Deep learning for resource usage prediction and automatic scaling in Kubernetes clusters: a systematic literature review.

I. Tables with Summarized Studies

It is important to highlight that the same study may be included in more than one category, resulting in a total count higher than the number of selected studies.

A. Deep learning architetures

TABLE I: Deep learning architectures used in Kubernetes.

Categories	**Studies**
36 - Recurrent Networks (RNNs)	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]
07 - Transformers and Variants	[37], [38], [39], [40], [41], [29], [42]
09 - Convolutional Networks (CNNs)	[43], [44], [40], [14], [45], [31], [33], [36], [46]
06 - Graph Neural Networks (GNNs)	[4], [47], [48], [49], [50], [34]
06 - Other DL architectures	[51], [52], [53], [54], [55], [56]
10 - Non-DL / clássico / comparativos	[57], [58], [59], [60], [61], [62], [63], [64], [46], [55]

B. Metrics used in the prediction

TABLE II: Deep learning metrics prediction.

Categories	Studies
38 - Resource Usage	[37], [38], [1], [52], [4],, [43], [5], [44], [6], [7], [39], [47], [9], [54], [10], [11], [12], [13], [48], [49], [57], [14], [58], [16], [17], [18], [19], [61], [62], [41], [22], [24], [25], [29], [56], [31], [32], [34]
16 - Workload	[51], [2], [3], [53], [8], [15], [18], [20], [21], [63], [55], [26], [27], [30], [33], [46]
03 - Scaling / Allocation	[25], [64], [36]
04 - Performance	[40], [50], [45], [42]
02- Custom Metrics	[60], [64]
03 - Other / Experimental	[59], [23], [28]

C. Table III: Integration Strategies in the Kubernetes Ecosystem

TABLE III: Integration Strategies in the Kubernetes Ecosystem

Categories	Studies
34 - Real Kubernetes Integration	[37], [51], [38], [2], [3], [53], [44], [7], [39], [40], [48], [49], [14], [15], [58], [59], [60], [50], [50], [19], [20], [61], [41], [22], [24], [26], [27], [64], [28], [29], [42], [30], [56], [35]
19 - Simulation or Other Platforms	[1], [43], [5], [6], [8], [47], [54], [10], [57], [21], [62], [63], [45], [23], [25], [31], [32], [36], [46]
11 - Prediction Only	[52], [4], [9], [11], [12], [13], [59], [18], [55], [33], [34]

D. Validation Metrics

TABLE IV: Metrics Validation

Categories	Studies
57 - Prediction Quality	[37], [51], [38], [1], [2], [52], [3], [4], [53], [43], [5], [44], [6], [7], [8], [39], [47], [9], [54], [40], [10], [11], [12], [13], [57], [14], [59], [60], [50], [17], [18], [21], [62], [41], [63], [55], [22], [45], [23], [25], [26], [27], [64], [29], [42], [30], [56], [31], [32], [33], [34], [36], [46]
34 - Application Performance (SLO/SLA)	[37], [38], [2], [3], [53], [43], [5], [44], [6], [39], [47], [40], [10], [48], [49], [57], [14], [15], [50], [17], [19], [20], [61], [21], [62], [41], [45], [26], [27], [64], [29], [42], [30], [35]
33 - Resource Efficiency	[37], [38], [1], [52], [3], [53], [43], [5], [44], [6], [39], [47], [54], [10], [48], [49], [57], [14], [15], [17], [19], [61], [62], [41], [55], [45], [23], [25], [26], [64], [29], [31], [36]
13 - Scaling Behavior	[37], [38], [2], [53], [39], [10], [48] [49], [60], [41], [63], [25], [64]
6 - Classification Metrics	[37], [43], [5], [45], [23], [56]

REFERENCES

[1] S. Zheng, F. Huang, C. Li, and H. Wang, "A cloud resource prediction and migration method for container scheduling," pp. 76–80, 2021.

[2] Y. Zhang, Y. Sun, C. Song, and P. Gao, "A container cloud elastic scaling method based on gru attention mechanism," pp. 290–293, 2024.

[3] M. Lakshmanan, "A theoretical framework for autoscaling and resource allocation in elastic stream processing," pp. 1072–1078, 2025.

