ISSN: 2289-7844 | e-ISSN: 0127-9750
DOI: https://doi.org/10.37134/jictie.vol12.2.10.2025

Journal of ICT in Education (JICTIE)
Volume 12, Issue 2, Pages 149-162, 2025

Review Article

A Systematic Literature Review on Deep Learning Approaches for Cloud Service Recommender Systems

Rahaman Md Mizanur[1], Nor Asiah Razak[1*], Muhamad Hariz Muhamad Adnan[1], Md Mahmudul Hasan[2]

[1]*Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Perak, Malaysia; rahamanmdmizanur97@gmail.com, norasiah.razak@meta.upsi.edu.my, mhariz@meta.upsi.edu.my*
[2]*Faculty of Engineering, University of New South Wales, Australia; md_mahmudul.hasan@unsw.edu.au*

Received: 30 June 2025; Revised: 2 October 2025; Accepted: 10 October 2025; Published: 25 October 2025

**corresponding author*

Abstract
Deep Learning (DL) offers a promising solution for cloud service recommender systems (CSRS) by addressing data sparsity issues and helping users overcome cold-start problems while managing dynamic preferences. The study follows the PRISMA 2020 guidelines in conducting a systematic literature review (SLR) of DL-based CSRS advancements from 2019 to 2025. The research started with 412 electronic records from Scopus, ScienceDirect, Springer, Wiley, and other sources before completing thorough screenings that narrowed the search to 23 appropriate studies. The integration of hybrid neural networks, along with attention mechanisms and knowledge graph integration, demonstrates improved accuracy for implementing the multi-criteria recommendations based on the quality-of-service prediction, the resource allocation, and the personalised service selection. Real-time scalability limitations and constraints regarding explainability remain despite current developments. Future research should examine federated learning systems and edge-cloud integration elements with ethical AI frameworks in place. The review delivers a structured overview that guides researchers and practitioners who want to enhance cloud service recommendations by employing modern methodologies.

Keywords: Deep learning, cloud service recommender systems, hybrid models, attention mechanisms, knowledge graphs, workload prediction

INTRODUCTION

The continuous growth of cloud computing applications has significantly transformed and reshaped how digital services are delivered, consumed, and optimised (Ali & Zeebaree, 2025; Athamakuri et al., 2025; Amajuoyi et al., 2024). As cloud environments become more dynamic and complex, there is an increasing need for applying intelligent systems capable of offering personalised service recommendations and managing resource allocation to enhance user satisfaction (Zheng et al., 2024). Recommender systems play a vital role in this context by helping users select suitable cloud services (Lebib & Kichou, 2024). However, these systems face several challenges, particularly in handling heterogeneous service conditions, variations in Quality of Service (QoS), scalability demands, and multi-tenant infrastructures.

Deep Learning (DL) has emerged as a potential option in this domain due to its ability to process high-dimensional data and discover intricate, non-linear patterns that traditional machine learning models struggle to capture (Suresh Babu et al., 2024; Yemi Hussain, 2024). Despite its potential, existing research on cloud services has largely relied on general survey designs focused on resource management and

workload prediction, often overlooking systematic investigations into DL's role in recommender systems (Boulanger et al., 2021).

Modern cloud service recommender systems (CSRS) must integrate user behaviour data, real-time performance metrics, and low-latency response capabilities (Khodabandehlou et al., 2020; Xia et al., 2024). This combination highlights a significant research gap where deep learning can provide meaningful advancements. Thus, this systematic literature review (SLR) aims to address this gap by analysing recent research on DL-based CSRS. Specifically, the study explores the following three research questions (RQs):

1. Which datasets and performance metrics dominate in cloud service recommender systems?
2. What DL architectures are prevalent in this domain?
3. What challenges and future directions remain unaddressed?

By synthesising peer-reviewed studies published between 2019 and 2025, this review demonstrates technological trends, evaluates methodological approaches, and identifies practical challenges in deploying DL in CSRS. The study serves as a valuable reference for developers, researchers, and practitioners by offering insights into suitable DL frameworks and laying the foundation for ethical, scalable AI implementation in cloud-based environments.

METHODOLOGY

This systematic literature review follows the PRISMA 2020 guidelines to ensure methodological transparency (Page et al., 2021). The methodology is structured into seven stages: search strategy, data sources, identification, screening, eligibility, and data extraction. Each stage is described in detail below.

