International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Recurrent Neural Network (RNN) Inference Cloud Computing

Deepshikha Saikia[1], Nihar Pratim Deka[2], Parusitom Brahma[3]

[1,2,3]Assam Downtown University

Abstract

The purpose of this technical article is to investigate the convergence of Recurrent Neural Networks (RNNs) with Cloud Computing considering the potential benefits and challenges that may be encountered in bringing these two systems together. Recurrent Neural Networks (RNNs) constitute a particular type of artificial neural network that is capable of processing data sequentially in such a way that it becomes well-suited for applications that involve handling time series or language processing. On the other hand, cloud computing provides computer infrastructures on a scalable basis as well as flexible computation services that are accessible whenever they are needed. This paper targets making RNN applications more efficient and faster with the use of RNN libraries on cloud computing platforms through which the cloud services can be used for things like data processing power and storage. By combining RNNs with Cloud Computing, this paper aims to enhance the efficiency and performance of RNN-based applications by leveraging the cloud's computational power and storage capabilities. The study investigates the impact of deploying RNN inference tasks in the cloud environment, analyzing factors such as latency, cost-effectiveness, and scalability.

Introduction

Recurrent neural networks (RNNs) have attracted a lot of attention across several domains due to their proficiency in effectively modeling sequential data. Despite being efficient in this regard, the high computation cost of training and deploying these models is often significant, more so if one has large datasets or intricate architectures. Cloud Computing provides a solution to this challenge by offering a scalable and cost-effective platform for running computationally intensive tasks. By offloading RNN inference tasks to the cloud, organizations can benefit from the cloud's vast computational resources without the need for extensive on-premises infrastructure.

Recurrent Neural Networks (RNNs) are a type of artificial neural network built to process sequences of data with the ability to recall the last inputs. In activities such as natural language understanding, spoken messages, and time analysis, this distinctive design makes RNNs outstanding. They are capable of retaining memory over directions which is not the case in normal feedforward neural networks. This capability makes RNNs well-suited for tasks where context and order are crucial for accurate predictions. The diagram below represents the various applications of RNNs.

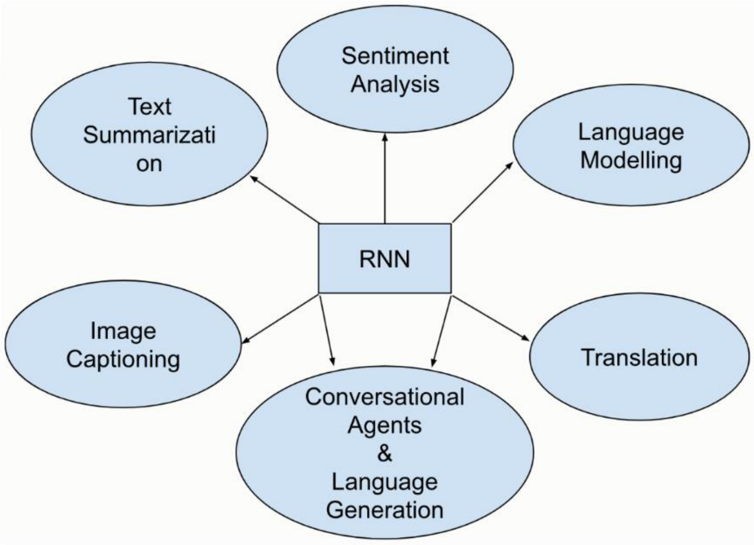


Fig 1: Applications of RNNs

Cloud Computing, on the other hand, revolutionizes the way computational resources are provisioned and utilized. Such an organization can use virtualized resources, including computing capacity, storage, and networking on-demand, through cloud computing. As compared to other solutions that cost lots of money, like fixed capital costs and huge operating expenses such as utility bills, among others when looking at the benefits cloud computing provides cheap solutions that are however effective in optimizing IT infrastructure hence reducing operational costs. This means it can be trusted if they are among those in need of accomplishing their goals in a short duration because it is one technology that helps people handle many or different tasks simultaneously. The diagram below represents the various characteristics of cloud computing.

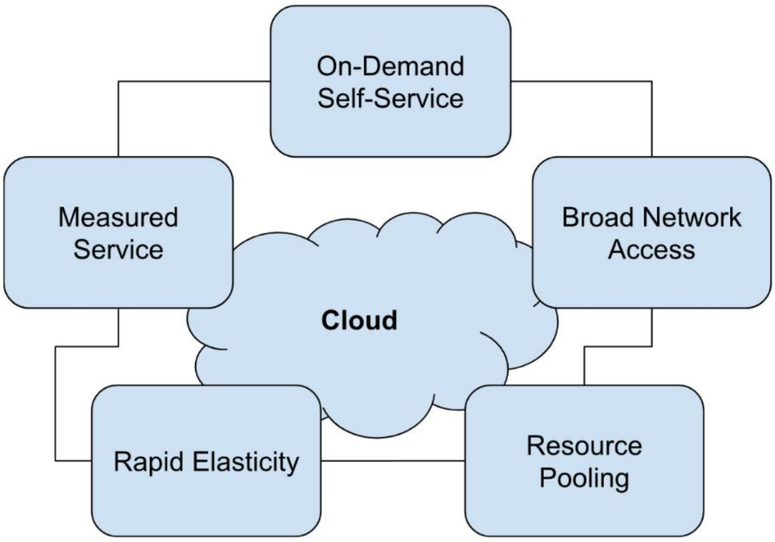


Fig 2: Characteristics of Cloud Computing

When RNNs are combined with Cloud Computing, they provide an exciting chance to boost RNN applications' performance, as well as their scaling abilities in a way that has never been done before.

This involves using cloud-based computing power for deploying models created under RNNs which in turn enables organizations to exploit enormous computational power in the cloud to increase model training and inference speeds. It creates an environment where such researchers work on intricate issues that require intensive data processing which would be hard or impossible without their merger.

This study aims to explore the coming together of RNNs and Cloud Computing by investigating the consequences of executing RNN inference tasks within a cloud environment. It thus seeks to give insight into the way forward with regard to improving efficiency and productivity when used in different applications through understanding various performance metrics, cost-effectiveness, and scalability of deploying RNN models in the cloud environment.

Methodology

We conducted an organized investigation in our review article to find out which research articles were relevant. We systematically searched various databases, including Science-Direct, IEEE Xplore, Springer, MDPI, and Google Scholar, among others. By using the keywords “Recurrent Neural Network (RNN) Inference Cloud Computing” and “Deployment of Recurrent Neural Network (RNN) in Cloud” we carefully scrutinized references that were found during our search process. Our criteria for inclusion were stringent: selected papers had to primarily focus on RNN inference tasks within cloud environments, be written in English, undergo peer review, and be accessible in full text. Every chosen article was thoroughly examined. We extracted important details including title, authors' name, publication year, methodologies employed, and main findings from those details. The careful procedure enabled a detailed comparison of the selected sets of papers and an investigation. Furthermore, a drawing showing the flow of our activity starting from the first paper identification up to the final selection stage was drawn.