[4] K. Mu, X. Lv, G. Chen, and X. Li, "Active elastic scaling strategy based on spatio-temporal graph neural network," pp. 01–04, 2025.

[5] S. S. Sefati, M. Keymasi, R. Craciunescu, S. Maiduc, M. Bayram, and B. Arasteh, "Adaptive resource scheduling in multi-cloud computing using recurrent neural forecasting and memory-based metaheuristic optimization: Adaptive resource scheduling in multi-cloud computing..." *J. Grid Comput.*, vol. 23.0, 2025. [Online]. Available: https://doi.org/10.1007/s10723-025-09812-7

[6] D. Bansal, P. Singh, and A. Singh, "An analysis of cloud cost optimization using machine learning," pp. 1–6, 2025.

[7] L. S. Hettiarachchi, S. V. Jayadeva, R. A. V. Bandara, D. Palliyaguruge, U. S. S. S. Arachchillage, and D. Kasthurirathna, "Artificial intelligence-based centralized resource management application for distributed systems," pp. 1–6, 2022.

[8] A. Bali, Y. E. Houm, A. Gherbi, and M. Cheriet, "Automatic data featurization for enhanced proactive service auto-scaling: Boosting forecasting accuracy and mitigating oscillation," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36.0, 2024. [Online]. Available: https://doi.org/10.1016/j.jksuci.2024.101924

[9] N. Zhang, Y. Su, B. Wu, X. Tu, Y. Jin, and X. Bao, "Cloud resource prediction model based on lstm and rbf," pp. 189–194, 2021.

[10] D. Chen, B. Shen, and Y. Chen, "Conlar: Learning to allocate resources to docker containers under time-varying workloads," in *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, 2021, pp. 458–469.

[11] H. Zhang, J. Zhang, J. Zeng, and Z. Zhu, "Container load prediction method based on a hybrid model of arima and bi-lstm," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 13692.0, p. Academic Exchange In, 2025. [Online]. Available: http://dx.doi.org/10.1117/12.3068497

[12] Amirullah and A. Saikhu, "Cpu usage forecasting for load balancing in kubernetes using lstm: A synthetic traffic simulation approach," pp. 277–282, 2025.

[13] M. P. J. Kuranage, L. Nuaymi, A. Bouabdallah, T. Ferrandiz, and P. Bertin, "Deep learning based resource forecasting for 5g core network scaling in kubernetes environment," *Proceedings of the 2022 IEEE International Conference on Network Softwarization: Network Softwarization Coming of Age: New Challenges and Opportunities, NetSoft 2022*, pp. 139 – 144, 2022. [Online]. Available: http://dx.doi.org/10.1109/NetSoft54395.2022.9844056

[14] L. Wang, X. Fan, Y. Li, Q. Du, J. She, and Q. Li, "Dynamic microservice resource optimization management based on mape loop," in *Proceedings of the 16th International Conference on Internetware*, ser. Internetware '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 578–588. [Online]. Available: https://doi.org/10.1145/3755881.3755898

[15] A. Lipari, G. P. Mattia, and R. Beraldi, "Dynamic and forecast-based containers autoscaling for kubernetes with reinforcement learning," *Proceedings - 2025 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2025*, pp. 1081 – 1088, 2025. [Online]. Available: http://dx.doi.org/10.1109/IPDPSW66978.2025.00169

[16] L. S. Hettiarachchi, S. V. Jayadeva, R. A. V. Bandara, D. Palliyaguruge, U. S. S. Arachchillage, and D. Kasthurirathna, "Expert system for kubernetes cluster autoscaling and resource management," pp. 174–179, 2022.

[17] P. Liu, W. Zhao, B. Zhang, and J. Wang, "Hybrid elastic scaling strategy for container cloud based on load prediction and reinforcement learning," *Journal of Physics: Conference Series*, vol. 2732.0, 2024. [Online]. Available: http://dx.doi.org/10.1088/1742-6596/2732/1/012014

[18] B. Vennala, D. Suresh, P. Samiya, M. Suresh, and B. Vamsi, "Integrating artificial intelligence with cloud platforms to optimize performance, scalability, and reliability in distributed computing systems," pp. 1–6, 2025.