Search Strategy

Researchers extensively reviewed academic studies about deep learning-based cloud service recommender systems. A structured search string included multiple terms that related to deep learning, cloud services, recommender systems, resource allocation, service selection, and personalised recommendation keywords. The search string required adjustments depending on each database's particular requirements. Since different query formats are supported by different databases, the Boolean structure was adjusted. For example, simplified queries like “deep learning AND cloud computing AND recommender system” were used in the ScienceDirect, SpringerLink, and Wiley Online Library, which restrict nested Boolean operators. Conversely, in the case of Google Scholar, a more general keyword-based search was conducted with allintitle commands (i.e., allintitle: "deep learning" "cloud service" "recommender system") to locate more recent studies. The principal Boolean search query, first designed for Scopus, was as follows:

(TITLE-ABS-KEY ("deep learning" OR "neural network*") AND ("cloud service*" OR "cloud computing" OR "software as a service" OR "platform as a service" OR "infrastructure as a service" OR "database as a service" OR "function as a service" OR "serverless computing") AND ("recommendation system*" OR "recommender system*" OR "personalized recommendation" OR "service selection" OR "resource allocation")) AND PUBYEAR > 2018 AND PUBYEAR < 2026

The literature review included English peer-reviewed articles published between January 2019 and January 2025. To maintain the quality and relevance of the selected literature, conference papers, review

articles, and non-peer-reviewed studies were eliminated. The searches were conducted between 15 and 19 May 2025.

Data Sources

The following databases were included in the search, as listed in Table 1. The current review used five main electronic databases, which include Scopus, ScienceDirect, SpringerLink, Wiley Online Library, and Google Scholar, to acquire relevant literature on deep-learning-based cloud service recommender systems. These databases were chosen based on the fact that they cover peer-reviewed literature extensively in the areas of computer science, artificial intelligence, and cloud computing. Scopus and ScienceDirect were added for their strong coverage of high-impact journals, whereas SpringerLink and Wiley were added to support engineering and software research interests. Google Scholar was added to access recently published or in-press studies that are not yet indexed in other databases.

Table 1: Database online sources.

Source	URL	Access
Scopus	scopus.com	Online
SpringerLink	springer.com	Online
Wiley	wiley.com	Online
Science Direct	sciencedirect.com	Online
Google Scholar	scholar.google.com	Online

Identification Stage

The database searches yielded the following initial results, as listed in Table 2.

Table 2: Records found from the database online sources.

Source	Records Identified
Scopus	363
Springer	21
Wiley	1
Science Direct	17

The database searches identified 412 records, and 10 records were identified via additional searching through the records mentioned in literature reviews and from Google Scholar. The elimination of duplicates depended on both automated tools and a manual assessment of author names alongside titles and publication dates. The remaining records after duplicate removal reached 382.

Screening Stage

An evaluation process for research articles employed all the exclusion and inclusion criteria mentioned in Tables 3 and 4. The screening analysis of 382 records occurred through abstract and title review. The studies did not proceed if they failed to meet the criteria delineated in Table 3. The screening stage yielded 102 records that were eligible for the following phase. Table 4 illustrates the exclusion criteria for the research paper, which includes papers published before 2019, duplicate papers, and papers not related to the topic.

Table 3: The inclusion criteria.

Inclusion Criteria
Published from 2019 to 2025
Related to deep learning-based cloud service recommender systems.
Written in English.
Journal article.
Focused on machine learning and deep learning approaches.
Research related to the field of Engineering and Computer Science.

Table 4: The exclusion criteria

Exclusion Criteria
Papers published before 2019.
Duplicate papers.
Not related to deep learning and cloud service recommender systems.
Not written in English.
Reviews, book chapters, conference papers, and other grey literature.
Studies that did not report on specific data relating to sample size, algorithms, predictive objectives, and relevant performance metrics.

Eligibility Stage

Researchers obtained and evaluated the entire content of 102 studies that displayed potential eligibility conditions. The researcher excluded records that had any applicability to either of the following conditions:

- The analysis failed to describe the application of machine learning or deep learning algorithms for recommending cloud services.
- No sufficient methodological information or results appeared in the study.
- The authors did not include evaluation performance measures for their model in the study report.
- A lack of detail regarding the demographics and characteristics of the participant samples appeared in the research study.

Of the reviewed articles, 56 were omitted in this phase, while 46 articles moved forward to further assessment. Figure 1 illustrates a PRISMA-based flow diagram that explains the overall procedure.