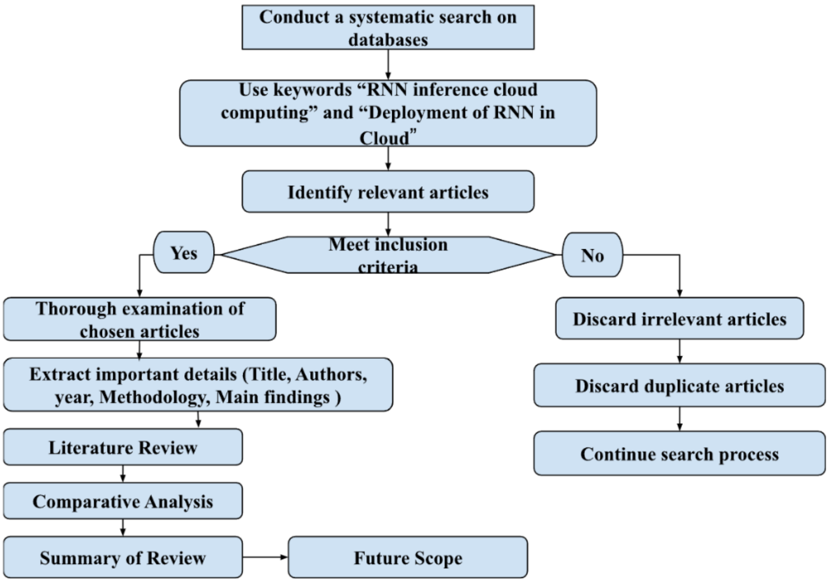

Conduct a systematic search on databases
Use keywords “RNN inference cloud computing” and “Deployment of RNN in Cloud”
Identify relevant articles
Meet inclusion criteria
Yes
No
Thorough examination of chosen articles
Extract important details (Title, Authors, year, Methodology, Main findings)
Literature Review
Comparative Analysis
Summary of Review
Future Scope
Discard irrelevant articles
Discard duplicate articles
Continue search process

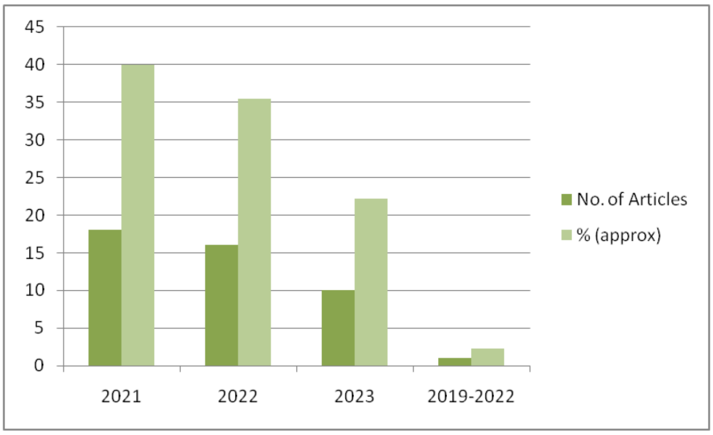
Fig 3: Workflow block diagram

Figure 3 depicts a succinct visual representation of our systematic review approach, which significantly enhances the clarity and transparency of our technique. This process block diagram clearly depicts the progression of papers from initial identification through examination to final inclusion, ensuring reliability and providing a clear understanding of our method.

Table 1: Article distribution by journal title

Journal Title	**No. of Articles**	**% (approx)**
IEEE	14	31.1
Science Direct	6	13.3
Springer	9	20
MDPI	15	33.3
Others	1	2.22
Total	45	100

Chart 1: Article distribution by journal year



Literature Review

The author Pillareddy Vamsheedhar Reddy and other authors (2023) of the research study [1] presents a multi-objective scheduling framework for cloud computing environment resource optimization. The approach aims to simultaneously handle numerous conflicting goals, improving performance and resource utilization. To improve generative AI applications such as deep learning models for natural language processing and image generation [2], YUN-CHENG WANG and other authors (2023) investigates how edge and cloud computing can be used. In research paper [3], the author Sowmya Sanagavarapu and others (2021) presents SDPredictNet, a novel framework for Software-Defined Networking (SDN) that improves network performance, efficiency, and reliability in SDN environments by utilizing traffic prediction analysis and neural routing algorithms.This will improve the large-scale deployment and maintenance of these models efficiently. Research [4], the author Dominic Karnehm and others evaluates and contrasts deep learning methods to estimate the state-of-charge of lithium-ion batteries in cloud-based battery management systems, evaluating how well they function and how long they last. To increase the efficacy and accuracy of automated speech recognition (ASR) systems, the author Chao-Han Huck Yang and others (2021) of the study [5] investigate a novel approach that decentralizes feature extraction using a quantum convolutional neural network (QCNN).

In research [6], the author Phuc Trinh Dinh and others (2021) provides a novel approach to detect Economic Denial of Sustainability (EDoS) assaults in cloud environments based on Software-Defined Networking (SDN) and using Generative Adversarial Networks (GANs) to address economic vulnerabilities instead of resource depletion. Once more, a new study [7], the author Zeinab Khodaverdian and others (2021) suggests an efficient way to choose virtual machines (VMs) for cloud migration using a Shallow Deep Neural Network (SDNN), which lowers energy consumption. In cloud data centers, this method lowers energy consumption and enhances resource allocation. The author Anil Verma and others (2021) present an IoT-influenced smart monitoring and reporting system for Education 4.0 in the paper [8], which improves decision-making by providing real-time insights and data analysis. To enhance resource allocation, performance, and scalability, the research [9] once more by author JORGE ARIZA and others (2021) suggests using deep learning techniques to optimize computing resources in cloud-based e-learning platforms. According to the study [10], by JEFFREY C. KIMMELL and other authors (2021) examining online behavior patterns, recurrent neural networks (RNNs) can be used to detect malware in cloud systems. By enabling the real-time identification of malware threats in cloud systems, this strategy seeks to improve security.

To improve system resistance, in research paper [11], the author PHUC TRINH DINH and others (2021) presents a novel technique called R-EDoS, which employs a Stochastic Recurrent Neural Network to detect Economic Denial of Sustainability (EDoS) assaults in Software-Defined Networking (SDN) cloud configurations. In a different study [12], a novel method called SecFedIDM-V1 is presented by the author EMMANUEL BALDWIN MBAYA and others (2023). It combines deep bidirectional Long Short-Term Memory networks with blockchain technology to improve the security of federated intrusion detection. The designers of this approach want to increase the accuracy of intrusion detection and data sharing security in federated systems. A deep learning model to anticipate large-scale cloud application failures has been built by a study [13], author Mohammad S. Jassas and others (2021), to improve app availability, performance, and reliability by anticipating possible problems. The author POOYAN KHOSRAVINIA and others (2023) of a study [14] provide a novel approach to improving road safety through the use of Graph Convolutional Recurrent Networks (GCRNs) to identify unsafe driving behaviors. By increasing the accuracy of identifying dangerous acts, this strategy seeks to lower the number of traffic accidents and promote safer driving practices. Using a deep recurrent neural network-based Hadoop framework, the study [15], the author D.B. Jagannadha Rao and others (2023) offers a novel prediction model for COVID-19 outcomes. This framework combines deep learning techniques with distributed computing to enable broad data analysis in cloud environments. The authors of the technical research study [16] Rojalina Priyadarshini and other authors(2022) propose deep learning as a potential defense against Distributed Denial of Service (DDoS) attacks in fog computing environments. Fog computing, a kind of cloud computing, leverages decentralized processing resources closer to the network's edge. The work proposes an intelligent system that effectively detects and blocks DDoS attacks in foggy conditions by using deep learning techniques. A novel approach to offloading work in mobile edge computing environments is proposed by the authors of a technical research paper. [17], "A RNN based offloading scheme to preserve energy and reduce latency using RNNBOS". The authors C.