[19] J. Violos, S. Tsanakas, T. Theodoropoulos, A. Leivadeas, K. Tserpes, and T. Varvarigou, "Intelligent horizontal autoscaling in edge computing using a double tower neural network," *Comput. Netw.*, vol. 217.0, 2022. [Online]. Available: https://doi.org/10.1016/j.comnet.2022.109339

[20] B. Tang, W. Xu, L. Zhang, B. Cao, M. Tang, and Q. Yang, "Location-aware dynamic scaling of microservices in mobile edge computing," *IEEE Transactions on Network and Service Management*, vol. 22.0, pp. 4288–4301, 2025.

[21] Z. Peng, B. Tang, W. Xu, Q. Yang, E. Hussaini, Y. Xiao, and H. Li, "Microservice auto-scaling algorithm based on workload prediction in cloud-edge collaboration environment," pp. 608–615, 2023.

[22] B. Kumar, A. Verma, and P. Verma, "Optimizing resource allocation using proactive scaling with predictive models and custom resources," *Comput. Electr. Eng.*, vol. 118.0, 2024. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2024.109419

[23] A. Kuity and S. K. Peddoju, "phpce: a hybrid power conservation approach for containerized hpc environment," *Cluster Computing*, vol. 27.0, p. 2611–2634, 2023. [Online]. Available: https://doi.org/10.1007/s10586-023-04105-8

[24] N. K. M R, A. B, H. J, S. Srinivasan, and S. S. Sand, "Pod scheduling and proactive resource management in an edge cluster using mcdm and federated learning," *J. Grid Comput.*, vol. 23.0, 2025. [Online]. Available: https://doi.org/10.1007/s10723-025-09811-8

[25] A. John, J. Kawash, and R. Alhajj, "Predictive container orchestration in the cloud using artificial intelligence techniques," *Computing*, vol. 107.0, 2025. [Online]. Available: http://dx.doi.org/10.1007/s00607-025-01505-z

[26] D.-D. Vu, M.-N. Tran, and Y. Kim, "Predictive hybrid autoscaling for containerized applications," *IEEE Access*, vol. 10.0, pp. 109 768–109 778, 2022.

[27] J. Dogani and F. Khunjush, "Proactive auto-scaling technique for web applications in container-based edge computing using federated learning model," *SSRN*, 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4358737

[28] X. Fan, Z. Huang, S. Huo, and X. Zhang, "Research on multi-protocol access method for iot platform based on microservice container scheduling," *2023 8th International Conference on Intelligent Computing and Signal Processing, ICSP 2023*, pp. 651 – 655, 2023. [Online]. Available: http://dx.doi.org/10.1109/ICSP58490.2023.10248790

[29] L. Wen, M. Xu, A. N. Toosi, and K. Ye, "Temposcale: A cloud workloads prediction approach integrating short-term and long-term information," pp. 183–193, 2024.

[30] J. Sheng, Z. Tang, J. Guo, and T. Wang, "Adaptive configuration selection for multi-model inference pipelines in edge computing," pp. 729–736, 2024.

[31] G. Indumathi and R. Sarala, "Fire hawk optimization-enabled deep learning scheme based hybrid cloud container architecture for migrating interoperability based application," *China Communications*, vol. 22.0, pp. 285–304, 2025.

[32] R. S and S. Soma, "Gradient beetle honey badger optimization based container migration and optimal key selection using firefly honey badger algorithm in cloud computing," pp. 1–6, 2023.

[33] J. Dogani, F. Khunjush, M. R. Mahmoudi, and M. Seydali, "Multivariate workload and resource prediction in cloud computing using cnn and gru by attention mechanism," *J. Supercomput.*, vol. 79.0, p. 3437–3470, 2022. [Online]. Available: https://doi.org/10.1007/s11227-022-04782-z

[34] R. She, X. Jia, J. Yan, and W. Li, "Research on cloud computing microservice resource management strategy based on graphgru," *2024 5th International Conference on Computer Engineering and Application, ICCEA 2024*, pp. 822 – 825, 2024. [Online]. Available: http://dx.doi.org/10.1109/ICCEA62105.2024.10604110

[35] H. Jia, Z. Li, G. Li, M. Xu, and K. Ye, "Sealos+: A sealos-based approach for adaptive resource optimization under dynamic workloads for securities trading system," pp. 1–9, 2025.

