# Edge Computing: AI on devices without cloud reliance

**Riya Rangwani**
Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

**Satwik Tripathy**
Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

**Abstract:**

The growing use of artificial intelligence (AI) in real-world applications has exposed several limitations of traditional cloud-centric architectures, including high inference latency, heavy bandwidth consumption, privacy concerns, and reduced reliability in environments with unstable or limited network connectivity. These issues are especially critical for time-sensitive and mission-critical systems that require real-time responses. Edge computing addresses these challenges by enabling AI inference directly on end devices such as IoT nodes, mobile platforms, and embedded systems, thereby reducing or eliminating continuous dependence on the cloud. This study examines the design and implementation of cloud-independent edge AI systems, focusing on efficient on-device inference, lightweight model architectures, and optimization techniques suitable for resource-constrained hardware. An end-to-end edge AI pipeline is presented, covering local data acquisition, on-device preprocessing, optimized model deployment, and real-time decision-making at the edge. Experimental results demonstrate that edge-based AI significantly reduces response latency, improves data privacy, and enhances system robustness while maintaining accuracy comparable to cloud-based solutions, particularly in scenarios with intermittent or no connectivity. These findings highlight edge AI as a practical and scalable solution for applications such as autonomous systems, healthcare monitoring, smart surveillance, and industrial automation, and position edge computing as a foundational technology for next-generation decentralized AI systems.

**Keywords:**

Decentralized intelligence, edge ai, edge computing, embedded systems, internet of things (iot), low-latency computing, on-device intelligence, real-time ai, resource-constrained ai.