Anuradha and others(2022) introduces the RNNBOS (Recurrent Neural Network-Based Offloading Scheme), which optimizes job offloading decisions to reduce latency and save energy. In a study[18], the authors Shitharth Selvarajan and others(2023) describe a novel security model that blends artificial intelligence (AI) capabilities with lightweight blockchain technology to address the specific security and privacy challenges that IIoT systems face. The proposed architecture uses artificial intelligence (AI) for threat detection and blockchain for secure data transport and storage to provide a robust defense against cyberattacks while protecting sensitive data in industrial settings. The authors Amirah Alshammari and others(2021) of study [19] look into the application of machine learning (ML) techniques for the identification and mitigation of malicious network traffic in cloud computing settings. The goal of the project is to use machine learning techniques to strengthen security measures in cloud infrastructures, which are vulnerable to various cyber threats. A cloud-based platform for intelligent self-diagnosis and department referral is presented by the authors of the research article study [20] Junshu Wang and others (2021). It utilizes Chinese medical BERT, a specialized language model trained on medical text. Based on the user's medical history and symptoms, the system provides exact diagnosis ideas and recommends relevant medical departments for further consultation. It aims to increase the precision and effectiveness of healthcare services, particularly in places with limited access to medical information. The process for creating a faulty knowledge graph tailored for multi-cloud IoT systems is described by the authors in a study [21] Wenqing Yang and others(2022). It explains the process of extracting defect-related data from many clouds and Internet of Things devices, organizing it into a knowledge graph, and facilitating more effective defect identification, resolution, and knowledge sharing within the ecosystem. The study also demonstrates how defect management practices in multi-cloud IoT installations may be improved in real-world scenarios by utilizing the generated knowledge graph. The authors of the research study Guanming Bao & Ping Guo(2022) [22] investigate the integration of federated learning into cloud-edge collaborative systems. Federated learning is a decentralized machine learning technique that allows models to be trained across multiple decentralized edge devices, hence eliminating the requirement for central data aggregation. This article examines the technology that underpins federated learning, along with its various applications and challenges.The research paper was published in 2023 by Theodoros Theodoropoulos and other authors the study paper [23] presents the findings on the use of GNNs to characterize intricate relationships between various resources in situations of multiplayer mobile gaming. By utilizing the graph structure that is inherent in resource interactions, it proposes a novel approach to modeling and forecasting patterns in resource utilization. The article conducts experiments using real-world datasets from mobile gaming platforms to demonstrate how well the proposed technique captures and predicts resource usage behaviors. The author Syed Mohamed Thameem Nizamudeen(2023) of a technical research study[24] proposes a revolutionary intrusion detection system made for multi-cloud and Internet of Things (IoT) scenarios. It uses a swarm-based deep learning classifier to increase the accuracy and efficacy of intrusion detection in this dynamic and heterogeneous context.

The authors Heng Zhou and others(2021) according to a study [25], This paper presents a fresh approach to ironmaking process optimization, based on digital twin technologies from cloud computing platforms. Using real-time sensor data and historical data to create a digital portrait of the ironmaking process, the system provides intelligent insights and recommendations for process optimization. The cloud-based service enables users to access and utilize the optimization service remotely, boosting ironmaking processes' productivity and efficiency. An innovative intrusion detection model and algorithm for the

Internet of Things (IoT) environment suggested in this research study is presented by the authors in a study [26]. It uses cloud computing, multi-feature extraction techniques, and Extreme Learning Machine (ELM) for efficient intrusion detection. The authors Yanfang Yin and others(2021) of research paper [27] propose the use of an edge-computing architecture to detect dangerous behavior in the power field. The method collects real-time data by putting sensors at various points across the electrical grid. The data is then processed locally at the edge devices rather than being routed to a centralized server. This facilitates the quicker analysis and resolution of potential safety concerns. The paper discusses this strategy's advantages and disadvantages.

The authors Abhishek Sharma and others(2022) of study report [28], The research study proposes a novel approach for intelligent risk assessment in cloud computing systems by leveraging artificial intelligence (AI) and supervised machine learning techniques. The proposed model aims to enhance the security and reliability of cloud systems by accurately predicting and mitigating risks. The authors Moses Ashawa and others(2022)of study report [29], The retracted technical research study attempted to increase the effectiveness of cloud computing systems by using a novel approach for resource allocation and load balancing. The study recommended using the Long Short-Term Memory (LSTM) machine learning technique for demand prediction and resource allocation optimization in cloud environments. The authors Mofei Song and others(2021) examine a novel deep-label distribution learning technique for visibility estimations in cloud systems in a study [30]. To enhance user experience and optimize resource allocation in cloud computing, the authors provide a system that uses advanced machine learning algorithms to estimate visibility levels. The author Edgar Cortés Gallardo Medina , Victor Miguel Velazquez Espitia and others (2021) of the study paper [31] discusses how to use parallelization techniques, distributed cloud computing, and object detection to enhance autonomous driving systems. It focuses on the difficulties of real-time object recognition and detection for secure navigation in changing situations. Through the transfer of calculation jobs to distant servers, distributed cloud computing lowers the compute burden onboard. Techniques for parallelization improve processing effectiveness and speed. Using evolutionary algorithms and machine learning approaches, the article [32] authors Sania Malik , Muhammad Tahir and others (2022) offer a prediction model for resource utilization forecasting in cloud data centers. This model uses historical data and real-time monitoring indicators for accurate forecasts, which optimizes resource allocation and increases efficiency in operating cloud infrastructures. The authors

Sardar Khaliq uz Zaman, Ali Imran Jehangiri and others (2022) of the study[33] investigate a novel approach to leveraging Long Short-Term Memory (LSTM) to offload computations in Mobile Edge Computing (MEC). By doing so, it becomes easier to decide when to offload calculations, allowing work to be transmitted in advance to edge servers and cutting down on delays.

To minimize losses from oversupply and conserve energy, a deep learning-based method for optimizing computer resource utilization is presented in the research paper [34] authors Marius Cioca and Ioan Cristian Schuszter and others (2022) by projecting future resource requirements, this method makes sure computers use less electricity and last longer. In order to guarantee that management efforts are not in vain and that computer resources are used sustainably, it also recommends the use of renewable resources. The authors Saud Alzughaibi and Salim El Khediri (2023) of the paper [35] describe a cloud intrusion detection system (IDS) that makes use of Particle Swarm Optimization (PSO), Backpropagation, and Deep Neural Networks (DNN). The system is intended to precisely identify and classify intrusions in cloud settings, enhancing security protocols and mitigating any risks. It has been tested on the CSE-CIC-IDS2018 dataset.

