AI-enabled Systems in Edge-cloud Environments

In-Young Ko[1], Michael Mrissa[2], Juan Manuel Murillo[3,4] and Abhishek Srivastava[5]

[1]*Korea Advanced Institute of Science and Technology, South Korea*
[2]*InnoRenew CoE, UP IAM & UP FAMNIT, University of Primorska, Slovenia*
[3]*University of Extremadura, Spain*
[4]*COMPUTAEX Foundation, Spain*
[5]*Indian Institute of Technology Indore, India*
E-mail: iko@kaist.ac.kr; michael.mrissa@innorenew.eu; juanmamu@unex.es; asrivastava@iiti.ac.in

In the era of artificial intelligence (AI), smart devices such as autonomous vehicles, drones, and service robots are increasingly collaborating to perform complex tasks for humans. However, the centralized control and optimization of these widely distributed devices suffer from significant scalability limitations. At the same time, traditional cloud infrastructures are struggling to meet the demands of collecting and processing massive volumes of data from countless devices, often leading to increased latency and reduced service responsiveness.

To address these challenges, edge-cloud computing has emerged as a promising infrastructure that enhances efficiency, scalability, and data privacy for delivering data-centric, AI-enabled services. In an edge-cloud ecosystem, multiple computing tiers – including edge devices, fog nodes, and centralized clouds – collaborate to support data collection, processing, and decision-making closer to data sources. These tiers must coordinate dynamically, considering available computing resources while ensuring service quality, safety, and accuracy.

The fifth edition of the International Workshop on Big Data-driven Edge Cloud Services (BECS 2025)[1] aims to bring together researchers and practitioners to exchange ideas, share experiences, and explore recent advances in developing intelligent, data-driven services for edge-cloud environments. The workshop was held in conjunction with the 25th International Conference on Web Engineering (ICWE 2025)[2].

Topics of interest for BECS 2025 include, but are not limited to: web services in edge-cloud environments; Web of Things and smart device integration; AI and machine learning at the edge (edge AI); reliable and user-friendly big data platforms; distributed data collection, analysis, and prediction; stream data processing in edge infrastructures; knowledge graphs for distributed edge-cloud computing; modeling, composition, and mashups of edge-cloud services; microservices architectures for edge-cloud environments; and strategies for edge-cloud collaboration and orchestration.

This special issue of the Journal of Web Engineering focuses on improving the safety and efficiency of AI-enabled systems by leveraging the unique characteristics of distributed edge-cloud environments. For this issue, we selected papers from BECS 2025 that propose conceptual frameworks and technical approaches aimed at addressing real-world challenges through edge-cloud computing.

The first article, "Data-driven Adaptive ML-Enabled Edge-cloud System Framework for Safe and Efficient Autonomous Systems," authored by Eunho Cho and In-Young Ko, presents an adaptive machine learning (ML) framework for autonomous driving systems (ADSs) operating in edge-cloud environments. Modern ADSs rely heavily on ML for perception, decision-making, and control; however, static models often struggle to cope with the diverse and unpredictable conditions encountered in real-world settings, particularly under the resource constraints of edge environments. Existing adaptive approaches typically assume white-box models, which limits their applicability to complex black-box systems such as deep neural networks. To address these limitations, the authors propose a two-phase edge-cloud collaboration framework that integrates cloud-based pre-runtime analysis with runtime adaptation across edge and cloud resources. This framework enables the dynamic selection of ML systems based on real-time environmental and system conditions, thereby improving both safety and efficiency. The proposed approach is validated through a prototype implementation on

[1] https://becs.kaist.ac.kr/iwbecs2025/

[2] https://icwe2025.webengineering.org/

the CARLA simulation platform, demonstrating the potential of combining cloud-scale computational power with edge-level responsiveness to support adaptive, black-box ML in ADSs.

In the second article, "Towards Real-time Underwater Object Detection and Identification: Integrating Acoustic Sensing with Edge Computing," Shekhar Tyagi, Akshat Shah, and Abhishek Srivastava present a novel approach to underwater object detection that addresses key limitations of existing methods. The authors observe that vision-driven machine learning models, such as CNN- and YOLO-based approaches, often underperform in low-visibility or deep-water environments. While traditional tools – including sonar systems, imaging technologies, and autonomous underwater vehicles – are widely used, they face challenges related to signal interference, high energy consumption, and environmental constraints. To overcome these issues, the authors propose an enhanced method based on underwater acoustic sensor (UAS) networks, integrating bathymetric analysis, path loss modeling, and advanced detection techniques such as Delaunay's Convex Hull-based point cloud reconstruction, the law of magnetic equilibrium, and the Doppler effect. These components enable accurate identification of object type, shape, location, and motion state, even in complex underwater environments. Real-time data transmission to surface anchors further supports efficient and reliable underwater monitoring, with potential applications in biodiversity research, defense, and environmental conservation.