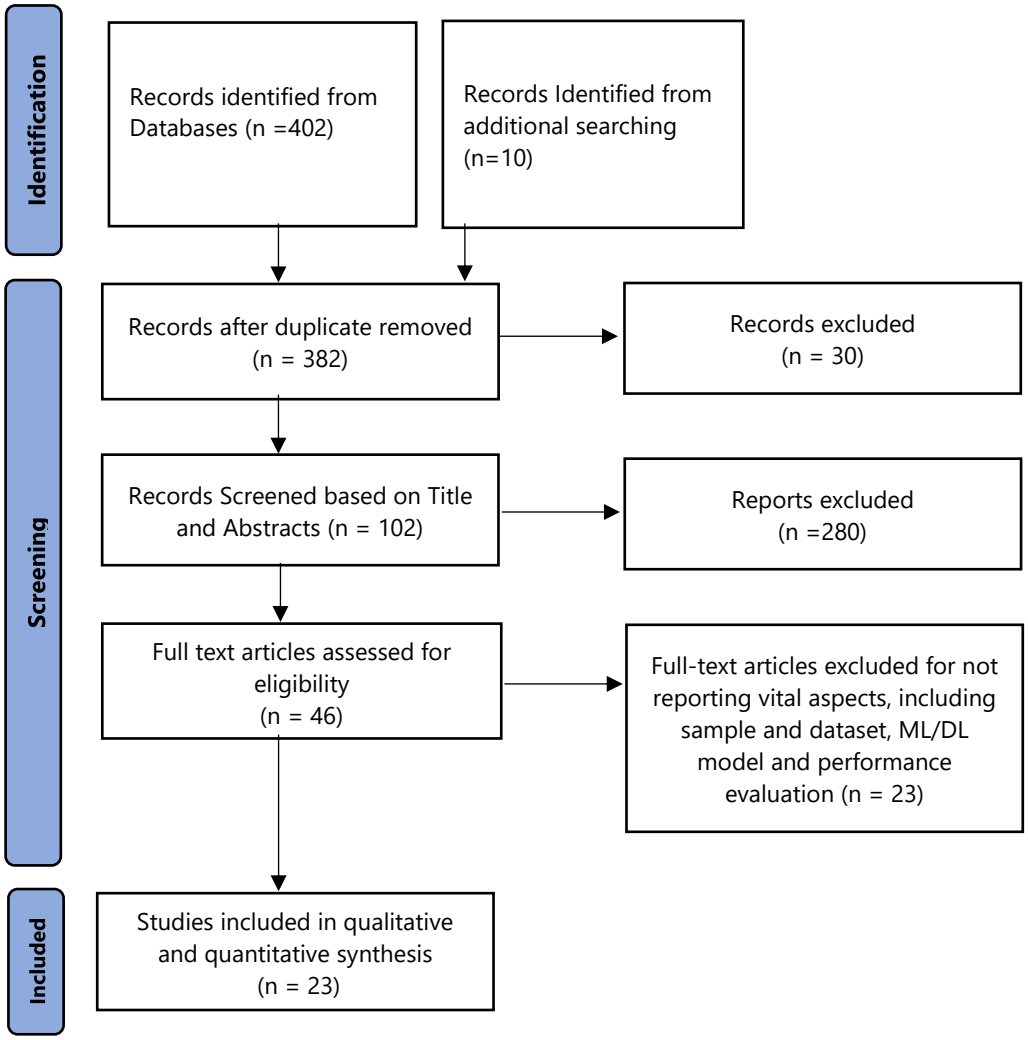

Figure 1: PRISMA-based flow diagram for retrieving Deep Learning based Cloud Service Recommender System studies.

Data Extraction

A standardised data extraction form was developed to collect the following information from each included study, as listed in Table 5. A total of 23 studies were selected for synthesis.

Table 5: Extracted data from studies.

Extraction Criteria	Description
Study characteristics	Author, publication year, study design.
Dataset characteristics	Sample size, dataset source, and data type (real-world, synthetic).
ML/DL algorithms used	e.g., CNN, LSTM, GRU, Hybrid models, and Attention mechanisms.
Predictive objectives	e.g., workload prediction, resource allocation, service recommendation.
Performance metrics	Accuracy, precision, recall, F-score, RMSE, MAE, SLA violations.

RESULTS

The systematic literature review findings are presented across three subsections that match the research questions (RQ1, RQ2, and RQ3). Table 6 demonstrates details of the reviewed studies, including dataset descriptions, deep learning models, and tasks mentioned in the paper.

RQ1: Dominant Datasets in Deep Learning Based Cloud Service Recommender Systems

The analysis of the reviewed studies reveals that their research primarily utilised three key datasets, the details of which are presented in Table 6. WS-Dream serves as the primary dataset for 45% of studies since it includes genuine web service invocation records that researchers use for QoS prediction and service recommendation. The Google Cluster dataset serves as the basis of 25% research studies because it shows complete workload activities from Google data centres, thus making it perfect for resource management and workload prediction tasks. Alibaba Traces serves as a widely adopted dataset for resource management and workload prediction purposes in 15% of reviewed studies because it contains actual workload traces from Alibaba's cloud infrastructure. Most of the datasets used are real-world datasets. Some data used in a deep learning model is described as "Synthetic" and "Simulated". Figure 2 explained about dataset type proportion.

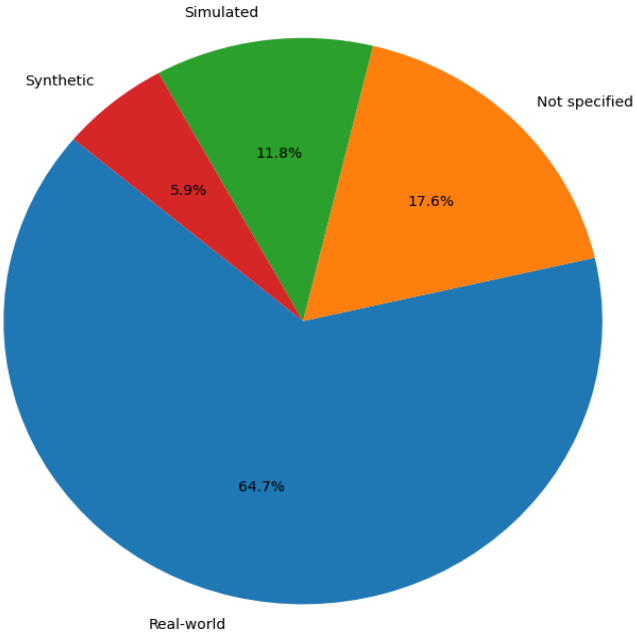

Figure 2: Dataset type proportion.