Once more, the study paper [36] authors Ahmed R. Nasser ,Ahmed M. Hasan and others (2021) presents a novel strategy for diabetes monitoring in healthcare that makes use of IoT and cloud computing. A wearable gadget that is linked to a deep learning system in the cloud has been developed for the control of diabetes. The gadget gathers and evaluates user data to deliver personalized diabetes care suggestions and real-time insights. The research study in [37] authors Abdul Razaque, Nazerke Shaldanbayeva and others (2022) introduces a novel big data analytics-based approach to guard against unwanted access to cloud computing resources. Using network traffic, log data, and user behavior, the approach addresses the growing concern about unauthorized access and security breaches in cloud environments by identifying anomalous activity and proactively mitigating potential hazards. The author Anil Verma , Divya Anand and others (2022) of the study[38] presents an irregularity-detection system for Industry 4.0 and Education 4.0 that is inspired by the Internet of Things. It detects anomalies in industrial and educational processes using machine learning, analytics, and real-time data collection. By anticipating and resolving anomalies, the system seeks to improve security, efficacy, and dependability. In the study paper[39], the authors A Angel Nancy,Dakshanamoorthy Ravindran and others (2022) provide a state-of-the-art IoT-cloud smart healthcare monitoring system that forecasts cardiac disease using deep learning. The system employs predictive modeling, cloud analytics, and real-time data collection to identify cardiovascular problems early and provide vulnerable people with individualized intervention strategies.

Zaakki Ahamed , Maher Khemakhem and others (2023) investigate how deep learning techniques can be used to enhance workload prediction in cloud computing [40], with an emphasis on capacity planning, resource allocation, and overall performance optimization. It highlights the shortcomings of current approaches and stresses how crucial precise workload planning is to the best possible cloud setups. In the research paper [41], the authors Tom Danino, Yehuda Ben-Shimol and Shlomo Greenberg (2023) provides Multi-Agent Deep Reinforcement Learning (MADRL), a novel approach for enhancing container allocation in cloud environments. With the help of this framework, many agents can autonomously distribute containers in response to shifting workload requirements and resource availability, which maximizes resource efficiency and lowers costs. In another research [42] author Masoud Emamian , Aref Eskandari and others (2022) a new monitoring system for photovoltaic plants is introduced in a different study which makes use of cloud computing, IoT, and machine learning. By keeping an eye on variables including solar irradiance, panel temperature, and energy output, this technology improves plant efficacy. The authors Chrysostomos Symvoulidis,George Marinos and others (2022) of a study [43] present healthFetch, a user-focused prefetch solution intended to improve user happiness and data retrieval in cloud-based health storage. To increase speed and reliability, it makes use of influence-driven and context-specific techniques that take user behavior, external variables, and the importance of the data into account. In another research [44], author Ammar Aldallal (2022) suggests merging CNNs, RNNs, and other models with deep learning to improve intrusion detection systems (IDS). The goal is to increase scalability, reduce false alarms, and improve detection accuracy in computer network security. A MINI-PC equipped with deep learning and an AI-QSR system is combined by MiniDeep, an AI platform, in the study [45] author Yuh-Shyan Chen, Kuang-Hung Cheng and others (2022) to improve decision accuracy in stand-alone establishments such as quick-service restaurants. Its lightweight, portable design makes it suitable for a wide range of uses, including decisions unique to the fast food sector.

Comparative Analysis

The table below gives a brief summary of each paper included in our research, as well as shows the key techniques, features, challenges and a outcome dealt with:

Table 2: Article Title, Techniques, Features, Challenges and outcome of each article

Article Title	**Techniques**	**Features**	**Challenges**	**Outcome**
A Multi-Objective Based Scheduling Framework for Effective Resource Utilization in Cloud Computing [1]	Anova-Recurrent Neural Network (A-RNN), UUID-BLAKE hashing technique, MD-PAM, LS-CSO	Improved resource used, Versatile optimization	Complicated	Based on the result analysis, the proposed LS-SCO outperformed when compared with the algorithms CSO, PSO and RR has achieved better performance.
An overview on Generative AI at Scale with Edge-Cloud Computing [2]	GenAI and Edge-Cloud Computing	Faster processing, Efficient resource usage	Management challenges	This paper points out several future research directions, such as domain-specific GenAI models, decomposition of large lan-guage models, green GenAI models, etc.
SDPredictNet-A Topology based SDN Neural Routing Framework with Traffic Prediction Analysis[3]	SDPredictNet, Seq2Seq, Artificial Neural Network (ANN)	Improved Routing Precision	Prediction Accuracy	SDPredictNet has achieved a RMSE score of 0.07 and an accuracy of 99.88% for traffic estimation and subsequent path determination.

Comprehensive Comparative Analysis of Deep Learning-based State-of-charge Estimation Algorithms for Cloud-based Lithium-ion Battery Management Systems [4]	EKF, FNN, LSTM, and GRU	Improved Accuracy, Enhanced Adaptability	It requires a lot of processing power and memory, especially for complex neural network structures.	The study concludes that the EKF method is the fastest and most accurate among all considered methods.
DECENTRALIZING FEATURE EXTRACTION WITH QUANTUM CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATIC SPEECH RECOGNITION [5]	Quantum Convolutional Neural Network (QCNN), Vertical Federated Learning (VFL)	Faster processing, Noise tolerance	Complexity	The proposed QCNN encoder attains a competitive accuracy of 95.12% in a decentralized model, which is better than using centralized RNN models with convolutional features.
Economic Denial of Sustainability (EDoS) Detection using GANs in SDN-based Cloud [6]	MAD-GAN Model	Flexibility, Early response	Inaccurate results	This paper shows that the proposed scheme is very efficient in both terms of accuracy and resource consumption.
A Shallow Deep Neural Network for Selection of	Convolutional Neural Network (CNN) and	Energy efficiency, automated decision making	Performance overhead	According to the empirical results, the proposed model has higher

Migration Candidate Virtual Machines to Reduce Energy Consumption [7]	Gated Recurrent Unit (GRU)			classification accuracy compared to other existing models for selecting the migration candidate virtual machines.
IoT Inspired Intelligent Monitoring and Reporting Framework for Education 4.0 [8]	IFC (IoT, fog and cloud computing), bi-directional learning, multi-layered bi-directional long-short term memory network.	Improved monitoring	Complex implementation	This paper verified that the proposed framework is capable of performing better in comparison to other contemporary decision-making methods for delay efficiency, data classification, irregularity prediction, and system stability.
Provisioning Computational Resources for Cloud-Based e-Learning Platforms Using Deep Learning Techniques [9]	Networks (RNNs), Long Short-Term Memory (LSTM) and Bidirectional (BIDI)	Customized allocation of resources	Privacy and security issues	This paper achieves high accuracy. The obtained results are promising, paving the way towards constructing software tools for provisioning computational resources on demand for e-learning platforms.

Recurrent Neural Networks Based Online Behavioural Malware Detection Techniques for Cloud Infrastructure [10]	deep learning, recurrent neural network, cloud IaaS	Instant detection	Difficult to find diverse datasets	Both our LSTM and BIDI models achieve high detection rates over 99% for different evaluation metrics.
R-EDoS: Robust Economic Denial of Sustainability Detection in an SDN-Based Cloud Through Stochastic Recurrent Neural Network [11]	software-defined networking, deep learning, cloud computing, network function virtualization, service function chaining.	Improved detection accuracy	False positives	This paper's comprehensive analysis of the results shows outstanding performance.
SecFedIDM-V1: A Secure Federated Intrusion Detection Model With Blockchain and Deep Bidirectional Long Short-Term Memory Network [12]	Blockchain, intrusion detection, deep learning, recurrent neural network.	Secure data sharing, Robust intrusion detection	increased computational overhead and delays	From the evaluation results of the intrusion classifier, the 80:10:10 BiLSTM network performed better than GRU with a Precision of 0.99624, Recall of 0.99906, F1 Score of 0.99614, False Positive Rate (FPR) of 0.00094, False Negative Rate (FNR) of 0.00395 and True Positive Rate

