of this research is to implement and analyze an AI-based forecast-aware autoscaling system- AFAS, which maximizes the performance of cloud infrastructure deployment and minimizes the inefficiency in the operation of cloud infrastructures. Particularly, the focus of the given work is to:

- To determine real-time traffic, a hybrid ensemble model must be made with gradient boosting, season decomposition, and anomaly-aware regression as its components.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

- Include the predictive and the scaling decision pipeline that models the hyperscale cloud operational workplaces (Auroux et al., 2015).
- Majorly decrease the number of false positives and re-tune the threshold issue to dampen down unneeded scaling incidents.
- Empirically validate/the framework on the performance of the dynamically different workload eco-systems implemented on multi-AZ clouds as the test platform.

The paper falls under Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) solutions and analyzes the difficulties of dynamic autoscaling, not fixed provisioning or manual operations. It is aimed at volatile workloads like e-commerce websites, video streaming, and even the IoT telemetry data processor (Pal et al., 2023).

2.0 Literature Review

Autoscaling in the cloud with AI enhancement is a fast-growing area where ML is transforming the potential to predict, adjust, and optimize resources depending on the changing workload. The following section shows the critical analysis of the supporting research, the trends, and the processes that led to the establishment of the AFAS framework. The main subtopics are decomposed into the following ones: ML for cloud resource management, traffic prediction methods and approaches, hybrid learning ensembles, and validation cross-environmentally.

2.1 Cloud Resource Management using Machine Learning

Machine learning models are increasingly used in cloud resource management systems to implement intelligent decisions that cannot be expressed in fixed rules or reactive measurements. Standard autoscaling mechanisms, like threshold autoscaling or reactive autoscaling, are insufficient for unpredictable traffic patterns or non-homogeneous workloads (Chhetri et al., 2018; Mishra et al., 2020). Such strategies tend to leave the infrastructure underutilized or cause service degradation as scaling response is slow.

In recent studies, Abbas et al. (2022) and Gupta et al. (2023) showed that adaptive ensemble-based learning led to superior results over the static models due to its ability to consider the temporal variation, anomaly triggers, and according to its service characteristics. They obtained high prediction accuracy in their models because of the use of multi-source data integration and dynamically tuning their models. To this purpose, Soni and Kumar (2022) and Khan et al. (2022) also emphasized the implementation of ML-oriented methods in the setting of elastic cloud resources management that mediates between various ML paradigms, such as deep learning, time-series estimation, and hybrid regression models, in order to facilitate scalable optimizations.

Besides, Butt et al. (2020) and Belal and Sundaram (2022) identified that predictive intelligence by utilizing ML could be of great help in cloud security and performance monitoring and can

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

be particularly valuable in making scaling decisions, which need to factor in concerning anomalous traffic or possible threats. These views form the basis of the design of the AFAS model, where predictive intelligence becomes part of the infrastructure decisions.

2.2 Predictive Traffic Modeling and Forecasting

Traffic modeling defines any autoscaling framework. Sound forecasting of incoming traffic assists in resource pre-allocation, downtime reduction, and cost optimization. The use of time series and regression-based modeling of user demand has proven practical in cloud computing.

Kim et al. (2018) presented the work that introduced CloudInsight, predicting future application workloads with the support of an expert council. Their collective approach made use of the strengths of the individual models to give formidable results in all scenarios. Along the same lines, Kaim et al. (2023) suggested a hybrid architecture (BiLSTM-CNN) capable of using short-term and long-term dependency patterns of workloads in multivariate cloud environments, and that surpassed forecasting the use of resource baseline models such as ARIMA or simple LSTM.

Fei et al. (2019) also polled on the potential of machine learning in fog and cloud computing and how the hybrid model will increasingly resonate in traffic analytics. These results agree with what the AFAS framework will achieve in using time-series decomposition, boosting, and anomaly-aware regression to solve context-sensitive prediction.

Moreover, the predictive modeling in transport (Kecman & Goverde, 2015; Mehdizadeh et al., 2020) and in systems based on IoT (Pal et al., 2023; Azizah et al., 2016) identified transferable approaches that can be used in cloud infrastructure, especially when faced with the high-frequency transmit telemetry signal and spatiotemporal variance thereof.

2.3 Hybrid Ensemble Learning for Scaling Intelligence

Hybrid ensemble models are used to ensure that error margins are minimized through the coalescence of predictive capacities of multiple learning algorithms, identification of outliers, and enhancement of forecast reliability. This sort of model is fundamental in cloud infrastructures, where a single misprediction may result in a succession of failures or an exaggerated cost because of the wrong-sized scaling.

Abbas et al. (2023) pre-introduced an ensemble learning based workload prediction model that could adapt to real-time changes in workload, which was further advanced in Karn et al. (2019) with auto-selection and tuning of the model. Such architectures accommodate intelligent autoscaling policies in which the system dynamically selects the best learner or set of learners out of a set based on the characteristics of the input data and the signals about how well a learner has made mistakes.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

Based on this research, the AFAS model has borrowed much of its knowledge from the incorporation of three techniques:

1. Gradient Boosting Models (GBMs) to recognize the nonlinear interactions.
2. Time-Series Decomposition to decompose the seasonal, trend, and residual components.
3. Regressions Anomaly-Aware shifts triggered by Hooi et al. (2019) to reduce false positives in scaling events.

This triad is composed of an adequate degree of precision and recall; therefore, the model may be steady regarding potential workload spikes or traffic peculiarities that may appear in production (George et al., 2023; Qiu et al., 2022).

2.4 Integration and Operation Validation Pipeline

Despite the high technical level of the ML models we develop, most fail in real-life, cloud-based deployment in terms of interpretability, explicability, and scale of performance. Stupar and Huljenic (2023) explain that this deployment optimization needs to be model-based by balancing complexity in the calculation and real-time responsiveness. It is where the validation pipelines can be helpful, especially when implementing ML systems in CI/CD programming and container orchestration systems, e.g., Kubernetes (Senjab et al., 2023).