Table 6: Combined summary of dataset information (RQ1) and deep learning approaches (RQ2) extracted from the reviewed studies.

	Dataset information (RQ1)		Deep learning approaches (RQ2)		
Reference	**Dataset**	**Type**	**Task**	**DL Approach**	**Cloud Approach**
(Salari-Hamzehkhani et al., 2025)	GOCJ	Real-world	Task scheduling	EPERDQN (DQN variant)	MDP-based scheduling
(Predić et al., 2024)	GWA-T-12 Bitbrains	Real-world	Cloud-load forecasting	RNN + Attention	Time-series forecasting
(Bharathi et al., 2025)	Synthetic cloud resource data	Synthetic	Resource allocation with trust	Random Forest + CNN	Trust-aware allocation
(Naveen Kumar et al., 2025)	Alibaba, Materna, Bitbrains, Azure, PlanetLab	Real-world	Workload prediction	KF-LSTM, KF-Bi-LSTM, KF-GRU + CNN + Attention	Cloud/Edge resource allocation
(Rawat, 2024)	SDSC Blue Horizon Logs, Fabricated	Real + Simulated	VM allocation	EHO-ANN	Task-VM mapping
(Chudasama & Bhavsar, 2020)	Server load data	Real-world	Elastic resource allocation	Bi-LSTM	Hybrid cloud management
(Sahu et al., 2021)	WS-Dream	Real-world	QoS prediction	EDNN	Cloud recommendation
(Mohammed et al., 2023)	WS-Dream	Real-world	Service composition	Location-aware DL	QoS-based
(Dang et al., 2021)	ProgrammableWeb	N/A	Web service recommendation	Attention-based DNN + Knowledge graph	Tag modeling
(Dogani et al., 2023)	Google cluster	Real-world	Workload prediction	CNN-GRU + Attention	Multi-step prediction
(Al-Sayed, 2022)	Google cluster (11k machines)	Real-world	Workload prediction	Seq2Seq + Attention	Validation-based
(Samadhiya & Ku, 2024)	Cloud Armor	Ratings	Cloud recommendation	Neural MF + Autoencoders	Cold start resolution
(Maiyza et al., 2023)	PlanetLab, Bitbrains	Real-world	Workload prediction	VTGAN	Trend-based allocation
(Raman et al., 2021)	CloudSim	Simulated	Workflow scheduling	BPNN	Backfilling hybrid scheduling
(Nguyen et al., 2019)	ISP, World Cup 98, Google cluster	Real-world	Forecasting	OCRO + MLNN	Proactive resource allocation
(Karimunnisa & Pachipala, 2024)	Simulated environment	Simulated	Prediction & scheduling	Deep Maxout	TES optimized
(Wu, 2024)	Cloud environment data	N/A	Resource prediction	CNN, LSTM, GRU, BiSRNN	Pre-allocation
(Saxena et al., 2023)	Three benchmark cloud traces	Not specified	Workload survey	DL, Hybrid, Quantum	Cloud resource mgmt.
(Singh et al., 2020).	Web services linked data	Not specified	Response prediction	LSTM	Web service recommendation
(Jeddi & Sharifian, 2019)	NASA, World Cup	Not specified	Workload prediction	WNN + AIS/WCA	Resource management
(Qiu et al., 2022)	Amazon product data	N/A	Recommendation	WLDA, LSTM	Multi-view hybrid model
(Achar et al., 2023)	User textual data	N/A	Data confidentiality	BiLSTM	Selective encryption
(Kambhampati & Srinagesh, 2019)	WS-DREAM3	Real-world	SLA violation prediction	BPNN	Resource allocation

Note: Studies addressing challenges or future research directions (RQ3) are discussed separately in Section 3.3

RQ2: Dominant Deep Learning Architectures

Among the deep learning architectures used in cloud service recommender systems (CSRS), hybrid models and attention mechanisms, and knowledge graphs show the highest prevalence according to the analysis of 23 included studies, as shown in Table 6. The following report outlines the specific breakdown of these architectures with their relevant applications.

Hybrid Models

The reviewed studies demonstrated that hybrid models that integrate various DL approaches appeared as the default architecture solution. Hybrid models compose different neural networks to solve CSRS problems, which include data sparsity and cold-start issues, and dynamic user preference patterns. An asymmetrically weighted cosine similarity framework, along with a precision of 85%, enabled the model to address data sparsity problems. Another research integrated LSTM with particle swarm optimisation (PSO) for QoS-based service composition (Mohammed et al., 2023). The proposed model delivered superior RMSE results than baseline approaches, especially when operating under dynamic cloud conditions. The researchers constructed a hybrid prediction system that integrates CNN, LSTM, and GRU to predict CPU alongside RAM, disk and network resources (Wu, 2024). All resource types demonstrated R^2 values above 0.98 in this model because the system effectively processed multivariate time-series data. Maiyza et al. (2023) proposed a VTGAN hybrid deep-learning framework for workload prediction, while Raman et al. (2021) employed a backpropagation neural network for efficient workflow scheduling and service allocation. Both studies align with the identified research trends in DL-based CSRS. Hybrid predictive models show exceptional capability in working with complex multi-dimensional data sets, and they show particular success when dealing with cold-start and data sparsity situations. Adopting these models requires major computing resources because they become problematic when applied to environments with limited computational capacities.