				(TPR) of 0.99605.
A Failure Prediction Model for Large Scale Cloud Applications using Deep Learning [13]	ANN deep learning algorithm	Early detection	Risks of False positives	The evaluation results show that the proposed model achieved a high precision, recall and f1 score.
Enhancing Road Safety Through Accurate Detection of Hazardous Driving Behaviors With Graph Convolutional Recurrent Networks [14]	Graph Convolutional Long Short-Term Memory networks	Complete analysis, Precise identification, Live tracking	Limited data access	The proposed model achieved an accuracy of 97.5% for public sensors and an average accuracy of 98.1% for non-public sensors, which shows that the proposed model can produce consistent and accurate results for both scenarios.
Deep recurrent neural network-based Hadoop framework for COVID prediction with applications to big data in cloud computing [15]	cloud, deep belief network, DBN, deep recurrent neural network.	Enhanced prediction	Data quality	This proposed method obtained minimal MSE and RMSE of 0.0523 and 0.2287 by considering affected cases. By considering death cases, the proposed method achieved minimal MSE and RMSE of 0.0010, and 0.0323 and

				measured minimum MSE of 0.0049 and minimum RMSE of 0.0702 for recovered cases.
A deep learning based intelligent framework to mitigate DDoS attack in fog environment [16]	Software Defined Network (SDN), (DL)-based detection methods, network traffic analysis mechanisms	Improve detection accuracy, real-time response, scalability	Data dependency, Being source intensive, vulnerability to adversarial attacks	The model has experimented on different parameters to get a set of optimized performance tuners. LSTM with 3 hidden layers, one dense layer, 128 input nodes and where a dropout rate is 0.2 for all the hidden layers is giving a good performance indicator in terms of growing accuracy and reduced error rate.The model is showing 98.88% of accuracy on the testing data set.
A RNN based offloading scheme to reduce latency and preserve energy using RNNBOS [17]	RNNBOS (Recurrent Neural Network Based Offloading scheme), recurrent neural network (RNN)	Minimizing Delays, Saving Power, Flexibility	Complexity, Training Overhead, Privacy Concerns	the authors simulated the RNN-based offloading scheme using a Python tool and observed that the RNN-based offloading scheme is effective in executing

				applications in MCC.
An artificial intelligence lightweight blockchain security model for security and privacy in IIoT systems [18]	Artificial Intelligence-based Lightweight Blockchain Security Model (AILBSM), Convivial Optimized Sprinter Neural Network (COSNN) , Lightweight Consensus Proof-of-Work (LCPoW)	Improved Security, Privacy Protection, Increased Efficiency	Complexity, Scalability, Integration Chalanges	By using the AILBSM framework, the execution time is minimized to 0.6 seconds, the overall classification accuracy is improved to 99.8%, and the detection performance is increased to 99.7%.
Apply machine learning techniques to detect malicious network traffic in cloud computing [19]	Machine learning (ML)model using a dataset constructed from both malicious and normal network traffic	Improved Detection Accuracy, Immediate Response to Threats, Flexibility in the Face of Changing Threats	Complexity of Implementation, Data Privacy Concerns, Overhead and Resource Consumption	The study presents a reliable model running in Real-time to detect malicious data flow traffic depending on the ML supervised techniques based on the ISOT-CID dataset that contains network traffic data features. The challenge in this research is to capture the deviations between the data instances so; malicious and

				normal properties categorize the data.
Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT [20]	Chinese medical BERT (Bidirectional Encoder Representations from Transformers), The framework is deployed in a cloud computing environment using container and micro-service technologies	Improved Precision, Streamlined Operations, Remote Availability	Dependency on Data Quality, Privacy Concerns, Limited Scope	The outcome of the research paper indicates that the proposed model outperformed traditional deep learning models and other pre-trained language models in terms of performance
Defect knowledge graph construction and application in multi-cloud IoT [21]	ontology design, fusion algorithms, and a knowledge graph reasoning method called GRULR (Gate Recursive Unit Logic Reasoning)	Comprehensive defect management, Enhance defect detection, Scalability	Complexity of Implementation, Integration Challenges, Security and Privacy Concerns	The experiment conducted shows that the GRULR method performs well in large-scale knowledge graphs and efficiently completes the reasoning task of the defect knowledge graph.
Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges	Federated learning in a collaborative cloud-edge architecture	Data Privacy, Edge Computing Efficiency, Scalability	Heterogeneity and Non-IID Data, Security Concerns, Communication and Synchronization Overhead	In this paper, the author elaborate on federated learning and cloud-edge collaborative architecture respectively. Then summarize the

[22]				key technologies, applications, and challenges of deploying federated learning in cloud-edge collaborative architecture.
Graph neural networks for representing multivariate resource usage: A multiplayer mobile gaming case-study [23]	Graph Neural Networks (GNNs), GNN-based Encoder-Decoder model	Creative strategy, Practical significance, Forecasting potential	Limited Scope, Data availability and quality, Interpretability	the proposed GCN-LSTM model outperforms other models in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), indicating its superior predictive accuracy
Intelligent intrusion detection framework for multi-clouds – IoT environment using swarm-based deep learning classifier [24]	The framework utilizes a swarm-based deep learning classifier to detect network and application-based attacks	Adaptability, Deep Learning, Swarm Intelligence	Complexity, Training Data, Scalability	The proposed framework achieved an accuracy of 95.20%, a false positive rate of 2.5%, and a detection rate of 97.24%
Intelligent Ironmaking Optimization Service on a Cloud Computing Platform by	The authors propose a multi-objective optimization framework based on cloud services and a	Improved Productivity, Convenient Remote Access, Smart Recommendations	Data Security Concerns, Dependency on Internet Connectivity, Implementation Complexity	The outcome of the research paper indicates that the application of this optimization service in a cloud factory led to

Digital Twin [25]	cloud distribution system.			significant improvements in iron production, coke ratio, and silicon content. Specifically, the iron production increased by 83.91 tons per day, the coke ratio decreased by 13.50 kg per ton, and the silicon content decreased by an average of 0.047%.
Internet of things intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine [26]	Internet of Things (IoT) intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine (MFE-ELM)	Improved Detection Accuracy, Scalability, Real-time Detection	Complexity, Resource Intensive, Dependency on Internet Connectivity	The experimental results showed that the proposed algorithm was effective in detecting and identifying most network data packets, achieving efficient intrusion detection for heterogeneous data of the IoT from cloud nodes. The algorithm enabled the cloud server to detect nodes with serious security threats in real-time, allowing for the implementation of optimal intrusion response strategies for the cloud cluster.