Maintaining quality of service (QoS) in edge-cloud environments is critical yet challenging due to their highly dynamic nature. The third article, "LLM-driven Multi-agent Architecture for QoS-aware Server Recommendation in Mobile–Edge–Cloud Environments," by Eunjeong Ju, Jeonghwa Lee, Duksan Ryu, Suntae Kim, and Jongmoon Baik, introduces a context-aware, multi-agent server recommendation framework for mobile edge computing (MEC) environments. The framework targets latency-sensitive and compute-intensive applications such as real-time streaming and augmented reality. Traditional server selection approaches often rely on static heuristics, such as proximity or average QoS metrics, which fail to capture user intent and application-specific requirements. To address this limitation, the authors propose a novel system that leverages the reasoning capabilities of large language models (LLMs) to infer user QoS preferences, predict future mobility and server conditions, and recommend optimal edge servers in a flexible and explainable manner. Experimental results demonstrate that the proposed framework effectively adapts to changing conditions and outperforms conventional strategies in both user satisfaction and selection accuracy.

Digital twins can be maintained and utilized more efficiently in distributed edge-cloud environments, particularly in critical domains such as healthcare. The fourth article, "Learning by Experiencing: An Immersive Digital Twin Tool for ECG Education," by Daniel Flores-Martin, Francisco Díaz-Barrancas, Pedro J. Pardo, Javier Berrocal, and Juan M. Murillo, introduces ECGTwinMentor, an interactive digital twin platform designed to enhance electrocardiogram (ECG) education. Despite advances in cardiology and AI-based diagnostics, medical trainees often face difficulties acquiring practical ECG interpretation skills due to limited data access, reduced clinical exposure, and outdated teaching methods. ECGTwinMentor addresses these challenges by providing a web-based, immersive learning environment that simulates real-world cardiac scenarios. The platform combines a deep learning model for cardiopathy prediction with dynamic ECG simulations, enabling students to practice diagnostic reasoning and pattern recognition in real time. Its modular and scalable design supports deployment on both cloud and edge devices, making hands-on learning accessible across diverse educational settings.

Ensuring software reliability is another critical challenge in edge-cloud environments, where data scarcity, rapid project evolution, and limited historical defect information can increase the risk of software defects. In the fifth article, titled "Project Evolution-aware Prompting of LLMs for Just-in-time Defect Prediction in Edge-cloud Systems," authors Inseok Yeo, Sungu Lee, Duksan Ryu, and Jongmoon Baik tackle this issue by focusing on just-in-time (JIT) defect prediction, which identifies potentially faulty code changes at the time of commit. JIT prediction is particularly valuable in distributed environments like edge-cloud systems, as it helps prioritize testing efforts. However, existing JIT models often depend on large labeled datasets and perform poorly in cross-project scenarios due to data scarcity and distribution shifts. To address these limitations, the authors propose PROPER-SDP, a prompt-based approach that leverages pretrained LLMs. By enriching prompts with project documentation and evolution history, PROPER-SDP captures project-specific context and delivers high-accuracy defect predictions without requiring any additional training data – outperforming existing baseline models.

The final article, "Spatio-temporal Mamba for User Mobility Prediction in Mobile Edge Computing," by Jeonghwa Lee, Eunjeong Ju, Duksan Ryu, Suntae Kim, and Jongmoon Baik, introduces BERT-M, a novel mobility prediction model built on the BERT architecture. Aimed at improving quality of service in MEC environments, BERT-M addresses the limitations

of traditional long short-term memory (LSTM)-based approaches, which often require complex preprocessing and explicit spatial distance matrices. By directly learning spatio-temporal patterns from user mobility data, BERT-M offers a more scalable and adaptable solution well suited for real-time edge-cloud systems.

Together, the articles in this special issue explore diverse approaches to enhancing the safety and efficiency of AI-enabled systems in edge-cloud environments. Beyond technical innovations, they offer valuable insights grounded in practical applications, highlighting the growing importance of edge-cloud computing in addressing real-world challenges.

Acknowledgment

The BECS 2025 workshop was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201795).