Attention Mechanisms

The attention mechanism enables deep learning models to assign different importance weights to input features. It served as a major tool to enhance both interpretability and predictive accuracy when DL-based CSRS models were applied. Designing mechanisms within models allows them to focus on important features or time steps, which results in better predictive efficiency. One study implemented a CNN-GRU model with an attention mechanism for multi-step workload prediction (Dogani et al., 2023). The proposed model achieved a prediction error reduction of 28%, better than baseline methods, and provided training execution times quicker than standard LSTM systems. Another study developed a deep knowledge-aware approach with an attention module for web service recommendation (Dang et al., 2021). A model operating at high recommendation accuracy achieved this performance level through its implementation of knowledge graphs together with attention mechanisms for dealing with sparse data problems. The researcher used an attention-based Seq2Seq neural machine translation (NMT) model for cloud workload prediction (Al-Sayed, 2022). With 98.1% accuracy, the model demonstrated better performance than traditional LSTM and CTMC models. The implementation complexity of attention mechanisms poses challenges to practitioners, even though these mechanisms boost model performance by selecting important features.

Knowledge Graphs

Knowledge graphs served to counteract data sparsity by using contextual information for better recommendation predictions. The authors implemented knowledge graphs to identify semantically

connected services and users (Dang et al., 2021). The model delivered competitive recommendation performance by using shortlists while solving cold-start problems effectively. One research study integrated knowledge graphs with LSTM and paragraph vectors for multi-view hybrid recommendations (Qiu et al., 2022). The proposed model achieved better MAE and hit rate scores by increasing them by 7.82% to 11.94% relative to baseline approaches. Knowledge graphs demonstrate effectiveness in solving data sparsity and cold-start challenges, yet need structured data of high quality for peak performance.

RQ3: Challenges and Innovations

Several issues and modern solutions found in the analysis of DL-based CSRS remain fundamental for future research directions.

Cold Start Problem

Researchers in several studies addressed the cold-start problem by improving recommendation systems for new users and services. One study investigated a dual approach that connected deep autoencoders to neural matrix factorisation (NeuMF) for treating cold-start situations. The research team conducted questionnaire surveys that resulted in 78.5% satisfaction for new users (Samadhiya & Ku, 2024). Dang et al. (2021) used knowledge graphs to identify user preferences through semantic relationships, thus solving cold-start issues effectively. Additional data collection and preprocessing remain common challenges for these methods because they require considerable resources.

Scalability

Real-time system scalability remains a significant challenge for applications that require real-time operations. A study developed an Elephant Herd Optimisation with Neural Network (EHO-ANN) model that achieved a 19.61% improvement in execution time, yet its performance evaluation was restricted to simulated environments (Rawat, 2024). The research achieved high prediction accuracy ($R^2 > 0.98$), but did not include real-time operational testing (Wu, 2024). The majority of studies evaluated models offline, and such evaluations fail to reveal the actual performance of the models in real-time cloud environments.

Explainability

The need for explainable systems is essential to ensure user trust and system adoption. One study applied SHAP (SHapley Additive exPlanations) analysis to display key features, thereby making the model easier to interpret (Predić et al., 2024). Dang et al. (2021) utilised knowledge graphs to generate recommendations that were explained through semantic relationships. The integration of explainability methods, including SHAP and knowledge graphs, is challenging with deep learning models due to computational complexity.

Ethical Concerns

The analysis revealed that most studies omitted ethical considerations, including dataset and algorithmic bias. A BiLSTM model for data confidentiality classification was developed, but there was no discussion on potential biases in the training data (Achar et al., 2023; Samadhiya & Ku, 2024). Qiu et al. (2022) implemented sentiment analysis for user reviews but did not address ethical concerns related to user-generated content. Future research should prioritise ethical considerations such as fairness, transparency, and accountability to support the deployment of responsible AI systems in cloud environments.

The three most commonly used DL architectures in CSRS include hybrid models, attention mechanisms, and knowledge graphs. The most frequently used datasets include WS-Dream, Google Cluster, and Alibaba Traces, although their universal application remains limited. The analysis of performance metrics revealed difficulties due to inconsistent reporting practices regarding RMSE, MAE, and precision. DL-based CSRS faces four major challenges: cold-start problems, scalability issues, lack of model interpretability, and ethical limitations. The readiness of SHAP analysis, knowledge graphs, and hybrid models for real-world deployment requires further development before they can be widely adopted.