Method for detection of unsafe actions in power field based on edge computing architecture [27]	End-to-end action recognition model, Temporal Convolutional Neural Network (TCN), Gate Recurrent Unit (GRU)	Real-time detection, Lowered response time, Improved protection	Limited computational resources, Maintenance Challenges, Data synchronization issues	The results demonstrate that the method exhibits better real-time performance and enhances the recognition accuracy of action segments. The proposed action cognition is verified using the MSRAction Dataset, further validating the effectiveness of the method in detecting unsafe actions.
Modeling of smart risk assessment approach for cloud computing environment using AI & supervised machine learning algorithms [28]	AI (Artificial Intelligence), AI-ML techniques such as Decision Tree (DTC), Randomizable Filter Classifier, and k-star with RMSE method to analyze threats within the cloud computing environment	Improved Security, Forecasting Ability, Streamlined Processes	Data Dependency, Overfitting, Complexity	The experimental outcomes indicated that dividing the dataset into a 95%-5% partitioning provided the best results. Furthermore, the DTC algorithm yielded the best outcomes among all the algorithms used in the experimental setups.
RETRACTED ARTICLE: Improving cloud	Long-Short Term Memory (LSTM) machine	Novel Approach, Potential Efficiency Improvements, Machine Learning Integration	Retraction, Reliability Concerns, Lack of Reproducibility	The proposed technique improves network efficiency by

efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm [29]	learning algorithm			reducing the time required for resource allocation and achieving a good predictive approach.
Visibility estimation via deep label distribution learning in cloud environment [30]	Recurrent Neural Network(RNN) , Convolutional Neural Network(CNN) , label distribution learning (LDL)	Accurate Visibility Estimation, Adaptability to Cloud Dynamics, Optimized Resource Allocation	Complexity, Data dependency, Training overhead	By training the deep neural network using deep label distribution, the model can estimate the visibility of images accurately. The learned model can generate the distribution of all possible visibilities for a given test image, and the visibility with the highest probability is taken as the predicted visibility.
Object Detection Systems in Autonomous Driving: Cloud Computing and Parallelization	autonomous vehicle,autonomous driving system, computer vision, neural networks, feature	Improved Performance,Scalability, Enhanced Safety.	Dependency on Network Connectivity,Security Concerns,Cost Considerations	Exploration of key and novel concepts like distributed systems and parallelization.

Techniques[31]	extraction, segmentation, assisted driving, cloud computing, parallelization			
A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques[32]	Functional Link Neural Network (FLNN), Hybrid Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) and Multi-resource utilization prediction	Enhanced Prediction Accuracy,Adaptive Optimization,Proactive Resource Management	Complexity,Data Dependency,Computational Overhead	The proposed model achieves better accuracy compared to traditional techniques.
COME-UP: Computation Offloading in Mobile Edge Computing with LSTM Based User Direction Prediction[33]	Long-short term memory (LSTM) based user direction prediction,Fitness function for server selection,weighted sum method	Improved Offloading Efficiency,Reduced Latency,Enhanced Resource Utilization	Training Data Requirements,Model Complexity,Scalability Challenges	32% reduction in latency,16% energy savings,9% increase in resource (CPU) utilization,Root mean square error of the LSTM model: 0.5
A System for Sustainable Usage of Computing Resources Leveraging Deep Learning Predictions[34]	Time-series analysis,Deep learning models,Learning rate scheduler,Huber loss function	Resource Optimization,Energy Efficiency,Cost Savings	Model Complexity,Data Dependency,,Scalability	LSTM model outperforms the other models in terms of prediction accuracy

A Cloud Intrusion Detection System Based on DNN Using Backpropagation and PSO on the CSE-CIC-IDS2018 Dataset [35]	Multi-Layer Perceptron (MLP) with Backpropagation (BP) ,Multi-Layer Perceptron (MLP) with Particle Swarm Optimization (PSO),Binary and Multi-Class Classification	Enhanced Detection Accuracy,Adaptability,Scalability	Computational Complexity,Data Dependence,Tuning Sensitivity	accuracy obtained for binary classification was 98.97%. ,accuracy obtained for multi-class classification was 98.41%
"IoT and Cloud Computing in Health-Care: A New Wearable Device and Cloud-Based Deep Learning Algorithm for Monitoring of Diabetes[36]	Deep Learning (DL),Cloud Computing,IoT	Continuous Monitoring,Remote Access	Data Privacy and Security, Dependence on Connectivity,Financial Factors	The proposed Cloud&DL- based wearable approach achieves an average accuracy value of 15.589 in terms of RMSE
Big Data Handling Approach for Unauthorized Cloud Computing Access[37]	Advanced Encryption Standard (AES),Attribute-Based Access Control (ABAC),Hybrid Intrusion Detection (HID) ,Random Forest and Neural Network	Detecting threats early on, Scalable solution,Predicting potential threats	Incorrect Alerts, Privacy Issues	The proposed HUDH scheme achieved high accuracy in user confidentiality, access privilege, and user secret key accountability (more than 97%)

IoT-Inspired Reliable Irregularity-Detection Framework for Education 4.0 and Industry 4.0[38]	Symbolic Aggregation Approximation (SAX),Kalman Filter,Learning Bayesian Network (LBN),Multi-Layered Bi-Directional Long Short-Term Memory (M-Bi-LSTM)	Live Monitoring,Combined Data Analysis,Anticipatory Analysis	Data Privacy and Security,Implementation Complexity, False Positives	The proposed framework provides a reliable assessment, irregularity detection, and alert generation system for Education 4.0
IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning[39]	Internet of Things (IoT),Cloud Computing,Fuzzy Inference System (FIS),Bidirectional Long Short-Term Memory (Bi-LSTM)	Ability to Monitor Patients Remotely,.Predictive Analysis	Data Privacy and Security, Model Interpretability,Integration Challenge	,The proposed system achieves an accuracy of 98.86%, a precision of 98.9%, a sensitivity of 98.8%, a specificity of 98.89%, and an F-measure of 98.86%.
Technical Study of Deep Learning in Cloud Computing for Accurate Workload Prediction[40]	Recurrent Neural Networks (RNN) ,Multilayer Perception (MLP) ,Long Short-Term Memory (LSTM) ,Convolutional Neural Networks (CNN),Mean Absolute Error (MAE),Root	Enhanced Prediction Accuracy,Scalability,flexibility	Need for abundant data,Complexity,.Lack of Interpretability,Risk of Overfitting	LSTM model exhibits the best performance compared to other models

	Mean Squared Error (RMSE)			
Container Allocation in Cloud Environments using Multi-Agent Deep Reinforcement Learning[41]	Actor-Critic framework, Long Short-Term Memory (LSTM) and Memory Augmented-based agents,Bin-packing heuristics, Kubernetes	Adaptive Resource Allocation,Scalability,Coordination	Learning curve,.Communication challenges,Adaptability	The paper demonstrates the effectiveness of using LSTM and DNC-based agents for solving the container allocation problem.
Cloud Computing and IoT Based Intelligent Monitoring System for Photovoltaic Plants Using Machine Learning Techniques [42]	Internet of Things (IoT),Cloud Computing,Machine Learning, Open-source and Lightweight Software,Cost-efficient Hardware	Live Monitoring,Anticipatory Repairs,Detection of Errors and Diagnosis	Data Privacy and Security,Integration Complexity, Model Interpretability	The proposed Intelligent Monitoring System (IMS) is able to predict the output power of the PV system with high accuracy
HealthFetch: An Influence-Based, Context-Aware PrefetchScheme in Citizen-Centered Health Storage Clouds[43]	Prefetching Scheme: ,Data Replication,Machine Learning	Improved User Experience,Context Sensitivity,Influence-Centric Prefetching	Privacy Concerns, Computational Overhead,Complexity of Implementation	The paper claims that the proposed prefetching scheme significantly improves download speeds compared to the storage cloud, especially when large data are exchanged.