[36] M. Kumar, P. SahooAshok, and K. T. Misra, "Task grouping and optimized deep learning based vm sizing for hosting containers as a service," *Journal of Cloud Computing*, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s13677-023-00441-7

[37] X. Huang, X. Liao, J. Yang, W. You, W. Wu, S. Min, and X. Ji, "5g-ppde: A novel adaptive scaling framework for enhancing the resilience of the 5g cloud core network," pp. 2025–2030, 2024.

[38] B. Kumar, A. Verma, and P. Verma, "A multivariate transformer-based monitor-analyze-plan-execute (mape) autoscaling framework for dynamic resource allocation in cloud environment," *Computing*, vol. 107.0, 2025. [Online]. Available: https://doi.org/10.1007/s00607-025-01426-x

[39] F. a. Meng, H. Dai, G. Cong, B. Zhu, and H. Zhao, "Catscaler: A convolution-augmented transformer scaling framework for cloud-native applications," *IEEE Transactions on Services Computing*, vol. 18.0, pp. 2659 – 2672, 2025. [Online]. Available: http://dx.doi.org/10.1109/TSC.2025.3592383

[40] R.-M. Ursu, N. Asadi, J. Zerwas, L. Wong, and W. Kellerer, "Comparative analysis between decentralized and centralized network digital twins of kubernetes clusters," pp. 137–145, 2025.

[41] J. Chen, X. He, H. Ye, F. Jiang, T. Zhang, J. Chen, and X. Gao, "Online ensemble transformer for accurate cloud workload forecasting in predictive auto-scaling," *arXiv*, 2025. [Online]. Available: http://dx.doi.org/10.48550/arXiv.2508.12773

[42] Y. Chen, J. Hao, Y. Peng, and H. Xia, "Transformer-based performance prediction and proactive resource allocation for cloud-native microservices," *Cluster Computing*, vol. 28.0, 2025. [Online]. Available: https://doi.org/10.1007/s10586-025-05237-9

[43] I. G and S. R, "Adaptive resource prediction in containerized cloud environments with cnn-lstm optimized by sca," pp. 1–6, 2024.

[44] S. Chouliaras and S. Sotiriadis, "An adaptive auto-scaling framework for cloud resource provisioning," *Future Gener. Comput. Syst.*, vol. 148.0, p. 173–183, 2023. [Online]. Available: https://doi.org/10.1016/j.future.2023.05.017

[45] Y. Peng, J. Hao, and Y. Chen, "Performance prediction and resource adaptive adjustment for cloud-native microservices," *Cluster Computing*, vol. 28.0, 2025. [Online]. Available: https://doi.org/10.1007/s10586-025-05437-3

[46] S. Stefan and V. Niculescu, "Microservice-oriented workload prediction using deep learning," *E-Informatica Software Engineering Journal*, vol. 16.0, 2022. [Online]. Available: http://dx.doi.org/10.37190/e-Inf220107

[47] Z. Qiu, W. Deng, J. Huang, X. Liu, B. Ren, and S. Qin, "Cloud platform resource capacity management and scheduling optimization based on data graph," p. 189–192, 2025. [Online]. Available: https://doi.org/10.1145/3727505.3727537

[48] C. Meng, S. Song, H. Tong, M. Pan, and Y. Yu, "Deepscaler: Holistic autoscaling for microservices based on spatiotemporal gnn with adaptive graph learning," p. 53–65, 2024. [Online]. Available: https://doi.org/10.1109/ASE56229.2023.00038

[49] Z. Wang, S. Zhu, J. Li, W. Jiang, K. K. Ramakrishnan, M. Yan, X. Zhang, and A. X. Liu, "Deepscaling: Autoscaling microservices with stable cpu utilization for large scale production cloud systems," *IEEE/ACM Trans. Netw.*, vol. 32.0, p. 3961–3976, 2024. [Online]. Available: https://doi.org/10.1109/TNET.2024.3400953

[50] J. Park, B. Choi, C. Lee, and D. Han, "Graph neural network-based slo-aware proactive resource autoscaling framework for microservices," *IEEE/ACM Trans. Netw.*, vol. 32.0, p. 3331–3346, 2024. [Online]. Available: https://doi.org/10.1109/TNET.2024.3393427