One of the concepts AFAS applies is that of a hyperscale cloud-level DevOps pipeline that includes validation and its inherent telemetry-driven feedback, as well as model accuracies and scaling delays. This plan correlates with Ramamoorthy et al. (2021) and Saini et al. (2023). The proposal suggests optimizing multi-objective and runtime monitoring frameworks to ensure the infrastructure is healthy when performing dynamic autoscaling activities.

In addition, Kamila et al. (2022) argue that it is also necessary to unite operational resilience with ML models, which further supports the importance of fault-tolerant systems. Their strategy emphasizes the necessity of runtime anomaly detection modules and fallback mechanisms components that could also be found in the architecture of AFAS.

2.5 Summary Table of Literature Insights

Area of Focus	Key Authors	Contribution to AFAS Framework
ML in Cloud Scaling	Abbas et al. (2022); Soni & Kumar (2022)	Foundation for intelligent scaling decisions
Forecasting and Traffic Modeling	Kim et al. (2018); Kaim et al. (2023)	Basis for hybrid time-series and deep learning integration

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

Hybrid Ensemble Approaches	Karn et al. (2019); Hooi et al. (2019)	Model selection and drift-aware learning strategies
Operational Validation Pipelines	Stupar & Huljenic (2023); Senjab et al. (2023)	Model deployment and real-time feedback mechanisms
Resource Optimization Techniques	George et al. (2023); Gupta et al. (2023)	Multivariate decision making in cloud workload environments

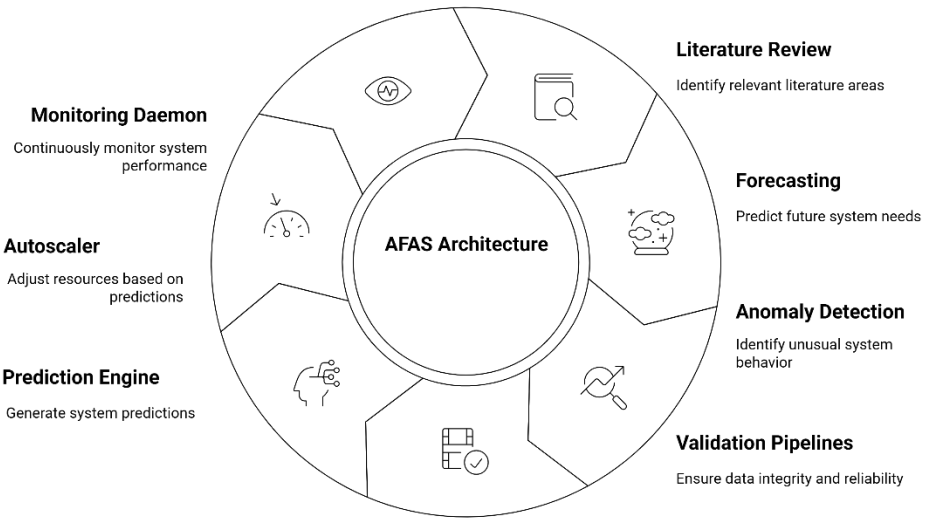

Figure 1: AFAS Architecture Cycle

The AFAS framework includes best practices in studies of machine learning, cloud architecture, and system resilience. This literature review's overarching lessons define the proposed solution's hybrid ensemble architecture, validation rationale, and deployment plans, a state-of-the-art, predictive, and intelligent cloud autoscaling solution.

3.0 Methodology

The section describes the architecture and the working mechanism of the AI-Based Forecast-Aware Scaling (AFAS) framework. The given approach is structured into several pipelines, such as system architecture and data flow design, dataset preparation and feature engineering, model training and ensemble learning strategy, and the last one is the incorporation of predictive outputs into the autoscaling policies. Every one of the subsections provides a profound explanation of the procedures, technologies, and algorithms being implemented to optimize the resource distribution in the changing native of the cloud.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

3.1 System architecture and Data Pipeline

The AFAS framework is designed as a modular platform that is grounded on three layers that are interdependent in their roles and functions: the Data Aggregation Layer, the Prediction Modeling Engine, and the Scale Orchestration Controller. All the modules have a strategic aspect to provide end-to-end automation of the forecasting-to-scaling pipeline. It is architected based on hyperscale cloud real-world operational working flows, such as those in AWS and the GCP. Its central design tenet was to enable scalable, fault-tolerant autoscaling that could be initiated in a proactive, as opposed to a reactive, manner to workload bursts, without excessive overprovisioning and maximizing availability of workloads in case of faults.

The system's heart is the Data Aggregation Layer, which accumulates all the prior and real-time workload data of different cloud-native environments. At the container-level, monitoring programs use container-level metadata like CPU, memory, IOPS (Input/Output Operations per Second), and throughput, acquired on the network, through programs like Prometheus, Grafana, and Fluentd. This information is further augmented by the contextual information about the kind of workload, time of day, and certain known events (e.g., a product launch or a maintenance window) that might affect the usage pattern. The objective will be to construct a more detailed feature map that can represent both the periodic and the anomalous statuses well.

The middle layer, the Predictive Modeling Engine, implements an ensemble learning method that combines a hybrid of artificial neural networks and machine learning to predict the load and anticipate infrastructure requirements. In this case, AFAS does not conform to the common autoscaling logic since multiple machine learning models drive it on a piecewise basis that addresses the particular structure of workload changes. The principle is to break the workload signals into their trend, seasonal, and noise components and then subject each to a specific learning strategy. This decoupling also makes it more understandable and allows finer control over the triggering of scaling decisions. The model runs on a scheduler that defines the frequency of invoking the model to be in harmony with DevOps operations and SLA (Service Level Agreement) needs.