Toward Efficient Intrusion Detection System Using Hybrid Deep Learning Approach [44]	Recurrent Neural Network (RNN),Long Short-Term Memory (LSTM),Gated Recurrent Unit (GRU),Pearson correlation feature selection,Multi packet Detection Mechanism (MPDM)	Improved Detection Accuracy,Lower False Alarms,Scalability	Complexity of Implementation: resources,Training Data Requirements, Interpretability	Improved accuracy in intrusion detection,Reduced false alarm rate,Enhanced computational efficiency,Effective identification of multiclass threats in cloud computing
MiniDeep: A Standalone AI-Edge Platform with a DeepLearning-Based MINI-PC and AI-QSR System[45]	MiniDeep, edge computing, deep learning, cloud computing, recommendation system	Small Size,Shop space Friendly,Customer-made solutions	Hardware Constraints,Implementation Complexity,Initial Costs	Improved recommendation performance,Successful AI-QSR implementation,Effective life cycle management

Table 2 presents a summary of the techniques employed, notable features investigated, challenges and a briefly explained outcome addressed in each paper. It serves as a comprehensive reference for understanding the breadth and depth of research covered in our study.

Summary of Review

The application of recurrent neural networks (RNNs) in the context of cloud computing is examined in a review of 45 research papers on Recurrent Neural Network Inference in Cloud Computing. This paper explores the use of RNNs for a range of applications in cloud environments, including natural language processing, sequence generation, and time series prediction. The authors offer a thorough examination of the advantages and difficulties of implementing RNNs in cloud environments, taking into account issues with scalability, resource usage, and model inference latency. The research also assesses current methods for improving RNN inference in cloud systems, emphasizing strategies to improve efficiency and performance.

The authors of the review provide a critical evaluation of the state-of-the-art approaches already in use and suggest topics for further research and development. To properly utilize RNNs in cloud computing applications, they stress the significance of resolving scalability challenges, optimizing resource allocation tactics, and limiting latency issues. Furthermore, the paper delves into new developments like federated learning and edge computing integration, which present viable paths to improve RNN inference performance in distributed environments.

Future Scope

Upcoming studies may concentrate on recurrent neural network (RNN) inference procedures that are tailored for cloud computing environments. Creating methods specifically designed to take advantage of the dispersed nature of cloud infrastructure could be one way to optimize this process and increase overall inference speed and resource consumption.

Researching the integration of RNN inference with edge devices could be a successful direction as edge computing continues to gain popularity. With this connection, real-time inference might be made possible for applications that need low-latency responses, including autonomous systems and Internet of Things devices, while also utilizing cloud resources' scalability and computing resources when needed.

With the dynamic nature of cloud workloads, it is imperative to modify RNN inference techniques to accommodate changes in data processing requirements and shifting resource demands. Future studies should investigate dynamic resource allocation methods and adaptive tactics to guarantee best practices and economy in cloud-based RNN inference systems.

It's critical to address security and privacy issues related to cloud-based RNN inference. Subsequent research endeavors may explore resilient encryption strategies, secure data transmission protocols, and privacy-enhancing strategies to protect confidential information and ensure compliance to regulatory requirements.

Although RNNs have proven successful in a number of applications, it may be useful to investigate alternative neural network architectures designed for cloud-based inference. Examining transformer-based models, convolutional neural networks (CNNs), or hybrid architectures in cloud environments may provide information about how to increase the efficiency and accuracy of inference.

Conclusion

To sum up, this paper has investigated the application of recurrent neural networks (RNNs) for cloud computing environments' inference tasks. We have shown the potential of RNNs in addressing complex sequential data processing tasks, like natural language processing, time series prediction, and speech recognition, within the context of cloud infrastructure through a thorough review of the existing literature and comparative analysis. Our findings demonstrate how cloud-based RNN inference systems are scalable, flexible, and computationally efficient, allowing for the deployment of advanced machine-learning models to meet the demands of large-scale data processing. In addition, we have covered a number of deployment approaches, optimization methods, and performance factors that are critical to optimizing RNN inference's efficiency in cloud contexts.

References

1. Y. Gao, S. Zhang, and J. Zhou, ‘‘A hybrid algorithm for multi-objective scientific workflow scheduling in IaaS cloud,’’ IEEE Access, vol. 7, pp. 125783–125795, 2019, doi: 10.1109/ACCESS.2019.2939294.
2. A. Ayad, M. Renner, and A. Schmeink, “Improving the communica-tion and computation efficiency of split learning for IoT applications,”in Proc. IEEE Glob. Commun.Conf. (GLOBECOM), 2021, pp. 1–6.

3. S. P. Sone, J. J. Lehtomaki, and Z. Khan, “Wireless Traffic Usage Forecasting Using Real Enterprise Network Data: Analysis and Methods,” IEEE Open Journal of the Communications Society, vol. 1, pp. 777–797, 2020.
4. M.-K. Tran, S. Panchal, T. D. Khang, K. Panchal, R. Fraser, and M. Fowler, “Concept review of a cloud-based smart battery management system for lithium-ion batteries: Feasibility, logistics, and functionality,” Batteries, vol. 8, no. 2, p. 19, 2022.
5. D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau,“Federated learning for keyword spotting,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6341–6345.
6. T. V. Phan, N. K. Bao and M. Park, "A Novel Hybrid Flow-Based andler with DDoS Attacks in Software-Defined Networking," 2016 Intl IEEE , Toulouse, 2016, pp. 350-357.
7. H. Sadr, M. M. Pedram, and M. Teshnehlab, "Multi-View Deep Network: A Deep Model Based on Learning Features From Heterogeneous Neural Networks for Sentiment Analysis," IEEE Access, vol. 8, pp. 86984-86997, 2020.
8. M. I. Ciolacu, L. Binder, P. Svasta, I. Tache, and D. Stoichescu, ‘‘Education 4.0—Jump to innovation with IoT in higher education,’’ in Proc. IEEE 25th Int. Symp. Design Technol. Electron. Packag. (SIITME), Oct. 2019, pp. 135–141, doi: 10.1109/SIITME47687.2019.8990825.
9. A. Y. Nikravesh, S. A. Ajila, and C.-H. Lung, ‘‘Measuring prediction sen-sitivity of a cloud auto-scaling system,’’ in Proc. IEEE 38th Int. Comput.Softw. Appl. Conf. Workshops, Jul. 2014, pp. 690–695.
10. M. R. Watson, N.-U.-H. Shirazi, A. K. Marnerides, A. Mauthe, and D. Hutchison, ‘‘Malware detection in cloud computing infrastructures,’’ IEEE Trans. Dependable Secure Comput., vol. 13, no. 2, pp. 192–205, Mar. 2016.
11. J. Rubio-Loyola, A. Galis, A. Astorga, J. Serrat, L. Lefevre, A. Fischer, A. Paler, and H. Meer, ‘‘Scalable service deployment on software-defined networks,’’ IEEE Commun. Mag., vol. 49, no. 12, pp. 84–93, Dec. 2011.
12. O. Alkadi, N. Moustafa, and B. Turnbull, ‘‘A review of intrusion detec-tion and blockchain applications in the cloud: Approaches, challenges and solutions,’’ IEEE Access, vol. 8, pp. 104893–104917, 2020, doi:10.1109/ACCESS.2020.2999715.
13. R. Jhawar, V. Piuri, and M. Santambrogio, “A comprehensive conceptual system-level approach to fault tolerance in cloud computing,” in 2012 IEEE International Systems Conference SysCon 2012. IEEE, 2012, pp. 1-5.
14. J. Liu, H. Guo, H. Nishiyama, H. Ujikawa, K. Suzuki, and N. Kato, ‘‘New perspectives on future smart FiWi networks: Scalability, reliability, and energy efficiency,’’ IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1045–1072, 2nd Quart., 2016, doi: 10.1109/COMST.2015.2500960.
15. Ambati, L.S. and Gayar, O.E. (2021) ‘Human activity recognition: a comparison of machine learning approaches’, Journal of the Midwest Association for Information Systems, Vol. 2021, No. 1, pp.49–60.
16. Priyadarshini, R. and Barik, R.K., 2022. A deep learning based intelligent framework to mitigate DDoS attack in fog environment. *Journal of King Saud University-Computer and Information Sciences*, *34*(3), pp.825-831.
17. Anuradha, C. and Ponnavaikko, M., 2022. A RNN based offloading scheme to reduce latency and preserve energy using RNNBOS. *Measurement: Sensors*, *24*, p.100429.

