Jurnal Teknik Informatika C.I.T Medicom, 16 (6) (2025) 363-373

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

A Comprehensive Review of Machine Learning Paradigms for Large-Scale Smart System

Morgan Jaden Liam
Electrical and Computer Engineering Department, Dalhousie University, Canada

Article Info

Article history

Received : Dec 28, 2024
Revised : Jan 28, 2025
Accepted : Jan 31, 2025

Key Words:

Machine Learning Paradigms;
Large-Scale Smart Systems;
Deep Learning;
Federated Learning;
Graph Neural Networks (GNNs).

Abstract

Large-scale smart systems such as smart cities, smart grids, smart healthcare, and IoT-based infrastructures generate massive volumes of complex, heterogeneous data that require intelligent analysis and real-time decision-making. Machine learning (ML) plays a central role in enabling these capabilities, yet the diversity of ML paradigms and the fragmented nature of existing studies make it difficult to determine which approaches are most effective for large-scale environments. This comprehensive review synthesizes and compares major ML paradigms, including supervised learning, unsupervised learning, reinforcement learning, deep learning, hybrid models, federated learning, and graph-based neural networks, across a wide range of smart system applications. The findings reveal that deep learning excels in processing high-dimensional and unstructured data, reinforcement learning performs best in autonomous and real-time control tasks, federated learning supports privacy-preserving analytics in distributed IoT ecosystems, and graph-based models offer superior performance in systems with interconnected network structures. The review also identifies key technological challenges such as data heterogeneity, computational complexity, communication bottlenecks, and privacy concerns that affect the scalability and deployment of ML in smart environments. By providing a unified comparison of ML paradigms and highlighting emerging trends, performance characteristics, and implementation challenges, this study offers valuable insights for researchers, system designers, engineers, and policymakers. The review further outlines future research directions aimed at enhancing scalability, robustness, interpretability, and real-time capability in next-generation smart systems.

Corresponding Author:

Morgan Jaden Liam
Electrical and Computer Engineering Department, Dalhousie University, Canada
6283 Alumni Crescent, Halifax, NS B3H 4R2, Canada
jadenliam @dal.ca

This is an open access article under the CC BY-NC license.

1. Introduction

The rapid development of digital technologies has driven the emergence of large-scale smart systems across various domains, including smart cities, smart grids, smart healthcare, and expansive Internet of Things (IoT) ecosystems. These systems are characterized by massive volumes of data generated continuously from interconnected devices, sensors, and platforms. As the scale of these systems increases, the data they produce becomes more complex and highly heterogeneous, encompassing

structured, semi-structured, and unstructured formats[1]. This complexity creates challenges for efficient data processing, system optimization, and real-time decision-making, all of which are critical for ensuring that smart systems operate reliably and effectively. Consequently, there is a growing need for intelligent automation tools capable of learning from data, adapting to dynamic environments, and supporting autonomous or semi-autonomous decision processes.

Machine Learning (ML) has become a foundational technology in addressing these needs[2]. Its ability to learn patterns, predict future states, optimize system behavior, detect anomalies, and facilitate autonomous control makes ML indispensable in the management of modern smart systems. Supervised learning enhances prediction accuracy, unsupervised learning enables pattern discovery, reinforcement learning supports adaptive decision-making, and deep learning excels in processing high-dimensional sensory data. However, deploying traditional ML approaches in large-scale smart environments presents new challenges. These include the computational difficulty of handling extremely large datasets, the need for continuous model updates in dynamic contexts, communication constraints in distributed systems, and limitations in model interpretability and latency. As smart systems expand in scale and complexity, conventional ML frameworks often struggle to maintain efficiency, scalability, and responsiveness.

Large-scale, cross-domain literature reviews have sought to map how machine learning (ML) is applied across smart systems[3]. For example, multiple recent surveys synthesize ML use in smart cities covering applications such as traffic management, air-quality monitoring, and smart health and discuss common challenges like data heterogeneity, privacy, and the need for scalable pipelines. Representative works include broad reviews published in the last few years that consolidate domain applications and highlight research needs for interpretability and real-time operation (e.g., Machine Learning for Smart Cities: A Comprehensive Survey, 2023).