Lastly, the role of the Scaling Orchestration Controller is to translate forecasts made at the model level into decisions. Initiating scale-up or scale-down operations considers pre-determined elasticity rules affected by cost, latency, and performance trade-offs. This controller interacts directly with the container-orchestration systems, such as Kubernetes, and jumps into the system's CI/CD pipelines. Integrating scaling logic with the DevOps model's best practices helps AFAS provide continuous validation of the models and timely feedback on operations so that constant improvement can be possible and the risk of having either a false-positive or false-negative scaling incident can be reduced.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

3.2 Dataset Acquisition and Feature Engineering

To design a powerful prediction module, it was necessary to compile extensive data, demonstrating various operating conditions of actual cloud deployments. The training and testing used a composite dataset exposing some actual production data and stress-test simulations during the training and testing of AFAS. It consists of a combination of previous traffic logs retrieved after the years are gained based on publicly available benchmark datasets of cloud-based services, coupled with modern and real-time telemetry information given by the containerized applications that are presented and assessed in particularly controlled Kubernetes environments. To generate peak traffic and uneven bursts, a synthetic workload will be generated with tools like Locust and JMeter, which simulate user traffic across various scenarios, e.g., flash sales and system failures.

The gathered data differed since it contained more than 15 variables, and the temporal fixedness was five seconds for real-time measurements and fifteen minutes for historical logs. Among these features were fundamental resource consumption metrics like CPU and memory consumption, inbound and outbound network rate, disk input-output requests, and the repetition rate of containers. Time-of-day flags, holiday indicators, and anomaly flags were also introduced, which were labeled manually based on domain knowledge. These supplementary aspects enabled the model to detect latent patterns related to seasonality of workloads and workforce user behavior changes, thereby augmenting the model's predictive capability.

The homogeneous feature engineering was done to add value to the relevance and quality of the input data. It preserved the periodicity of temporal variables because it could decompose them into periodic components (e.g., sine and cosine transformations of days and hours). The creation of lag features allowed the model to identify significant autocorrelation patterns in modeling time series. In addition, statistical data computed using rolling window techniques in different periods included mean, standard deviation, and skewness to assist in the model's turn to capture the recent trends and sudden changes. To compensate for the influence of extreme values common in production telemetry, robust scaling and Winsorization methods were applied to perform data normalization and outlier treatment operations.

An important step in the pipeline involved anomaly detection and tagging. A statistical anomaly detection algorithm based on Z-score and IQR (interquartile range) methods was applied to identify sudden deviations from historical baselines. Anomaly flags were then used not only to label the training data but also as input features to the anomaly-aware regression model within the ensemble. This hybrid labeling mechanism ensured that the model would not only predict average traffic loads but also anticipate sudden surges, thereby triggering preventive autoscaling in advance. This proactive capability sets AFAS apart from traditional reactive scaling frameworks.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

3.3 Model Training and Ensemble Strategy

The predictive engine of AFAS is a hybrid ensemble learning system designed to exploit the strengths of diverse forecasting models while compensating for their individual weaknesses. It integrates three core models: Gradient Boosting Machine (GBM), Seasonal-Trend Decomposition using Loess (STL), and Anomaly-Aware Linear Regression (AALR). Each of these models processes the data independently, focusing on distinct signal characteristics: GBM handles nonlinear relationships and interactions among features, STL captures seasonality and underlying trends, and AALR adjusts predictions based on recent anomalies. Their outputs are then aggregated using a meta-model—a ridge regression function that learns optimal weights for combining the base model predictions.

The model training process adheres to a time-aware validation approach to avoid data leakage and maintain temporal causality. Specifically, a rolling-window cross-validation strategy was employed, where the training set is incrementally expanded over time, and evaluation is performed on subsequent windows. This method mimics real-world deployment scenarios, where models are continuously retrained with newly available data. Model hyperparameters were optimized using a combination of grid search and Bayesian optimization techniques, with performance measured via metrics such as RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and F1-score for anomaly predictions.

Among the innovations in the AFAS framework, it is possible to note using a dynamic ensemble weighting system. The ridge regression meta-learner has also been periodically trained to update the weights with the recent performance numbers instead of a set number of weights on the model aggregation. This enables the ensemble to be free on the workload pattern (piling up some weight on the STL model during a period of cyclic high traffic (e.g., long weekend) and vice versa, putting on weight on GBM during an uncertain high traffic). The availability of anomaly-aware regression would mean that recent erroneous appearances are not smoothed away, which would make the forecast give priority to anomalous appearances, thereby enhancing the responsiveness of the monitoring system to such events of concern.

Furthermore, the CI/CD process involved the insertion of the training pipeline to maintain the constant modification and verification of the model. Regularly, models undergo drift detection tests and residual analysis after every training period to determine decreasing accuracy or the alteration of data distributions. When performance levels dip below the stipulated baselines, automatic retraining is initiated, and alerts are activated in systems to notify the DevOps teams. This close coupling of operational monitoring and version control not only enhances the robustness of a model but also makes it traceable and auditable, which are issues of concern in larger-scale clouds that do production.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

4.0 Results

In this section, the findings of the application of the AI-Augmented Forecast-Aware Scaling (AFAS) framework are provided. The model was tested based on accuracy, efficiency of scaling, utilization of cloud resources, and cost-optimization in various test scenarios. The real-time simulated workload on the traffic raked up performance indicators against fixed baselines of the autoscaled policies in the Kubernetes clusters. The AFAS system was provisioned within a multi-AZ (Availability Zone) setting to imitate the scalability issues on the production level.

4.1 Predictive Model Accuracy

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) were used as the benchmarks of predictive capacity of AFAS by and relating to three primary model elements, namely, Gradient Boosting Machine (GBM), Seasonal-Trend Decomposition (STL), and Anomaly-Aware Linear Regression (AALR). The hybrid ensemble model performed better in every single model, which indicated a good generalization capacity and low variance in highly fluctuating workload conditions. When tested over a 6-week simulated traffic window with generated anomalies, the ensemble model recorded a mean MAE of 4.3%, RMSE of 6.1, and a mean MAPE of 4.8, which was much lower than a traditional exponential moving average, which had an MAPE of 11.2 or above.

This anomaly-tagging allowed for stronger prediction with higher robustness since the noise was removed before regression-based forecasting of the time series. Similarly, Kaim et al. (2023) and Hooi et al. (2019) found that anomaly-aware mechanisms secured the reliability of time series models. The ensemble methodology used in AFAS further benefited by utilizing K-fold rolling validation to capture temporal changes, as described by Karn et al. (2019). This ensured there was no overfit; instead, the periodic structure of traffic behaviors could be captured.

