Hochschule München University of Applied Sciences

Courses in English
Course Description

Department	07 Computer Science and Mathematics
Course title	**Edge AI and TinyML**
Course number	
Hours per week (SWS)	4
Number of ECTS credits	5
Course objective	Students acquire knowledge about methods, tools, and hardware platforms for developing, optimizing, and deploying AI applications on edge systems. They understand the differences between Edge AI on more powerful embedded platforms and TinyML on extremely resource-constrained microcontrollers. Upon successful completion of the module, students will be able to: explain the fundamentals of Edge AI and TinyML, categorize their areas of application, and evaluate the use of edge, cloud, and hybrid inference with respect to latency, energy efficiency, data privacy, and resource requirements, analyze the characteristics of different edge hardware platforms and select appropriate hardware and software solutions for specific application scenarios, specifically optimize machine learning models for resource-constrained edge systems using quantization, pruning, clustering, and knowledge distillation, develop embedded AI applications, implement the preprocessing of sensor data, and deploy optimized models on edge devices and microcontrollers, • measure, analyze, and evaluate the performance of edge AI applications with regard to accuracy, latency, memory requirements, and energy efficiency, and derive appropriate optimization measures.
Prerequisites	Basic programming skills (Python, C/C++); basic knowledge in the field of Machine Learning / Deep Learning; basic knowledge of microcontroller technology and computer architectures; experience in using software development tools (IDE, CLI, debugging).
Recommended reading	
Teaching methods	
Assessment methods	oral exam
Language of instruction	English
Name of lecturer	
Email	
Link	https://zpa.cs.hm.edu/public/module/532/
Course content	Fundamentals of Edge AI and TinyML: Terms, concepts, differentiation from cloud AI, and fields of application; comparison of architectures; evaluation based on latency, energy efficiency, data privacy, and resource requirements Edge Hardware Platforms: Microcontrollers, NPUs/TPUs, single-board computers; comparison of processing power, memory, and energy consumption Software Ecosystems and Toolchains: Frameworks (e.g., TensorFlow Lite Micro, Edge Impulse, CMSIS-NN, ONNX Runtime), development workflow Model Optimization: Quantization, pruning, clustering, knowledge distillation; analysis regarding size, accuracy, and speed Sensor Data Processing: Acquisition and preprocessing (audio, image, acceleration), feature extraction under resource constraints Development and Deployment: Model training, conversion, integration; deployment on microcontrollers/edge devices Performance Analysis: Accuracy, latency, memory, energy consumption; benchmarking methods; derivation of optimization measures
Remarks	