In the smart-grid and power-systems domain, dedicated reviews examine how ML and AI methods improve forecasting, anomaly detection, demand-response optimization, and asset management. Recent papers (e.g., Banad et al., 2025; Niraula, 2023) review both classical ML models and contemporary deep learning techniques, and they emphasize practical constraints in deployment such as latency, reliability, and regulatory requirements. These reviews also call attention to hybrid solutions combining model-based and data-driven approaches to handle the scale and critical safety requirements of electrical networks.

Federated learning (FL) has emerged as a core research direction for privacy-preserving, distributed model training in large IoT and mobile ecosystems. The seminal work by McMahan et al. (2016/2017) introduced practical federated averaging for decentralized deep networks and demonstrated strong communication-efficiency and robustness to non-IID data; this paper effectively seeded a large body of FL research aimed at resource-constrained, widely distributed devices. Subsequent surveys and application-focused studies have extended FL to IoT environments examining heterogeneity, communication bottlenecks, and system-level orchestration (surveys in 2019-2025 provide detailed taxonomies and open challenges).

Graph neural networks (GNNs) and graph convolutional networks (GCNs) are another fast-growing direction because many smart systems naturally form graph structures (sensor networks, distribution grids, transportation graphs). The influential work by Kipf and Welling (2016/2017) on semi-supervised classification with graph convolutional networks demonstrated scalable spectral/spatial graph convolutions and has led to wide adoption of GNNs for tasks such as network anomaly detection, traffic forecasting, and topology-aware prediction in smart infrastructures. Recent applied research leverages GNNs to encode relational structure at scale and to improve generalization across interconnected subsystems.

Deep reinforcement learning (DRL) has been widely studied for control and sequential decision problems in intelligent systems. Mnih et al.'s (2015) deep Q-network (DQN) paper established that deep architectures can learn policies directly from high-dimensional sensory inputs, catalyzing many DRL applications in robotics, autonomous vehicles, and adaptive resource allocation for smart systems. Later work adapts DRL to large-scale control problems addressing issues such as sample efficiency,

safety constraints, and multi-agent coordination that are critical when scaling to real operational environments.

Edge AI and TinyML research addresses the opposite extreme of scale: deploying compact ML at billions of constrained endpoints to achieve low latency and reduced communication to the cloud. Recent surveys (e.g., Heydari et al., 2025) summarize techniques for on-device inference, model compression, and energy-aware design; these works argue that combining TinyML with hierarchical/federated strategies is essential for enabling real-time intelligence in large, geographically distributed sensor fleets. The literature highlights trade-offs between model size, accuracy, and system-level scalability.

Despite the significant advancements in ML applications, a clear understanding of which ML paradigms are most suitable for large-scale smart systems remains underdeveloped[4]. The current body of literature tends to be fragmented, with many studies focusing on specific ML categories such as supervised, unsupervised, reinforcement learning, deep learning, or hybrid approaches without offering a unified perspective. Moreover, existing reviews typically center on a single application domain, such as smart grids or smart transportation, resulting in limited insight into how different ML paradigms perform across diverse large-scale environments. As a result, researchers, engineers, and policymakers often lack a comprehensive framework for selecting ML approaches that are both scalable and robust for broad smart system applications.

This fragmentation highlights a clear research gap: while numerous ML techniques have been explored in isolated contexts, there is insufficient holistic comparison of their strengths, weaknesses, scalability considerations, and suitability for large-scale deployment. The absence of such comparative analysis challenges stakeholders seeking to design, implement, or improve smart systems that rely on ML-based intelligence. There is thus a pressing need for a systematic, cross-domain review that evaluates ML paradigms through a unified analytical lens.

The purpose of this study is to address these gaps by providing a comprehensive review and categorization of major ML paradigms used in large-scale smart systems. This review aims to analyze the strengths and limitations of each paradigm, evaluate their scalability and performance in large distributed environments, and identify patterns and trends that influence ML effectiveness across application domains. Furthermore, the study seeks to uncover future research directions that can support the development of more scalable, efficient, and adaptive ML solutions for next-generation smart systems. Through this work, the study offers a structured foundation for guiding future ML integration in complex, data-intensive smart environments.

2. Research Methodolgy

This study employs a comprehensive systematic literature review approach to examine machine learning paradigms used in large-scale smart systems. Given the breadth of research across domains such as smart cities, smart grids, smart healthcare, and IoT-based infrastructures, a structured review method was necessary to ensure that relevant studies were identified, screened, and analyzed consistently. The methodology follows established guidelines for evidence-based review studies, focusing on transparency, reproducibility, and methodological rigor.