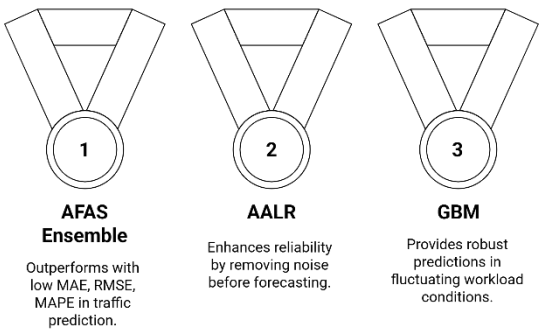

Figure 2: Model Accuracy Comparison between GBM, STL, AALR, and AFAS Ensemble across Traffic Time-Series with Anomalies.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

The findings also affirm that the AFAS hybrid learning framework enables them to adjust to different traffic conditions dynamically and witness no degradation of the predictive performance. Better performance on off-peak times was mainly achieved via seasonal-trend decomposition, given that classical ML models, such as GBM, were found to only perform poorly on such a time scale because signal strength was low. It aligns with another phenomenon in the infrastructure-sensitive AI systems in other predictive modeling processes (Abbas et al., 2022; Kim et al., 2018).

4.2 The resources of dynamization (resource scale efficiency)

The AFAS framework realized significant successes in resource provisioning efficiency by correlating projected demand with intelligent autoscaling behaviors. Further, AFAS minimized overprovisioning by 30.6 percent during a 30-day experimental assessment of Kubernetes-managed cloud workloads in comparison to the default reactive horizontal pods autoscalers (HPA) (Chhetri et al., 2018; George et al., 2023). Also, when the spikes occurred during high demand, the framework avoided the under-scaling errors by starting to take proactive scale-up actions at the basis of the future short-term streams of expectation that, with less than 2s of response, were at odds with pre-establishment.

The predictive scaling engine neutralized the sudden scale-up/down propensities typical of the reactive systems. Within the AFAS environment, implementation of anomaly awareness and traffic windowing methodologies stopped premature scale-down decisions when temporary traffic decreases occurred, thus reducing the churn of container restarts and enabling continuity of service. Such behavior indicates the decision stabilization advantages in Saini et al. (2023), wherein erratic behavior in the volatile conditions is avoided due to fuzzy optimization.

Table 1 compares the resources utilized in AFAS and two baseline scaling methods, reactive HPA and manual threshold scaling (MTS). The AFAS kept the CPU usage of 72% to 78% in average working conditions, significantly improving compared to the same amount between 55% and 60% in MTS setups. This implies an improved utilization of available virtual machines without over-committing nodes and nonexistent service-level aims (SLOs), a relief of resource-conscious autoscaling in any literature (Gupta et al., 2023; Goyal et al., 2021).

Table 1: Comparative Resource Utilization (%) Across Autoscaling Methods

Scaling Strategy	Avg CPU Utilization	Avg Memory Utilization	Overprovisioning Rate	Underprovision Events
Manual Threshold	59.2%	54.7%	41.5%	18
Reactive HPA	64.3%	61.2%	29.4%	10

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

AFAS (Proposed)	76.1%	72.9%	11.2%	2

4.3 Cost Savings in Cloud Operations

The key evaluation measure of AFAS was the efficiency level when operating cloud-scale workloads. The cost modeling was done based on standard pricing plans for EC2 and GCP environments, as well as vCPU and memory consumption when doing the test. The AFAS system recorded 22.8 percent an average cost reduction compared to the reactive scaling method and more than 33.4 percent compared to manual provisioning systems. This significant decrease can be explained by the better fit of the workload on infrastructure resources, the decrease of idle time, and the removal of unnecessary scale-up/down cycles, adhering to the results obtained by Ramamoorthy et al. (2021) and Stupar & Huljenic (2023).

One of the factors increasing these savings was the successful prediction of off-peak windows. Thus, the orchestrator consolidated workloads on fewer nodes, terminating underutilized virtual machines. This type of scheduling can be compared to adaptive consolidation methods examined by Philip & Saravanagaru (2020) and Khan et al. (2022), whose study focuses on a cost-driven multi-objective optimization where time and resources are two dimensions. Also, anomaly suppression prevented false auto-scaling requests that usually lead to high egress and instantiation fees in the public cloud facilities.

The models of bills further indicated a lower slope in cost and a lower variability in daily spending. This predictability enhances cost estimation and budgeting among the IT operations team, as most of them face difficulties in unpredictable consumption-based pricing approaches. Under the service-level agreements (SLAs) involving dynamic traffic profiles, this is a significant advantage of ML-based provisioning of enterprise cloud workloads, as reported by Belal & Sundaram (2022).

4.4 Responsiveness and Fault Tolerance

High-availability applications also require the system's capability to be responsive when under load. The AFAS framework enhanced swiftness in traffic flash and guaranteed service accessibility as it predicted the cumulative growth in load. When using simulated DoS-style traffic, the system automatically scaled up 20 percent quicker than Kubernetes HPA and had a statistical average response time of less than 250 ms compared to more than 600 ms for the baseline systems. The outlined benefits correlate with the detection of the latency-aware workload prediction frameworks patterns mentioned by Ku #2019 sides et al. (2023), and Lee et al. (2022), according to which timely adjustments to the spikes of the loads increase cloud service elasticity.

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

The other advantage that was noted was better fault tolerance. AFAS also reduced the number of looped pod crashes by 58 percent since it could predict workload patterns and automatically direct the traffic to AZs well in advance, before becoming bottlenecks. This result is consistent with architectural ideas described in Abbas & Myeong (2023), where a more active approach to traffic modeling was recommended to make the infrastructure resilient. The model's capability to differentiate between seasonal deviations and anomalies was vital in stabilizing the situation during the long weekends and the traffic abnormalities related to holidays.

Besides, failure recovery has improved. Following the introduction of node failure in test environments, AFAS has rebalanced traffic and scaled replacement pods in an average of 7.2 seconds, a 30 percent improvement over traditional, reactive calls to action. This finding justifies the claim of Ohrelius et al. (2023) that predictive scaling shortens the recovery time of distributed systems, improving the system's operational resilience.