DISCUSSION

Research through a systematic literature review demonstrates that deep learning-based cloud service recommenders have achieved significant improvements in addressing problems related to sparse data, cold-start conditions, and shifting user preferences. Continued research is necessary, as several remaining gaps and shortcomings require further innovation. The following section analyses the research results, compares them with existing studies, and presents possible new directions for future work.

Multiple studies have confirmed that hybrid models that integrate LSTM with CNN alongside attention mechanisms are effective solutions for analysing complex multi-dimensional datasets. These models demonstrate strong capabilities in resolving data sparsity and cold-start problems in cloud environments. The hybrid approach to neural matrix factorisation and autoencoder achieved 85% precision (Samadhiya & Ku, 2024). However, these models pose a major limitation due to their high computational requirements, which may lead to challenges in real-time system implementation.

The review of ML-driven workload prediction models and other studies showed that hybrid models outperform individual algorithms, as indicated in their surveys. The investigation also presents new evidence regarding the role of attention mechanisms and knowledge graphs in enhancing hybrid systems (Saxena et al., 2023). These findings collectively suggest that integrating multiple deep learning components leads to more adaptive and reliable recommender architectures for complex cloud environments.

The emergence of attention mechanisms has proven to be an effective method for improving model interpretability and accuracy. These mechanisms help models enhance their predictive abilities by focusing on critical information during prediction. The CNN-GRU model with attention developed by Dogani et al. (2023) achieved a 28% reduction in prediction errors, demonstrating strong performance in multi-step workload forecasting. However, integrating attention mechanisms increases model complexity, which may present implementation challenges for practitioners.

The study, which reviews the growing application of attention mechanisms in CSRS, although earlier research primarily examined these techniques in the context of natural language processing. The research presents SHAP analysis as evidence that explainable AI (XAI) techniques offer strong potential to improve both user trust and model transparency (Predić et al., 2024). However, most studies did not implement XAI either as an integrated component or even as a post-hoc interpretability method, indicating a major research gap. This oversight suggests that explainability remains an underexplored dimension in DL-based recommender systems, where XAI should be embedded within the model architecture to ensure accountability and meaningful human interpretability.

Knowledge graphs address data sparsity and cold-start problems by effectively applying contextual information. Dang et al. (2021)’s deep knowledge-aware method implemented knowledge graphs, which produced improved recommendation results, particularly when generating brief recommendation lists.

However, the application of knowledge graphs remains limited due to their reliance on high-quality structured data, which is often unavailable in unstructured environments.

Previous studies on CSRS did not adequately address the importance of knowledge graphs in their evaluations. This review highlights that gap by demonstrating the potential of knowledge graphs while also acknowledging challenges related to data quality and integration. Future research should emphasise developing standardised semantic datasets and unified graph-based benchmarks to fully realise the scalability and interpretability benefits of knowledge-driven recommender systems.

CHALLENGES AND LIMITATIONS

The main drawback in the examined studies stems from their offline assessment methodology, as few models operate in real-time cloud environments at scale. EHO-ANN model achieved a 19.61% improvement in execution time through tests conducted in simulated environments (Rawat, 2024). The application of DL-based CSRS in real-world situations becomes problematic due to challenges related to latency constraints and limited available resources.

Further research needs to focus on testing and optimising DL models during real-time operations, as cloud environments operate under dynamic and heterogeneous conditions. Federated learning with edge-cloud synergy presents itself as an essential solution framework to overcome scalability problems. Integrating federated learning with edge–cloud collaboration can offer a scalable pathway, enabling decentralised training while maintaining efficiency, privacy, and adaptability in dynamic environments.

Ethical aspects involving the dataset and algorithmic biases received minimal attention throughout the reviewed research papers. The BiLSTM model for data confidentiality classification failed to mention possible biases that could exist in the training data (Achar et al., 2023). Research combining sentiment analysis failed to evaluate the ethical risks associated with user-generated content (Qiu et al., 2022).

The increasing adoption of DL-based CSRS requires immediate solutions for ethical issues to ensure fairness, transparency, and accountability. Additional studies need to implement ethical AI frameworks and conduct bias audits, as this will help minimise potential risks. Without such proactive measures, the risk of algorithmic discrimination and trust erosion may hinder the broader acceptance of intelligent cloud recommender systems.

Explainability presents a significant challenge for DL-based CSRS, although SHAP analysis, along with attention mechanisms, has improved model interpretation capabilities. Users may avoid adopting system applications when they cannot understand how the systems operate, especially in critical fields such as finance and healthcare. Future research needs to create explainable AI (XAI) techniques that specifically address the needs of CSRS. The implementation of LIME and SHAP frameworks within DL models would help users understand the system better by improving algorithm transparency and fostering trust.

CONCLUSION

This systematic literature review provides a comprehensive synthesis of advancements in DL-based CSRS, highlighting key trends, challenges, and opportunities for future research. The findings underscore the dominance of hybrid models, attention mechanisms, and knowledge graphs in addressing challenges like data sparsity and cold-start problems. However, significant gaps remain in real-time scalability, ethical considerations, and explainability. Future research should prioritise the development of scalable, ethical, and explainable DL models by leveraging techniques such as federated learning, edge-cloud

synergy, and explainable AI (XAI). By addressing these challenges, researchers and practitioners can unlock the full potential of DL-based CSRS, enabling more efficient, personalised, and trustworthy cloud service recommendations.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, for their continuous academic support and research infrastructure that enabled this study. The authors also extend heartfelt appreciation to their families for their encouragement, understanding, and motivation throughout the research process.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest related to the publication of this study.