The literature search was conducted across multiple reputable academic databases to ensure broad and reliable coverage of scientific publications[5]. These databases include Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Elsevier. These sources were chosen because they collectively index a wide range of journals, conference proceedings, and technical reports in computer science, artificial intelligence, data analytics, and smart system technologies. Searching across these platforms minimizes the possibility of publication bias and ensures that the review captures both foundational and recent advancements in machine learning applications for large-scale environments.

A systematic search strategy was implemented using combinations of predefined keywords and Boolean operators[6]. Keywords included “machine learning,” “large-scale smart systems,” “smart city analytics,” “IoT analytics,” “large-scale distributed systems,” “deep learning frameworks,”

"reinforcement learning," "graph neural networks," and "edge intelligence." These terms were selected to reflect the primary themes of the research topic while capturing the diversity of ML paradigms and their applications. Boolean operators such as AND, OR, and NOT were used to refine the searches and filter out irrelevant literature. Searches were conducted iteratively to incorporate emerging publications relevant to the review's objectives.

To ensure consistency and relevance, a set of inclusion and exclusion criteria was applied. The inclusion criteria required that studies be published between 2015 and 2025, appear in peer-reviewed journals or high-quality conference proceedings, and explicitly discuss the application of machine learning methods in large-scale smart system contexts. Only studies written in English and containing empirical, methodological, or conceptual contributions were considered[7]. Conversely, exclusion criteria removed publications that were non-English, outdated, unverified preprints without adequate review status, and studies limited to small-scale laboratory experiments not generalizable to real-world large-scale systems. Editorials, opinion papers, and non-scientific articles were also excluded to maintain academic rigor.

The selected papers were then analyzed using a structured framework to extract relevant information[8]. The analysis began with the categorization of studies based on machine learning paradigms, such as supervised learning, unsupervised learning, reinforcement learning, deep learning, graph-based models, federated learning, and hybrid methods. For each study, critical details were recorded, including algorithm type, data scale, system architecture, and application domain. Next, key performance metrics such as accuracy, latency, energy efficiency, model complexity, and system scalability were extracted to identify strengths and limitations across paradigms. Special attention was given to evaluating the scalability of ML models, including their ability to operate under large data volumes, distributed infrastructures, and real-time constraints.

Finally, a comparative analysis was conducted to synthesize findings across studies. This step involved comparing how different ML paradigms perform in various large-scale smart system applications, identifying common challenges, and highlighting emerging trends and opportunities. Through this structured methodology, the review not only summarizes existing knowledge but also provides a coherent understanding of the suitability, performance, and scalability of machine learning paradigms in modern smart systems.

3. Results and Discussion

Classification of ML Paradigms for Smart Systems

The findings of this review reveal that machine learning (ML) paradigms used in large-scale smart systems fall into several major categories, each offering distinct advantages and limitations depending on the operational requirements, data characteristics, and system architecture. These paradigms demonstrate a diverse spectrum of capabilities from predictive modeling and pattern recognition to real-time decision-making and distributed intelligence highlighting the multidimensional nature of smart system development. The discussion below synthesizes the major ML categories and evaluates their performance, scalability, and suitability across various smart system environments.

Supervised learning remains one of the most widely adopted paradigms in large-scale smart systems, particularly for tasks requiring accurate prediction, classification, and regression[9]. Applications such as traffic flow prediction, energy demand forecasting, medical diagnosis, and intrusion detection rely heavily on supervised models due to their strong predictive capability. Algorithms like Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines perform well when sufficient labeled data is available. However, supervised learning faces challenges in large-scale contexts, especially due to high data labeling costs, concept drift in dynamic environments, and scalability limitations when models must be frequently retrained to accommodate new data streams. Despite these limitations, supervised learning remains a foundational component of many smart system architectures.

Unsupervised learning plays an essential role in analyzing unlabeled, heterogeneous, and high-dimensional data commonly found in IoT ecosystems and distributed smart networks. Techniques

such as clustering, dimensionality reduction, and anomaly detection allow systems to uncover hidden patterns, detect abnormal behaviors, and group similar data points without requiring human intervention[10]. In large-scale smart systems, unsupervised learning is particularly valuable for fault detection in power grids, behavioral clustering in smart cities, and unsupervised feature extraction in sensor-heavy environments. Nevertheless, these methods sometimes struggle with interpretability and may exhibit reduced performance when handling noisy or extremely large datasets without careful preprocessing or model tuning.