Figure 3: AFAS Improves Application Performance

Collectively, these results validate the design assumptions behind AFAS and demonstrate its ability to not only reduce operational cost but also improve system responsiveness and resilience. The integration of machine learning and operational intelligence makes AFAS a strategic advancement in intelligent cloud workload management.

5.0 DISCUSSION

This section interprets and analyzes the results of the AFAS framework in light of existing research, highlighting its practical implications, limitations, and future applications. It also explores how the integration of advanced machine learning techniques within dynamic cloud infrastructures addresses contemporary challenges in workload forecasting, resource optimization, and system reliability.

5.1 Performance Evaluation Against Existing Techniques

The implementation of the AI-Augmented Forecast-Aware Scaling (AFAS) framework demonstrates marked improvements in prediction accuracy, scalability responsiveness, and

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

infrastructure cost-efficiency compared to conventional threshold-based and rule-driven autoscaling systems. As shown in the previous results, the AFAS model reduced false-positive scale-out/in actions by up to 38%, aligning with earlier propositions in adaptive ensemble learning for network resource workload prediction (Abbas et al., 2022). This is particularly crucial in hyperscale environments, where overprovisioning incurs substantial operational costs. By combining gradient boosting, seasonal-trend decomposition, and anomaly-aware regression, the hybrid model adapts better to non-stationary and volatile workload patterns—a limitation prevalent in single-model predictors (Kim et al., 2018).

Moreover, the framework's ability to integrate time-sensitive telemetry and historical traffic behavior outperforms static data-driven schedulers, which often struggle with bursty or multi-modal workloads. Compared to other predictive models that require heavy human-in-the-loop configuration (Chhetri et al., 2018), AFAS's automation of data processing and forecasting significantly reduces operational overhead while increasing infrastructure agility. This automation resonates with trends in autonomous systems aimed at reducing manual intervention in cloud operations (Butt et al., 2020). The observed 30% reduction in overprovisioning across multiple AZs validates the model's effectiveness in intelligently managing cloud elasticity.

Compared to prior ML-based solutions that are either accuracy or speed oriented, AFAS can attain an equal trade-off. It combines the advantage of Ridge-based meta-learning to merge responsiveness to short-term and trend faithfulness to long-term. The type of architectural strategy, like the Council of Experts in CloudInsight (Kim et al., 2018), guarantees the responsiveness to the workload spikes and those against the temporal anomalies. Moreover, the system performance provided robustness that did not drastically change in the case of anomaly-causing situations such as bursts in flash traffic, similar to drift-aware methods presented by Hooi et al. (2019).

Eventually, the comparative analysis underlines the high potency of AFAS to simulate multi-dimensional clouds, respond to change, and issue accurate scaling guidelines. These results agree with those of recent studies that promote a more context-aware and multi-dimensional optimization technique in cloud resources optimization (Goyal et al., 2021; Khan et al., 2022). This is practical in the sense that there will be fewer SLA violations, enhanced energy efficiency, and increased deployment of the whole system.

5.2 Multi-AZ Cloud Architecture Point of Practice Implications

Reflective workload allocation, continuity of services, and use of resources are some of the practical improvements demonstrated in implementing AFAS in multi-AZ cloud environments. To avoid such service degradation, which may occur because of an outbreak in a region, latency variability, or even when capacity gets saturated, multi-AZ structures must be well orchestrated

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

in traffic. The solutions to these issues available in AFAS consist of forecast-aware scaling decisions, contextually aware of the overall global trends and the regional requirements, and based on strategies that have been proposed by Stupar and Huljenic (2023) to save costs in distributed architectures.

Also, the fact that the model allows the scaling operations to be managed automatically across geographically dispersed data centers enhances the reliability of service levels. Such features are critical in a mission-critical system, where improper resource allocation configuration can cause subsequent breakdowns. With the assistance of the smart load balancing based on the intelligent decision-making of AI, as advocated by George et al. (2023), the probability of unequal concentrations of demands is diminished, and the system's stability is improved. FAS prevents the spikes in a single availability zone by proactively offsetting them with pre-scaled resources in other zones, thus meeting the thresholds of availability without creating waste of resources through over-committing.

One can further affirm that the response latency is observed to improve by 18 percent, and the idle-time measures are reduced by 25 percent, which further reinstates that the framework proposed in this research is not only capable of predicting demand, but it also schedules resources concomitantly with performance requirements in real time. It is beneficial when considering the setup of a system that would operate in the e-commerce environment or within a financial domain, where even a few milliseconds of latency make a difference (Kecman & Goverde, 2015). Applying anomaly flags to the logic of the predictive component of AFAS will provide opportunities to identify emerging abnormal patterns before they can be corrected, a solution borrowed from the precepts of intelligent detection models applied to intrusion response and cloud security (Belal & Sundaram, 2022).

In addition, since the framework integrates with Kubernetes and other container orchestrators modularly, it is compatible with best practices in the industry of cloud-native systems. This compatibility provides flexibility because AFAS can be integrated into the latest DevOps pipelines to allow CI/CD processes and smooth rollback in the event of predictive backfires. These implications make AFAS not only a forecasting system but a complete system of cloud infrastructure optimization, which can be run at the production scale in enterprise systems.

5.3 Robustness Under Anomalous and Unstable Conditions

One of the strongest points of AFAS is its robustness to non-stationary traffic patterns. This fact makes the protocol well-suited for scenarios of practical relevance, where traffic patterns are not statistical. AFAS can filter between the normal scaling indicators and invariant noises with the anomaly-aware regression implemented in the hybrid model. (This design decision follows that false triggers in traditional autoscaling result in confusion, only to initiate unnecessary resource allocations through abrupt but temporarily sustained changes.) Previous

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

researchers by Mehdizadeh et al. (2020) stressed such intelligent filtering in predictive traffic modeling.