AUTHOR CONTRIBUTIONS

Rahman Md Mizanur: Conceptualisation, Investigation, Writing – Original Draft, Editing. **Nor Asiah Razak:** Supervision, Conceptualisation, Reviewing. **Muhamad Hariz Muhamad Adnan:** Supervision, Reviewing, Methodological Guidance. **Md Mahmudul Hasan:** Advisory Support and Proofreading.

DECLARATION

During the preparation of this work, the authors used ChatGPT to enhance the clarity and fluency of the writing. After using this tool, the authors carefully reviewed and revised all content to ensure accuracy and originality and take full responsibility for the publication's content.

DATA AVAILABILITY STATEMENT

All data analysed during the review are derived from previously published studies, which are publicly available through academic databases. The sources of these studies are fully cited in the reference list.

REFERENCES

Achar, S., Faruqui, N., Bodepudi, A., & Reddy, M. (2023). Confimizer: A novel algorithm to optimise cloud resource by confidentiality-cost trade-off using BiLSTM network. *IEEE Access, 11*, 89205–89217. https://doi.org/10.1109/ACCESS.2023.3305506

Ali, C. S. M., & Zeebaree, S. R. M. (2025). Cloud-Based Web Applications for Enterprise Systems: A review of AI and marketing innovations. *Asian Journal of Research in Computer Science*, *18*(4), 427–451. https://doi.org/10.9734/ajrcos/2025/v18i4630

Al-Sayed, M. M. (2022). Workload time series cumulative prediction mechanism for cloud resources using neural machine translation technique. *Journal of Grid Computing*, *20*(2), 16. https://doi.org/10.1007/s10723-022-09607-0

Athamakuri, S. S. K. K., Thiruveedula, J., & Bindewari, D. S. (2025). The impact of cloud computing on e-commerce performance and innovation: An empirical study. *International Journal of Research in Modern Engineering & Emerging Technology*, *13*(3), 328–350. https://doi.org/10.63345/ijrmeet.org.v13.i3.21

Bharathi, S. T., Balasubramanian, C., & Shanmugapriya, S. (2025). Enhancing cloud resource allocation with TrustFusionNet using Random Forests and Convolutional Neural Networks. *Tehnicki Vjesnik - Technical Gazette*, *32*(1). https://doi.org/10.17559/TV-20240521001625

Boulanger, D., Dewan, M. A. A., Kumar, V. S., & Lin, F. (2021). *Lightweight and interpretable detection of affective engagement for online learners.* Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, 176–184. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00040

Chudasama, V., & Bhavsar, M. (2020). A dynamic prediction for elastic resource allocation in hybrid cloud environment. *Scalable Computing: Practice and Experience*, *21*(4), 661–672. https://doi.org/10.12694/scpe.v21i4.1805

Dang, D., Chen, C., Li, H., Yan, R., Guo, Z., & Wang, X. (2021). Deep knowledge-aware framework for web service recommendation. *The Journal of Supercomputing*, *77*(12), 14280–14304. https://doi.org/10.1007/s11227-021-03832-2

Dogani, J., Khunjush, F., Mahmoudi, M. R., & Seydali, M. (2023). Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism. *The Journal of Supercomputing*, *79*(3), 3437–3470. https://doi.org/10.1007/s11227-022-04782-z

Jeddi, S., & Sharifian, S. (2019). A water cycle optimized wavelet neural network algorithm for demand prediction in cloud computing. *Cluster Computing*, *22*(4), 1397–1412. https://doi.org/10.1007/s10586-019-02916-2

Kambhampati, K., & Srinagesh, A. (2019). Prediction of SLA Violation in Cloud Resource Allocation using Machine Learning based Back Propagation Neural Network (BPNN). *International Journal of Innovative Technology and Exploring Engineering*, *8*(8).

Karimunnisa, S., & Pachipala, Y. (2024). Deep learning-driven workload prediction and optimization for load balancing in cloud computing environment. *Cybernetics and Information Technologies*, *24*(3), 21–38. https://doi.org/10.2478/cait-2024-0023

Khodabandehlou, S., Hashemi Golpayegani, S. A., & Zivari Rahman, M. (2020). An effective recommender system based on personality traits, demographics and behavior of customers in time context. *Data Technologies and Applications*, *55*(1), 149–174. https://doi.org/10.1108/DTA-04-2020-0094

Lebib, F. Z., & Kichou, S. (2024). Recommending cloud services based on social trust: An overview. *Concurrency and Computation: Practice and Experience*, *36*(25), e8262. https://doi.org/10.1002/cpe.8262