Reinforcement learning (RL) has gained prominence for its capability to support adaptive decision-making in environments where system states evolve continuously. RL models especially in their deep learning-enhanced forms have been widely applied in smart energy management, autonomous vehicles, resource allocation, and dynamic scheduling. RL's strength lies in its ability to learn optimal policies through interaction with an environment, making it suitable for real-time, sequential decision-making[11]. However, RL's performance in large-scale systems is often hindered by its high computational demands, slow convergence, and sensitivity to environmental unpredictability. In complex systems with multi-agent interactions, techniques such as Multi-Agent Reinforcement Learning (MARL) have emerged, but they remain challenging to scale reliably.

Deep learning (DL) represents one of the most transformative ML paradigms for large-scale smart systems due to its exceptional ability to process complex, high-dimensional data such as images, audio, sensor readings, and streaming data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers have significantly advanced applications like computer vision in smart surveillance, speech recognition in smart assistants, and time-series forecasting in smart transportation. Deep learning's hierarchical feature learning provides robustness to noisy and heterogeneous data, though these benefits come with high computational costs[12]. In large-scale deployments, DL models may suffer from latency issues, energy constraints, and difficulties in integrating with distributed edge devices, prompting research into lightweight or compressed models.

Hybrid and federated learning approaches have gained momentum as solutions to issues of scalability, data privacy, and distributed intelligence. Hybrid ML systems combine multiple learning paradigms such as integrating deep learning with reinforcement learning or blending supervised and unsupervised components to leverage their complementary strengths. Federated learning (FL), in particular, enables collaborative model training across distributed devices without centralizing data, making it highly suitable for privacy-sensitive applications in healthcare, transportation, and smart homes. Despite these advantages, FL faces challenges such as non-uniform data distribution, communication overhead, and model synchronization problems, especially in large heterogeneous networks.

Edge-based and distributed ML models address the pressing need for real-time processing and reduced dependence on centralized cloud servers. By performing computation directly on edge devices or in hierarchical layers between edge and cloud environments, these models significantly reduce latency, improve scalability, and enhance system resilience. Edge ML is crucial for applications requiring immediate response, such as industrial automation, wearable devices, and autonomous navigation. However, edge devices often have limited processing power and memory, necessitating the development of optimized, lightweight models such as TinyML, quantized neural networks, and pruning-based techniques. Distributed ML systems attempt to balance computational loads across many devices, but issues such as communication bottlenecks and system orchestration remain active research challenges.

Graph-based models, particularly Graph Neural Networks (GNNs), have emerged as powerful tools for representing and learning from relational and interconnected data structures. Many large-scale smart systems such as power distribution networks, transportation systems, and sensor communication networks naturally form graph topologies. GNNs leverage these structures to improve prediction accuracy, enhance anomaly detection, and support topology-aware optimization. Their capacity to capture spatial and relational dependencies makes them highly effective for modeling interactions within complex infrastructures. Nonetheless, GNNs may face scalability issues when

dealing with extremely large graphs, leading to ongoing research into sampling-based, distributed, and hierarchical GNN methods to reduce computational complexity.

Overall, the classification and comparative analysis of ML paradigms demonstrate that no single approach is universally superior for all large-scale smart system applications. Instead, the suitability of a paradigm depends on the specific system requirements, data availability, real-time constraints, and computational resources. The increasing integration of hybrid, federated, and graph-based models suggests a shift toward more adaptive, scalable, and context-aware ML solutions for next-generation smart systems.

Findings About Performance

The analysis of existing literature reveals that accuracy remains one of the most widely reported performance indicators for evaluating machine learning models in large-scale smart systems. Supervised learning and deep learning models generally achieve high accuracy in tasks such as demand forecasting, intrusion detection, image recognition, and anomaly classification[13]. Models such as CNNs, gradient boosting, and ensemble techniques consistently outperform simpler methods when trained on large, well-labeled datasets. However, high accuracy often comes at the cost of computational complexity, especially in deep neural networks that require substantial training time and memory. In dynamic environments where data distribution changes rapidly, performance can degrade unless models are frequently retrained or supported by adaptive learning mechanisms. This highlights the need for more robust, accuracy-driven models that maintain performance across varying conditions.