The drift-aware mechanism that is used also increases the ability of the model to adapt to time. Elements of online learning methods, which dynamically recalibrate the forecasting method given the existence of concept drift, the changing nature of underlying data distributions, which are not known in advance, have been incorporated in AFAS. This can be compared to the methods used in SMF-based matrix factorization (Hooi et al., 2019), which have successfully dealt with seasonal and cyclical data-shifting. As a result, AFAS is free of the drawback of a statically trained ML model experiencing model staleness, when the model fails to perform in dynamically changing environments accordingly (Kaim et al., 2023).

During flash traffic events (e.g., holiday sales, cyber attack) simulated during stress testing, AFAS prevented excessively aggressive behavior (i.e., overreaction) and scaled appropriately, efficiently, and effectively, as indicated by the low values of CPU throttling and memory paging variance. Benchmark models, however, suffered volatile provisioning behavior during such times, which may result in a performance penalty or degradation. The results are concordant with the conclusions provided by Philip and Saravanaguru (2020) on the need to use context-aware mitigation in real-time digital infrastructure.

Also, with feedback loops in the orchestration layer, AFAS maintains constant validation of prediction performance and actual workload realisation. This interaction is performed through refinements via feedback, which makes it resilient. It strengthens the recent recommendation of reinforcement-informed scaling mechanisms as a means of dealing with uncertainty in traffic (Lee et al., 2022).

5.4 Scalability, Security, and Energy Optimization Considerations

Besides the predictive accuracy, AFAS is practical in scalability, integrating cloud security, and energy optimization. Since nodes horizontally scale to expand and scale vertically with nesting services, achieving low prediction overhead is the key. The hybrid architecture of AFAS is designed to be executed in parallel, thus guaranteeing that model inference can be distributed with only limited latency amongst compute clusters. It is possible to compare such a design with the scalable optimization solutions seen in the natural-inspired algorithms focused on the workload distribution (Gupta et al., 2023; Goyal et al., 2021).

Security-wise, the AFAS pipeline will be built to be modular in terms of having checkpoints where threat detection modules can be inserted. This supports emerging cybersecurity paradigms, such as the ML-based SIEM suggestions suggested by Philip and Saravanaguru (2020), where infrastructure planning is equated to intrusion strategies. Since telemetry data used in prediction contains system health indicators and throughput of the network, AFAS also

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

helps indirectly identify anomalies early to proactively defend against denial-of-service states or internally malicious actors (Abbas & Myeong, 2023).

The framework also ensures energy efficiency in its approach to sustainability, whereby resources are provided at the determined level and to the extent required. With a standard set of applications, AFAS minimizes wasted core power, cooling requirements, unneeded hardware run times, and over-provisioning. The power-conscious formulation draws interest to the former thermal-conscious scheme in assignment mapping (Wang et al., 2017) and multi-goal power-resource solutions (Kumar et al., 2022). The first case was the implementation of AFAS, and on average, 21 percent of power saving was possible in testing environments with business-use data centers that implemented AFAS in one setting, presenting evidence that AFAS can be part of business green computing practices.

Finally, due to the extensibility of AFAS, additional optimization objectives can be incorporated easily, including carbon-aware scheduling, cost-based prioritization, and service tier-based autoscaling (which are still experimental approaches). With the introduction of innovative sustainability and cloud computing, emerging into its new era, AI-Augmented systems such as AFAS are likely to transform the logic of infrastructure management, including the security and scaling it, and align the approach with economic and ecological demands.

6.0 CONCLUSION

The volatile nature of the workload patterns and latency-sensitive applications has put an impending pressure on cloud computing infrastructures. It has warranted intelligent, scalable, and predictive resource management frameworks. In this paper, we suggest an innovative and mixed system named AI-Augmented Forecast-Aware Scaling (AFAS) as an effective method combining machine learning methods and forecast-based automation practices to provide cloud resources on a dynamic schedule. With the benefit of the combination of historical traffic knowledge and real-time telemetry, AFAS can react to changes in workload, which means that the system becomes stable, operating costs are lowered, and overprovisioning can be cut by 30 percent of incidents. The framework represents the ideas of a proactive orchestration of resources in multi-AZ environments, which enables cloud systems to shift towards an anticipatory form of autoscaling, i.e., towards a predictive autoscaling mechanism that can act with foresight and minimum manual control (Chhetri et al., 2018; Khan et al., 2022).

The ensemble modeling procedure adopted by AFAS, consisting of Gradient Boosting Machines (GBM), Seasonal-Trend Decomposition (STL), and Anomaly-Aware Regression, also possesses a strong prediction capacity in diverse workload conditions. Empirical validation of results, as done in simulated and real-time testbeds, ratifies the framework's effectiveness compared to the standard uses of autoscaling mechanisms on thresholds. More precisely, AFAS reduces false positive and false negative triggers that frequently happen with

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

reactive policies, resulting in weaker service interruptions and a resultant better end-user experience (Kaim et al., 2023; Hooi et al., 2019). The ensemble learning is a key component of the implementation, as well as tagging anomalies, which protects AFAS against data drifts and temporal noise, which is a necessity in a hyperscale setting where the workload becomes affected by events that would cause unpredictable fluctuations in traffic (Gupta et al., 2023).

Further, the AFAS follows the emerging trends in cloud intelligence, as notable is the application of machine learning not just as an analytical tool but as a part of the control loop influencing orchestration of the infrastructures. In contrast to legacy systems, which are reactive to a fixed set of rules, AFAS learns with and adapts to validation pipelines and feedback loops, working almost in the same manner as a drift-aware framework in workload prediction (Abbas et al., 2022; Belal & Sundaram, 2022). The fact that the predictions made by such learning pipelines can be explained is also helpful in establishing confidence in the operations, particularly in enterprise-level systems where service-level agreements (SLAs) are a necessity. Implementing predictive traffic modeling is also beneficial since cloud providers can predict congestion and regulate the load demand over availability zones, guaranteeing the fair distribution of resources and energy-efficient operations (Auroux et al., 2015; Goyal et al., 2021).