18. Selvarajan, S., Srivastava, G., Khadidos, A.O., Khadidos, A.O., Baza, M., Alshehri, A. and Lin, J.C.W., 2023. An artificial intelligence lightweight blockchain security model for security and privacy in IIoT systems. *Journal of Cloud Computing*, *12*(1), p.38.
19. Alshammari, A. and Aldribi, A., 2021. Apply machine learning techniques to detect malicious network traffic in cloud computing. *Journal of Big Data*, *8*(1), p.90.
20. Wang, J., Zhang, G., Wang, W., Zhang, K. and Sheng, Y., 2021. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. *Journal of Cloud Computing*, *10*(1), p.4.
21. Yang, W., Li, X., Wang, P., Hou, J., Li, Q. and Zhang, N., 2022. Defect knowledge graph construction and application in multi-cloud IoT. *Journal of Cloud Computing*, *11*(1), p.59.
22. Bao, G. and Guo, P., 2022. Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges. *Journal of Cloud Computing*, *11*(1), p.94.
23. Theodoropoulos, T., Makris, A., Kontopoulos, I., Violos, J., Tarkowski, P., Ledwoń, Z., Dazzi, P. and Tserpes, K., 2023. Graph neural networks for representing multivariate resource usage: A multiplayer mobile gaming case-study. *International Journal of Information Management Data Insights*, *3*(1), p.100158.
24. Nizamudeen, S.M.T., 2023. Intelligent intrusion detection framework for multi-clouds–IoT environment using swarm-based deep learning classifier. *Journal of Cloud Computing*, *12*(1), p.134.
25. Zhou, H., Yang, C. and Sun, Y., 2021. Intelligent ironmaking optimization service on a cloud computing platform by digital twin. *Engineering*, *7*(9), pp.1274-1281.
26. Lin, H., Xue, Q., Feng, J. and 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), pp.111-124.
27. Yin, Y., Lin, J., Sun, N., Zhu, Q., Zhang, S., Zhang, Y. and Liu, M., 2021. Method for detection of unsafe actions in power field based on edge computing architecture. *Journal of Cloud Computing*, *10*, pp.1-14.
28. Sharma, A. and Singh, U.K., 2022. Modelling of smart risk assessment approach for cloud computing environment using AI & supervised machine learning algorithms. *Global Transitions Proceedings*, *3*(1), pp.243-250.
29. Ashawa, M., Douglas, O., Osamor, J. and Jackie, R., 2022. RETRACTED ARTICLE: Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm. *Journal of Cloud Computing*, *11*(1), p.87.
30. Song, M., Han, X., Liu, X.F. and Li, Q., 2021. Visibility estimation via deep label distribution learning in cloud environment. *Journal of Cloud Computing*, *10*(1), p.46.
31. Cortes Gallardo Medina, E., Velazquez Espitia, V.M., Chipuli Silva, D., Fernandez Ruiz de las Cuevas, S., Palacios Hirata, M., Zhu Chen, A., Gonzalez Gonzalez, J.A., Bustamante-Bello, R. and Moreno-García, C.F., 2021. Object detection, distributed cloud computing and parallelization techniques for autonomous driving systems. *Applied sciences*, *11*(7), p.2925.
32. Malik, S., Tahir, M., Sardaraz, M. and Alourani, A., 2022. A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques. *Applied Sciences*, *12*(4), p.2160.

33. Zaman, S.K.U., Jehangiri, A.I., Maqsood, T., Umar, A.I., Khan, M.A., Jhanjhi, N.Z., Shorfuzzaman, M. and Masud, M., 2022. COME-UP: Computation offloading in mobile edge computing with LSTM based user direction prediction. *Applied Sciences*, *12*(7), p.3312.
34. Cioca, M. and Schuszter, I.C., 2022. A system for sustainable usage of computing resources leveraging deep learning predictions. *Applied Sciences*, *12*(17), p.8411.
35. Alzughaibi, S. and El Khediri, S., 2023. A cloud intrusion detection systems based on dnn using backpropagation and pso on the cse-cic-ids2018 dataset. *Applied Sciences*, *13*(4), p.2276.
36. Nasser, A.R., Hasan, A.M., Humaidi, A.J., Alkhayyat, A., Alzubaidi, L., Fadhel, M.A., Santamaría, J. and Duan, Y., 2021. Iot and cloud computing in health-care: A new wearable device and cloud-based deep learning algorithm for monitoring of diabetes. *Electronics*, *10*(21), p.2719.
37. Razaque, A., Shaldanbayeva, N., Alotaibi, B., Alotaibi, M., Murat, A. and Alotaibi, A., 2022. Big data handling approach for unauthorized cloud computing access. *Electronics*, *11*(1), p.137.
38. Verma, A., Anand, D., Singh, A., Vij, R., Alharbi, A., Alshammari, M. and Ortega Mansilla, A., 2022. IoT-Inspired Reliable Irregularity-Detection Framework for Education 4.0 and Industry 4.0. *Electronics*, *11*(9), p.1436.
39. Nancy, A.A., Ravindran, D., Raj Vincent, P.D., Srinivasan, K. and Gutierrez Reina, D., 2022. Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. *Electronics*, *11*(15), p.2292.
40. Ahamed, Z., Khemakhem, M., Eassa, F., Alsolami, F. and Al-Ghamdi, A.S.A.M., 2023. Technical study of deep learning in cloud computing for accurate workload prediction. *Electronics*, *12*(3), p.650.
41. Danino, T., Ben-Shimol, Y. and Greenberg, S., 2023. Container allocation in cloud environment using multi-agent deep reinforcement learning. *Electronics*, *12*(12), p.2614.
42. Emamian, M., Eskandari, A., Aghaei, M., Nedaei, A., Sizkouhi, A.M. and Milimonfared, J., 2022. Cloud computing and IoT based intelligent monitoring system for photovoltaic plants using machine learning techniques. *Energies*, *15*(9), p.3014.
43. Symvoulidis, C., Marinos, G., Kiourtis, A., Mavrogiorgou, A. and Kyriazis, D., 2022. Healthfetch: An influence-based, context-aware prefetch scheme in citizen-centered health storage clouds. *Future Internet*, *14*(4), p.112.
44. Aldallal, A., 2022. Toward efficient intrusion detection system using hybrid deep learning approach. *Symmetry*, *14*(9), p.1916.
45. Chen, Y.S., Cheng, K.H., Hsu, C.S. and Zhang, H.L., 2022. MiniDeep: A Standalone AI-Edge Platform with a Deep Learning-Based MINI-PC and AI-QSR System. *Sensors*, *22*(16), p.5975.