Scalability emerges as a critical challenge for ML paradigms applied to smart systems, especially those involving large IoT deployments, dense sensor networks, and distributed infrastructures[14]. Centralized ML models often struggle to scale due to bottlenecks in communication, storage, and computation. Federated learning, distributed training frameworks, and graph-based compression techniques provide promising solutions, enabling large-scale models to be trained more efficiently across heterogeneous devices. Yet, scalability remains uneven across paradigms; while GNNs and RL models demonstrate strong potential, they also introduce higher computational costs as graph sizes or state-action spaces increase. The literature consistently underscores the importance of models that can scale both horizontally (across devices) and vertically (with increasing data volume).

In terms of real-time capability, the findings indicate substantial variation among ML paradigms. Edge-based ML and TinyML models exhibit superior performance in latency-sensitive applications because they process data locally and reduce dependence on cloud-based computation. Conversely, deep learning architectures particularly complex CNNs and transformers often struggle with real-time constraints due to their heavy computational demands. Reinforcement learning models designed for real-time control in smart grids or autonomous vehicles face convergence delays that hinder deployment unless optimized or paired with approximate or surrogate models. Overall, real-time performance is strongly influenced by hardware constraints, model structure, and the degree of required responsiveness.

Energy efficiency is another significant factor shaping ML deployment in smart systems, particularly in battery-powered or resource-limited devices[15]. Traditional deep learning models are resource-intensive, consuming substantial energy during both training and inference. This restricts their use on low-power devices unless model compression techniques such as pruning, quantization, and knowledge distillation are applied. Federated learning can save energy by limiting data transmission but may increase computational load at the device level. Findings indicate that edge ML solutions, lightweight neural networks, and adaptive model-switching strategies offer the best trade-off between model complexity and energy consumption, making them suitable for long-term deployment.

Regarding robustness to noise, unsupervised learning and deep architectures especially autoencoders and GNN-based anomaly detectors demonstrate strong capability in identifying irregular patterns in noisy or incomplete data. Smart environments often suffer from sensor failures, communication delays, and data corruption, making robustness a key requirement. While supervised

learning models may perform well under ideal conditions, their performance degrades quickly when exposed to noise unless robust feature engineering or adversarial defense methods are used. Reinforcement learning models, on the other hand, show moderate robustness but are highly sensitive to the design of reward functions and environmental stochasticity.

Finally, adaptability in dynamic environments is essential for smart systems that operate under continuously changing conditions, such as fluctuating energy loads, evolving traffic patterns, and mobile user behavior. Reinforcement learning and online learning models excel in this area due to their ability to update policies and parameters based on ongoing interactions[16]. Federated learning also enhances adaptability by aggregating diverse local updates from distributed nodes, making the resulting model more reflective of real-world variability. Deep learning models, although powerful, typically lack inherent adaptability unless supported by continual learning or domain adaptation techniques. Across the literature, adaptability emerges as a defining characteristic of modern ML paradigms designed for real-world smart system deployment, underscoring the need for models that can evolve alongside their environments.

Application-Specific Findings

a. Smart Grid Systems: Reinforcement Learning for Dynamic Load Balancing
   In smart grid environments, electrical demand fluctuates continuously due to consumer behavior, weather conditions, and the integration of renewable energy. Traditional optimization models often struggle to make real-time decisions under uncertainty. Reinforcement Learning (RL) provides a significant performance advantage in this domain because it allows the system to learn optimal load-balancing strategies through continuous interaction with the grid environment. RL enables smart grids to dynamically adjust power distribution, reduce energy losses, and prevent overloads[17]. Algorithms such as Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and multi-agent RL excel at coordinating distributed energy resources. Empirical studies show that RL achieves faster adaptation to peak loads and enhances grid resilience during unexpected disturbances. This demonstrates that RL is especially effective when the system requires autonomous, real-time control with minimal human intervention.
b. Smart Transportation: Deep Learning for Demand Prediction and Traffic Optimization
   Smart transportation systems rely heavily on accurate demand forecasting, route optimization, and congestion management. Deep Learning (DL) models have proven superior for capturing the complex temporal and spatial patterns that characterize traffic flow and transportation demand[18]. Recurrent Neural Networks (RNN), LSTM, CNN-LSTM hybrids, and Transformer-based architectures are widely adopted because they can model long-term dependencies and nonlinear relationships in mobility data. In large-scale urban mobility networks, DL improves prediction accuracy for ride-hailing demand, public transport scheduling, and traffic incident detection. These models also allow transportation authorities to preemptively allocate resources, manage traffic lights, and adaptively route vehicles. Consequently, DL contributes to reduced congestion, lower travel times, and more efficient transportation infrastructure, illustrating its strengths in handling large, heterogeneous datasets.
c. Smart City Network Management: Graph Neural Networks for Interconnected Sensor Systems
   Smart cities deploy massive networks of heterogeneous sensors for monitoring environmental conditions, public infrastructure, mobility, energy systems, and public safety. Since these sensors are inherently interconnected, traditional ML approaches fail to fully exploit the relational information within the data. Graph Neural Networks (GNNs) are especially well-suited for modeling this type of connectivity. GNNs outperform conventional models because they can learn how information flows across a network and capture complex spatial correlations. Applications include air-quality monitoring, water distribution system analysis, traffic network modeling, and infrastructure fault detection. In large-scale deployments, GNNs have shown higher robustness, superior noise tolerance, and improved detection accuracy for anomalies in sensor networks. These results indicate that GNNs are the most effective ML paradigm for any smart system characterized by interconnected or graph-structured data.