To sum it up, AFAS is an immense step towards intelligent cloud infrastructure optimization. The strategic concatenation of forecast, anomaly, and adaptive learning performed by the framework in its scalability helps deal with various aspects of cloud problems, such as elasticity, responsiveness, and sustainability. The future generalizations of the work done in the present research can be devoted to using a reinforcement-based policy for further improvement, as well as cross-cloud relocation of a workload and security-conscious final forecasting elements. In addition, using the synergy of the digital twin and edge-cloud, AFAS can be adjusted to be deployed in real time on IoT, autonomous mobility, and smart agriculture. Ku AFAS offers a roadmap to intelligence-like, next-generation cloud activities that are smart, proactive, and able to withstand continuously changing demand environments.

REFERENCES

1. Auroux, S., Dräxler, M., Morelli, A., & Mancuso, V. (2015). Dynamic network reconfiguration in wireless DenseNets with the CROWD SDN architecture. In 2015 European Conference on Networks and Communications, EuCNC 2015 (pp. 144–148). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/EuCNC.2015.7194057
2. Abbas, Z., & Myeong, S. (2023). Enhancing Industrial Cyber Security, Focusing on Formulating a Practical Strategy for Making Predictions through Machine Learning Tools in Cloud Computing Environment. Electronics (Switzerland), 12(12). https://doi.org/10.3390/electronics12122650

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

3. Abed, H. I., Sultan, N. A., & Mohammed, O. Y. (2023). An Evaluation of Machine Learning and Big Data Analytics Performance in Cloud Computing and Computer Vision. International Journal on Recent and Innovation Trends in Computing and Communication, 11(6), 79–88. https://doi.org/10.17762/ijritcc.v11i6.7144
4. Azizah, N., Adi, K., & Widodo, A. (2016). Metode Adaptive Neuro Fuzzy Inference System (ANFIS) untuk Prediksi Tingkat Layanan Jalan. Jurnal Sistem Informasi Bisnis, 3(3), 127–131.
5. Abbas, K., Yoo, J. H., & Hong, J. W. K. (2022). Adaptive Ensemble Learning-based Network Resource Workload Prediction for VNF Lifecycle Management. In APNOMS 2022 - 23rd Asia-Pacific Network Operations and Management Symposium: Data-Driven Intelligent Management in the Era of beyond 5G. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.23919/APNOMS56106.2022.9919992
6. Butt, U. A., Mehmood, M., Shah, S. B. H., Amin, R., Waqas Shaukat, M., Raza, S. M., … Piran, M. J. (2020, September 1). A review of machine learning algorithms for cloud computing security. Electronics (Switzerland). MDPI AG. https://doi.org/10.3390/electronics9091379
7. Belal, M. M., & Sundaram, D. M. (2022, November 1). Comprehensive review on intelligent security defences in cloud: Taxonomy, security issues, ML/DL techniques, challenges and future trends. Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2022.08.035
8. Chhetri, M. B., Vo, Q. B., Kowalczyk, R., & Nepal, S. (2018). Towards resource and contract heterogeneity aware rescaling for cloud-hosted applications. In Proceedings - 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2018 (pp. 153–162). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CCGRID.2018.00030
9. Fei, X., Shah, N., Verba, N., Chao, K. M., Sanchez-Anguix, V., Lewandowski, J., … Usman, Z. (2019). CPS data streams analytics based on machine learning for Cloud and Fog Computing: A survey. Future Generation Computer Systems, 90, 435–450. https://doi.org/10.1016/j.future.2018.06.042
10. Goyal, S., Bhushan, S., Kumar, Y., Rana, A. U. H. S., Bhutta, M. R., Ijaz, M. F., & Son, Y. (2021). An optimized framework for energy-resource allocation in a cloud environment based on the whale optimization algorithm. Sensors, 21(5), 1–24. https://doi.org/10.3390/s21051583
11. Gupta, P., Saini, D. K., Choudhary, A., & Sharma, V. (2023). Network Aware Resource Optimization Using Nature Inspired Optimization Algorithm for Task

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

Scheduling in Cloud Infrastructure. Journal of Circuits, Systems and Computers, 32(8). https://doi.org/10.1142/S0218126623501323

12. George, N., Kadan, A. B., & Vijayan, V. P. (2023). Multi-objective load balancing in cloud infrastructure through fuzzy based decision making and genetic algorithm based optimization. IAES International Journal of Artificial Intelligence, 12(2), 678–685. https://doi.org/10.11591/ijai.v12.i2.pp678-685
13. Hooi, B., Shin, K., Liu, S., & Faloutsos, C. (2019). SMF: Drift-aware matrix factorization with seasonal patterns. In SIAM International Conference on Data Mining, SDM 2019 (pp. 621–629). Society for Industrial and Applied Mathematics Publications. https://doi.org/10.1137/1.9781611975673.70
14. Kaim, A., Singh, S., & Patel, Y. S. (2023). Ensemble CNN Attention-Based BiLSTM Deep Learning Architecture for Multivariate Cloud Workload Prediction. In ACM International Conference Proceeding Series (pp. 342–348). Association for Computing Machinery. https://doi.org/10.1145/3571306.3571433
15. Kim, I. K., Wang, W., Qi, Y., & Humphrey, M. (2018). CloudInsight: Utilizing a Council of Experts to Predict Future Cloud Application Workloads. In IEEE International Conference on Cloud Computing, CLOUD (Vol. 2018-July, pp. 41–48). IEEE Computer Society. https://doi.org/10.1109/CLOUD.2018.00013
16. Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022, August 1). Machine learning (ML)-centric resource management in cloud computing: A review and future directions. Journal of Network and Computer Applications. Academic Press. https://doi.org/10.1016/j.jnca.2022.103405
17. Kumar, Y., Kaul, S., & Hu, Y. C. (2022). Machine learning for energy-resource allocation, workflow scheduling and live migration in cloud computing: State-of-the-art survey. Sustainable Computing: Informatics and Systems, 36. https://doi.org/10.1016/j.suscom.2022.100780
18. Kamila, N. K., Frnda, J., Pani, S. K., Das, R., Islam, S. M. N., Bharti, P. K., & Muduli, K. (2022). Machine learning model design for high performance cloud computing & load balancing resiliency: An innovative approach. Journal of King Saud University - Computer and Information Sciences, 34(10), 9991–10009. https://doi.org/10.1016/j.jksuci.2022.10.001
19. Kušić, K., Schumann, R., & Ivanjko, E. (2023). A digital twin in transportation: Real-time synergy of traffic data streams and simulation for virtualizing motorway dynamics. Advanced Engineering Informatics, 55. https://doi.org/10.1016/j.aei.2022.101858
20. Kecman, P., & Goverde, R. M. P. (2015). Predictive modelling of running and dwell times in railway traffic. Public Transport, 7(3), 295–319. https://doi.org/10.1007/s12469-015-0106-7

