III B.Tech II Semester

		L	T	P	C
23A39602	**AI FOR EDGE COMPUTING** (**Professional Core**)	**3**	**0**	**0**	**3**

Course Objectives

- Introduce students to the fundamentals and applications of Edge Computing and how AI is integrated at the edge.
- Enable learners to understand the design and deployment of AI models on edge devices.
- Familiarize students with energy-efficient, latency-aware, and privacy-preserving AI systems in edge environments.
- Equip students with hands-on skills in frameworks, tools, and platforms that support edge-based AI deployments.
- Address challenges like model compression, edge-cloud collaboration, and inference acceleration at the edge.

Course Outcomes

- Describe the fundamentals of Edge Computing and its relationship with Artificial Intelligence.
- Analyze edge-centric architectures and frameworks suitable for AI workloads.
- Design and deploy optimized AI models on edge devices considering resource constraints.
- Apply real-time data analytics and AI inference on edge nodes with minimal latency.
- Explore future directions and open research areas in edge-based AI systems and applications.

UNIT I – Introduction to Edge Computing and AI

Evolution of Computing Paradigms: Cloud, Fog, and Edge, Introduction to Edge AI – Concepts and Motivation, Architecture of Edge Computing Systems, Differences between Edge AI and Cloud AI, Use Cases of Edge AI – Smart Cities, Healthcare, IoT, and Industry 4.0, Hardware for Edge AI – Edge GPUs, TPUs, FPGAs, Types of Edge Devices – Raspberry Pi, Jetson Nano, Coral, Challenges in Deploying AI on Edge

UNIT II – AI Models and Edge Inference

Types of AI Models Suitable for Edge Deployment, Model Optimization Techniques: Quantization, Pruning, Distillation, Transfer Learning for Edge AI, Inference Acceleration using Edge Hardware, Lightweight Models: MobileNet, SqueezeNet, TinyML, Frameworks for Edge Deployment: TensorFlow Lite, ONNX, OpenVINO, Compilation Tools: TVM, Glow, Energy and Latency-aware Inference.

UNIT III – Edge-Centric Architectures and Data Management

Distributed AI Architectures: Edge, Fog, and Cloud, Collaborative Intelligence – Edge-Cloud Offloading Strategies, Data Lifecycle in Edge AI Systems, Real-time Stream Processing at the Edge, Data Compression and Fusion Techniques, Caching and Scheduling Mechanisms, Privacy-Preserving Edge AI (Federated Learning, Differential Privacy), Case Study: Real-time Video Analytics using Edge Devices,

UNIT IV – Security, Privacy, and Ethical Aspects of AI at the Edge

Security Threats in Edge Environments, Privacy Concerns with AI Inference on Personal Devices, Federated Learning: Concepts and Frameworks (e.g., Flower, TensorFlow Federated), Data Anonymization and Encryption Techniques, Blockchain for Secure Edge AI, Explainable AI for Edge Decisions, Regulations and Ethical Challenges in Edge AI, Case Studies on Privacy-Aware AI Systems.

UNIT V – Applications and Future Trends in Edge AI

Edge AI in Autonomous Vehicles, Industrial Automation and Predictive Maintenance, AI-Driven Surveillance and Smart Homes, Edge AI in Healthcare Monitoring Systems, 5G and Edge AI Integration, Emerging Trends: TinyML, Neuromorphic Computing, Benchmarking Tools for Edge AI Performance, Future Research Directions and Innovation Opportunities.

Textbooks

1. "Artificial Intelligence at the Edge" by Daniel Situnayake & Pete Warden
2. "Edge Computing: A Primer" by Jie Cao, Weisong Shi, and Qun Li
3. "TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers" by Pete Warden and Daniel Situnayake

Reference Books

1. "Designing Distributed Systems" by Brendan Burns (O'Reilly)
2. "Hands-On Edge Analytics with Azure IoT" by Abhishek Kumar
3. Recent IEEE and ACM journal publications on Edge AI and Federated Learning

Online Courses

1. TinyML Specialization – Harvard & Google (edX)
2. AI for Edge Computing – NPTEL
3. Federated Learning – Coursera (Intel & University of Illinois)