d. Smart Healthcare: Hybrid Machine Learning for High-Precision Anomaly Detection
Smart healthcare systems rely on continuous patient monitoring, medical imaging, wearable sensors, and electronic health records[19]. The high dimensionality and sensitivity of medical data require models that are both accurate and interpretable. Hybrid ML approaches combining DL, statistical learning, and rule-based methods provide the best performance in this setting. For example, integrating CNNs with decision trees or LSTM with probabilistic models can significantly improve anomaly detection in vital signs, disease prediction, and early warning systems for critical conditions. Hybrid models reduce false positives, enhance timely detection, and improve patient outcomes. They also allow the system to incorporate expert knowledge (such as clinical guidelines) while still benefiting from the predictive power of modern ML architectures. As a result, hybrid ML is becoming the preferred approach for complex healthcare data environments, where both precision and reliability are paramount.

Technological Challenges

Large-scale smart systems face a series of complex technological challenges that significantly influence the performance, reliability, and scalability of machine learning applications. One of the most persistent issues is data heterogeneity. Smart systems often integrate data from numerous sources sensors, devices, platforms, and services all operating under different standards, formats, and sampling rates. This diversity complicates data preprocessing, fusion, and synchronization. Machine learning models frequently struggle to generalize across non-uniform datasets, leading to inconsistencies in system behavior. As smart environments evolve, the challenge of harmonizing continuously growing, multimodal data streams becomes even more critical[20].

Another major issue is the high computational cost required to train and deploy advanced ML models, particularly deep learning and reinforcement learning paradigms. Large-scale smart systems generate massive volumes of real-time data, demanding continuous retraining or adaptation of models to maintain accuracy. This requires substantial GPU or TPU resources, which may not be feasible for all organizations, especially in resource-constrained regions. Moreover, distributed systems such as smart grids, autonomous vehicles, and healthcare monitoring networks often operate with strict latency requirements, making it difficult to balance computational load and real-time responsiveness.

Smart systems also experience communication bottlenecks, especially when deployed across geographically distributed or densely networked environments[21]. As devices transmit large volumes of high-frequency data, network congestion becomes inevitable. Limited bandwidth, unstable connectivity, and latency spikes can cause delays in data transmission, reducing the efficiency of ML-driven decision-making. Communication bottlenecks are particularly problematic for applications that rely on coordinated actions or real-time analytics, such as traffic control or emergency response systems. These issues highlight the need for advanced edge computing and communication-efficient algorithms.

A further challenge involves model interpretability, which becomes increasingly important as smart systems influence decisions that affect safety, resource allocation, and public welfare. Many powerful ML models, especially deep learning architectures, operate as black boxes, providing limited insight into how they generate predictions[22]. Lack of interpretability undermines trust, limits regulatory compliance, and complicates error diagnosis. In domains like healthcare or energy management, transparent decision-making is essential. Developing interpretable or explainable AI methods remains an ongoing research priority, but achieving a balance between model accuracy and interpretability continues to be difficult.

In addition, privacy and security concerns pose serious risks to smart systems, which frequently collect sensitive data such as personal information, mobility patterns, health metrics, and energy usage behavior. Cyberattacks including data breaches, model poisoning, adversarial attacks, and network intrusions can compromise system integrity and endanger users. Ensuring robust authentication, encryption, and secure communication protocols is crucial, yet many smart devices lack sufficient protection due to limited hardware capabilities or outdated firmware. Privacy regulations such as

GDPR further complicate data handling and model training, demanding strict compliance mechanisms.