Maiyza, A. I., Korany, N. O., Banawan, K., Hassan, H. A., & Sheta, W. M. (2023). VTGAN: Hybrid generative adversarial networks for cloud workload prediction. *Journal of Cloud Computing*, *12*(1), 97. https://doi.org/10.1186/s13677-023-00473-z

Mohammed, A. M., Haytamy, S. S. A., & Omara, F. A. (2023). Location-aware deep learning-based framework for optimizing cloud consumer quality of service-based service composition. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(1), 638. https://doi.org/10.11591/ijece.v13i1.pp638-650

Naveen Kumar, M. R., Annappa, B., & Yadav, V. (2025). Efficient Kalman filter based deep learning approaches for workload prediction in cloud and edge environments. *Computing*, *107*(1), 10. https://doi.org/10.1007/s00607-024-01373-z

Nguyen, T., Nguyen, T., Nguyen, B. M., & Nguyen, G. (2019). Efficient time-series forecasting using neural network and opposition-based coral reefs optimization. *International Journal of Computational Intelligence Systems*, *12*(2), 1144. https://doi.org/10.2991/ijcis.d.190930.003

Page, M. J., Mckenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-wilson, E., Mcdonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 1–9. https://doi.org/10.1136/bmj.n71

Predić, B., Jovanovic, L., Simic, V., Bacanin, N., Zivkovic, M., Spalevic, P., Budimirovic, N., & Dobrojevic, M. (2024). Cloud-load forecasting via decomposition-aided attention recurrent neural network tuned by modified particle swarm optimization. *Complex & Intelligent Systems*, *10*(2), 2249–2269. https://doi.org/10.1007/s40747-023-01265-3

Qiu, G., Song, C., Jiang, L., & Guo, Y. (2022). Multi-view hybrid recommendation model based on deep learning. *Intelligent Data Analysis*, *26*(4), 977–992. https://doi.org/10.3233/IDA-215988

Raman, N., Wahab, A. B., & Chandrasekaran, S. (2021). Computation of workflow scheduling using backpropagation neural network in cloud computing: A virtual machine placement approach. *The Journal of Supercomputing*, *77*(9), 9454–9473. https://doi.org/10.1007/s11227-021-03648-0

Rawat, P. S. (2024). Virtual machine allocation using optimal resource management approach. *Wireless Personal Communications*, *137*(2), 1313–1332. https://doi.org/10.1007/s11277-024-11465-w

Sahu, P., Raghavan, S., & Chandrasekaran, K. (2021). Ensemble deep neural network based quality of service prediction for cloud service recommendation. *Neurocomputing*, *465*, 476–489. https://doi.org/10.1016/j.neucom.2021.08.110

Salari-Hamzehkhani, B., Akbari, M., & Safi-Esfahani, F. (2025). Introducing an improved deep reinforcement learning algorithm for task scheduling in cloud computing. *The Journal of Supercomputing*, *81*(1), 295. https://doi.org/10.1007/s11227-024-06668-8

Samadhiya, S., & Ku, C. C.-Y. (2024). Hybrid approach to improve recommendation of cloud services for personalized qos requirements. *Electronics*, *13*(7), 1386. https://doi.org/10.3390/electronics13071386

Saxena, D., Kumar, J., Singh, A. K., Ieee, S., & Schmid, S. (2023). Performance analysis of machine learning centered workload prediction models for cloud. *IEEE Transactions on Parallel and Distributed Systems*, *34*(4), 1313–1330. https://doi.org/10.1109/TPDS.2023.3240567

Singh, V. P., Pandey, M. K., Singh, P. S., & Karthikeyan, S. (2020). An LSTM based time series forecasting framework for web services recommendation. *Computación y Sistemas*, *24*(2). https://doi.org/10.13053/cys-24-2-3402

Suresh Babu, K. T., Ashok Tingare, B., Khedkar, V., Diwan, T. D., William, P., & Badholia, A. (2024). Deep learning and its applications: A novel approach to machine learning with encoding algorithms. *2024 International Conference on Intelligent and Innovative Practices in Engineering and Management*, 1–6. https://doi.org/10.1109/IIPEM62726.2024.10925693

Wu, R.-C. (2024). Developing a deep learning-based multimodal intelligent cloud computing resource load prediction system. *EAI Endorsed Transactions on Internet of Things*, *10*. https://doi.org/10.4108/eetiot.6296

Xia, L., Huang, C., Xu, Y., Dai, P., & Bo, L. (2024). Multi-behavior graph neural networks for recommender system. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(4), 5473–5487. https://doi.org/10.1109/TNNLS.2022.3204775

Yemi Hussain, N. (2024). Deep learning architectures enabling sophisticated feature extraction and representation for complex data analysis. *International Journal of Innovative Science and Research Technology*, 2290–2300. https://doi.org/10.38124/ijisrt/IJISRT24OCT1521

Zheng, H., Xu, K., Zhang, M., Tan, H., & Li, H. (2024). Efficient resource allocation in cloud computing environments using AI-driven predictive analytics. *Applied and Computational Engineering*, *82*(1), 17–23. https://doi.org/10.54254/2755-2721/82/2024GLG0055