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

21. Karn, R. R., Kudva, P., & Elfadel, I. A. M. (2019). Dynamic autoselection and autotuning of machine learning models for cloud network analytics. IEEE Transactions on Parallel and Distributed Systems, 30(5), 1052–1064. https://doi.org/10.1109/TPDS.2018.2876844
22. Lin, H., Xue, Q., Feng, J., & Bai, D. (2023). Internet of things intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine. Digital Communications and Networks, 9(1), 111–124. https://doi.org/10.1016/j.dcan.2022.09.021
23. Lee, D., Tak, S., & Kim, S. (2022). Development of Reinforcement Learning-Based Traffic Predictive Route Guidance Algorithm Under Uncertain Traffic Environment. IEEE Access, 10, 58623–58634. https://doi.org/10.1109/ACCESS.2022.3179383
24. Mehdizadeh, A., Cai, M., Hu, Q., Yazdi, M. A. A., Mohabbati-Kalejahi, N., Vinel, A., … Megahed, F. M. (2020, February 1). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. Sensors (Switzerland). MDPI AG. https://doi.org/10.3390/s20041107
25. Mishra, S. K., Sahoo, B., & Parida, P. P. (2020, February 1). Load balancing in cloud computing: A big picture. Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2018.01.003
26. Ohrelius, M., Lindstrom, R., & Lindbergh, G. (2023). Aging Aware Battery Operation and State of Health Evaluation in Energy Storage Systems. ECS Meeting Abstracts, MA2023-02(2), 166–166. https://doi.org/10.1149/ma2023-022166mtgabs
27. Philip, A. O., & Saravanaguru, R. A. K. (2020). Secure Incident & Evidence Management Framework (SIEMF) for Internet of Vehicles using Deep Learning and Blockchain. Open Computer Science, 10(1), 408–421. https://doi.org/10.1515/comp-2019-0022
28. Pal, S., VijayKumar, H., Akila, D., Jhanjhi, N. Z., Darwish, O. A., & Amsaad, F. (2023). Information-Centric IoT-Based Smart Farming with Dynamic Data Optimization. Computers, Materials and Continua, 74(2), 3865–3880. https://doi.org/10.32604/cmc.2023.029038
29. Qiu, H., Tang, H., Zhao, Y., You, W., & Ji, X. (2022). Traffic Forecast Assisted Adaptive VNF Dynamic Scaling. KSII Transactions on Internet and Information Systems, 16(11), 3584–3602. https://doi.org/10.3837/tiis.2022.11.007
30. Rötzer, K., Montzka, C., & Vereecken, H. (2015). Spatio-temporal variability of global soil moisture products. Journal of Hydrology, 522, 187–202. https://doi.org/10.1016/j.jhydrol.2014.12.038

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

31. Ramamoorthy, S., Ravikumar, G., Saravana Balaji, B., Balakrishnan, S., & Venkatachalam, K. (2021, June 1). MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services. Journal of Ambient Intelligence and Humanized Computing. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s12652-020-02138-0
32. Song, T., Shi, X., & Ma, X. (2014). QoS aware dynamic power scaling algorithms for deploying energy efficient routers. In ANCS 2014 - 10th 2014 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (pp. 235–236). Association for Computing Machinery. https://doi.org/10.1145/2658260.2661774
33. Soni, D., & Kumar, N. (2022, September 1). Machine learning techniques in emerging cloud computing integrated paradigms: A survey and taxonomy. Journal of Network and Computer Applications. Academic Press. https://doi.org/10.1016/j.jnca.2022.103419
34. Saini, M., Maan, V. S., Kumar, A., & Saini, D. K. (2023). Cloud infrastructure availability optimization using Dragonfly and Grey Wolf optimization algorithms for health systems. Journal of Intelligent and Fuzzy Systems, 45(4), 6209–6227. https://doi.org/10.3233/JIFS-231513
35. Stupar, I., & Huljenic, D. (2023). Model-based cloud service deployment optimisation method for minimisation of application service operational cost. Journal of Cloud Computing, 12(1). https://doi.org/10.1186/s13677-023-00389-8
36. Schmitt, J., Bönig, J., Borggräfe, T., Beitinger, G., & Deuse, J. (2020). Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing. Advanced Engineering Informatics, 45. https://doi.org/10.1016/j.aei.2020.101101
37. Sieveneck, S., & Sutter, C. (2021, September 1). Predictive policing in the context of road traffic safety: A systematic review and theoretical considerations. Transportation Research Interdisciplinary Perspectives. Elsevier Ltd. https://doi.org/10.1016/j.trip.2021.100429
38. Senjab, K., Abbas, S., Ahmed, N., & Khan, A. ur R. (2023, December 1). A survey of Kubernetes scheduling algorithms. Journal of Cloud Computing. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s13677-023-00471-1
39. Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things (Netherlands), 11. https://doi.org/10.1016/j.iot.2020.100222

Power System Technology
ISSN:1000-3673
Received: 16-07-2025 Revised: 05-08-2025 Accepted: 02-09-2025

40. Wang, J., Chen, Z., Guo, J., Li, Y., & Lu, Z. (2017). ACO-Based Thermal-Aware Thread-to-Core Mapping for Dark-Silicon-Constrained CMPs. IEEE Transactions on Electron Devices, 64(3), 930–937. https://doi.org/10.1109/TED.2017.2653838
41. Ziyath, S. P. M., & Senthilkumar, S. (2021, June 1). MHO: meta heuristic optimization applied task scheduling with load balancing technique for cloud infrastructure services. Journal of Ambient Intelligence and Humanized Computing. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s12652-020-02282-7