Lastly, deploying ML solutions in a distributed environment introduces significant operational challenges. Training, updating, and maintaining ML models across numerous devices or nodes requires careful coordination. Variations in computational capacity, energy availability, and network stability make uniform deployment nearly impossible. Furthermore, distributed systems are prone to synchronization issues, heterogeneous hardware failures, and inconsistencies in local model updates. Federated learning and edge AI offer potential solutions, but their implementation remains complex and costly. These factors collectively hinder the seamless deployment and long-term sustainability of ML-enabled smart systems.

Comparative Discussion

Machine learning paradigms offer distinct strengths when applied to large-scale smart systems, and comparing them reveals why certain approaches outperform others under specific conditions. Deep learning has emerged as a dominant paradigm due to its superior ability to handle high-dimensional, unstructured, and multimodal data[23]. Smart systems especially in transportation, city surveillance, and healthcare generate vast streams of images, signals, sensor logs, and text, all of which require sophisticated feature extraction. Deep neural networks automatically learn hierarchical representations, making them highly effective for tasks such as traffic flow prediction, medical imaging analysis, and environmental monitoring. Unlike traditional supervised learning models, deep learning does not require manual feature engineering, allowing it to scale efficiently as data complexity increases. This capability explains why DL consistently outperforms classical ML in environments characterized by large, rapidly evolving datasets.

In contrast, reinforcement learning (RL) is particularly suited for autonomous control, sequential decision-making, and real-time optimization. RL excels where agents must learn optimal actions through trial-and-error interactions with their environment, such as in smart grids, autonomous vehicles, or dynamic resource allocation systems[24]. These applications involve continuous adaptation to changing conditions peak energy demand, traffic disruptions, or environmental fluctuations which rule-based or supervised techniques struggle to handle. RL's ability to learn policies that maximize long-term rewards makes it invaluable for systems requiring sustained operation under uncertainty. Moreover, advanced RL variants like deep reinforcement learning further enhance adaptability in complex, high-dimensional state spaces.

Federated learning (FL) addresses one of the most critical barriers in large-scale smart systems: privacy and data governance. Many smart environments, such as healthcare networks, financial systems, and smart homes, cannot share raw data due to confidentiality concerns or legal restrictions. FL enables collaborative model training without transferring sensitive data to central servers. This paradigm preserves privacy, reduces data exposure, and minimizes communication overhead by updating models locally and aggregating only the learned parameters. For distributed systems with heterogeneous devices, such as IoT sensor networks, FL provides an efficient and scalable framework that aligns with modern privacy regulations like GDPR and HIPAA. Its ability to operate across diverse hardware environments makes it a cornerstone for future decentralized learning architectures.

Meanwhile, graph-based models, particularly Graph Neural Networks (GNNs), demonstrate exceptional performance in smart systems characterized by interconnected, relational, and non-Euclidean structures. Smart cities, transportation networks, communication infrastructures, and power grids inherently form graph-like topologies[25]. Traditional ML models assume independence among inputs, making them unsuitable for capturing the complex dependencies and interactions within these systems. GNNs, however, are designed to exploit relational structures by learning node, edge, and graph-level representations[26]. This enables more accurate prediction of traffic congestion, energy demand propagation, anomaly detection in sensor networks, and resilience analysis of infrastructure systems. Their ability to model spatial, temporal, and structural dependencies makes GNNs uniquely suited to represent the interconnected nature of large-scale smart environments.

Overall, these comparative insights highlight that no single ML paradigm universally dominates across all smart system scenarios. Instead, each paradigm offers unique advantages: deep learning for high-dimensional feature learning, reinforcement learning for continuous autonomous decision-making, federated learning for privacy-preserving distributed training, and graph-based models for networked data representations. The optimal choice depends on system requirements, computational constraints, and the structure of the data involved[27].

4. Conclusion

This review concludes that different machine learning paradigms offer distinct advantages for large-scale smart systems, with no single approach dominating all scenarios. Deep learning performs best for high-dimensional and unstructured data, while reinforcement learning is ideal for autonomous, real-time decision-making in dynamic environments. Federated learning provides strong privacy protection for distributed IoT systems, and graph-based models are most effective for applications involving interconnected networks such as smart cities and communication infrastructures. The review contributes a unified comparison of major ML paradigms, mapping their strengths, limitations, and scalability characteristics across multiple smart system domains. It also identifies key challenges including data heterogeneity, computational demands, and privacy risks that influence real-world deployment. Practically, the findings support system designers, engineers, policymakers, and urban planners in selecting appropriate ML models for specific smart-system needs. Despite limitations related to literature scope, emerging techniques, and variability in evaluation metrics, the review provides a comprehensive and timely synthesis that can guide both future research and the development of intelligent, efficient, and secure smart systems.