[51] B. Jeon, C. Wang, D. Arroyo, A. Youssef, and I. Gupta, "A house united within itself: Slo-awareness for on-premises containerized ml inference clusters via faro," p. 524–540, 2025. [Online]. Available: https://doi.org/10.1145/3689031.3696071

[52] S. Chen, X. Mo, and W. Wu, "A flexible model for predicting workloads of massive microservices," *CACML 2025 - 2025 4th Asia Conference on Algorithms, Computing and Machine Learning*, pp. IEEE –, 2025. [Online]. Available: http://dx.doi.org/10.1109/CACML64929.2025.11010955

[53] L. M. D. da Silva, P. V. A. Alves, S. N. Silva, and M. A. C. Fernandes, "Adaptive horizontal scaling in kubernetes clusters with ann-based load forecasting," *Cluster Computing*, vol. 28.0, 2025. [Online]. Available: https://doi.org/10.1007/s10586-024-04887-5

[54] S. Park and H. Bahn, "Combining genetic algorithms and bayesian neural networks for resource usage prediction in multi-tenant container environments," *Cluster Computing*, vol. 28.0, 2025. [Online]. Available: http://dx.doi.org/10.1007/s10586-024-04832-6

[55] I. G and S. R, "Optimization enabled deep learning methods for container-based cloud computing environment," in *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, 2023, pp. 1–6.

[56] J. S. Camargo, E. Coronado, W. Ramirez, D. Camps, S. S. Deutsch, J. Pérez-Romero, A. Antonopoulos, O. Trullols-Cruces, S. Gonzalez-Diaz, B. Otura, and G. Rigazzi, "Dynamic slicing reconfiguration for virtualized 5g networks using ml forecasting of computing capacity," *Comput. Netw.*, vol. 236.0, 2023. [Online]. Available: https://doi.org/10.1016/j.comnet.2023.110001

[57] L. Vinícius, L. Rodrigues, M. Torquato, and F. A. Silva, "Docker platform aging: a systematic performance evaluation and prediction of resource consumption," *J. Supercomput.*, vol. 78.0, p. 12898–12928, 2022. [Online]. Available: https://doi.org/10.1007/s11227-022-04389-4

[58] K. Q. Pham and T. Kim, "Elastic federated learning with kubernetes vertical pod autoscaler for edge computing," *Future Gener. Comput. Syst.*, vol. 158, no. C, p. 501–515, Sep. 2024. [Online]. Available: https://doi.org/10.1016/j.future.2024.04.047

[59] I. Mahajan and D. Nadig, "Enhancing workload predictions using service interactions in cloud-native microservices," *2024 IEEE 13th International Conference on Cloud Networking, CloudNet 2024*, p. Comite Gestor da Int, 2024. [Online]. Available: http://dx.doi.org/10.1109/CloudNet62863.2024.10815917

[60] Y. X. Chia, C. K. Seow, K. Chen, and Q. Cao, "Exploring resource prediction models based on custom kubernetes auto-scaling metrics," pp. 47–52, 2024.

[61] A. Rubak and J. Taheri, "Machine learning for predictive resource scaling of microservices on kubernetes platforms," 2024. [Online]. Available: https://doi.org/10.1145/3603166.3632165

[62] S. Bawa, P. S. Rana, and R. Tekchandani, "Multivariate time series ensemble model for load prediction on hosts using anomaly detection techniques," *Cluster Computing*, vol. 27.0, p. 10993–11016, 2024. [Online]. Available: https://doi.org/10.1007/s10586-024-04517-0

[63] T. P. d. Silva, A. R. Neto, T. V. Batista, F. C. Delicato, P. F. Pires, and F. Lopes, "Online machine learning for auto-scaling in the edge computing," *Pervasive Mob. Comput.*, vol. 87.0, 2022. [Online]. Available: https://doi.org/10.1016/j.pmcj.2022.101722

[64] G. Marques, C. Senna, S. Sargento, L. Carvalho, L. Pereira, and R. Matos, "Proactive resource management for cloud of services environments," *Future Gener. Comput. Syst.*, vol. 150.0, p. 90–102, 2024. [Online]. Available: https://doi.org/10.1016/j.future.2023.08.005