References

[1] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *Int. J. Eng. Bus. Manag.*, vol. 11, p. 1847979019890771, 2019.

[2] N. Baker *et al.*, "Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence," USDOE Office of Science (SC), Washington, DC (United States), 2019.

[3] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, "Approaches to cross-domain sentiment analysis: A systematic literature review," *Ieee access*, vol. 5, pp. 16173–16192, 2017.

[4] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson, "Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions," *Inf. Softw. Technol.*, vol. 127, p. 106368, 2020.

[5] M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources," *Res. Synth. Methods*, vol. 11, no. 2, pp. 181–217, 2020.

[6] W. M. Bramer, G. B. De Jonge, M. L. Rethlefsen, F. Mast, and J. Kleijnen, "A systematic approach to searching: an efficient and complete method to develop literature searches," *J. Med. Libr. Assoc. JMLA*, vol. 106, no. 4, p. 531, 2018.

[7] D. S. R. Vosgerau and J. P. Romanowski, "Review studies: conceptual and methodological implications," *Rev. Diálogo Educ.*, vol. 14, no. 41, pp. 165–189, 2014.

[8] Z. Yu, B. C. M. Fung, and F. Haghighat, "Extracting knowledge from building-related data—A data mining framework," in *Building simulation*, Springer, 2013, pp. 207–222.

[9] Q. Chen *et al.*, "A survey on an emerging area: Deep learning for smart city data," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 3, no. 5, pp. 392–410, 2019.

[10] M. Fahim and A. Sillitti, "Anomaly detection, analysis and prediction techniques in iot environment: A systematic literature review," *IEEE Access*, vol. 7, pp. 81664–81681, 2019.

[11] S. Ramstedt and C. Pal, "Real-time reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[12] G. Zhong, X. Ling, and L. Wang, "From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 1, p. e1255, 2019.

[13] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A

survey," *Appl. Sci.*, vol. 9, no. 20, p. 4396, 2019.
[14] A. El-Mougy, I. Al-Shiab, and M. Ibnkahla, "Scalable personalized IoT networks," *Proc. IEEE*, vol. 107, no. 4, pp. 695–710, 2019.
[15] J. Hempstead *et al.*, "Low-cost photodynamic therapy devices for global health settings: Characterization of battery-powered LED performance and smartphone imaging in 3D tumor models," *Sci. Rep.*, vol. 5, no. 1, p. 10093, 2015.
[16] X. Bai, J. Guan, and H. Wang, "A model-based reinforcement learning with adversarial training for online recommendation," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[17] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, 2018.
[18] S. Wang, J. Cao, and S. Y. Philip, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3681–3700, 2020.
[19] M. M. Baig and H. Gholamhosseini, "Smart health monitoring systems: an overview of design and modeling," *J. Med. Syst.*, vol. 37, no. 2, p. 9898, 2013.
[20] Q. Cai, H. Wang, Z. Li, and X. Liu, "A survey on multimodal data-driven smart healthcare systems: approaches and applications," *IEEE Access*, vol. 7, pp. 133583–133599, 2019.
[21] Y. Tsado, D. Lund, and K. A. A. Gamage, "Resilient communication for smart grid ubiquitous sensor network: State of the art and prospects for next generation," *Comput. Commun.*, vol. 71, pp. 34–49, 2015.
[22] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.
[23] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, 2020.
[24] V. Shah, "Reinforcement learning for autonomous software agents: Recent advances and applications," *Rev. Esp. Doc. Cient.*, vol. 14, no. 1, pp. 56–71, 2020.
[25] C.-H. Lee *et al.*, "Building a generic platform for big sensor data application," in *2013 IEEE International Conference on Big Data*, IEEE, 2013, pp. 94–102.
[26] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. neural networks Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2020.
[27] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt, "A review of modern computational algorithms for Bayesian optimal design," *Int. Stat. Rev.*, vol. 84, no. 1, pp. 128–154, 2016.