Minor in AI

TinyML

Introduction

May 21, 2025

1 TinyML

Definition

TinyML stands for *Tiny Machine Learning*, a subfield of machine learning that focuses on developing models capable of running on small, low-power hardware devices such as microcontrollers, typically using less than 1 mW of power. These devices often have:

- Memory constraints (as low as 32 KB RAM)
- Low computational power (MHz-range processors)
- No operating system or only lightweight RTOS

TinyML enables real-time, on-device inference without the need to communicate with the cloud, allowing for ultra-low latency and privacy-preserving applications.

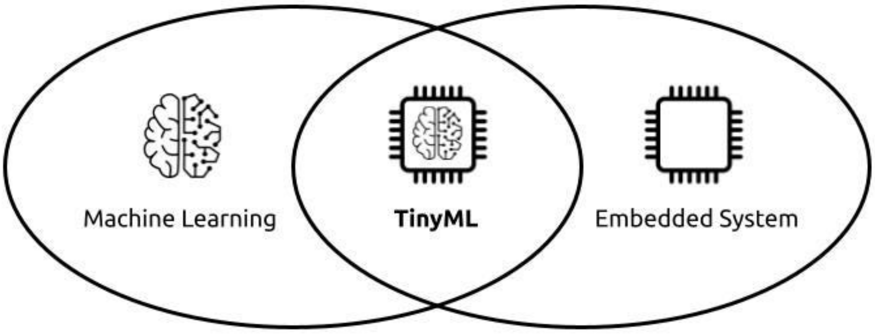

Figure 1: TinyML

2 Characteristics

1. **Low Power Consumption**: Models are optimized to run on less than 1 mW, making them ideal for battery-operated or energy-harvesting devices.
2. **Small Memory Footprint**: Models must fit within kilobytes of RAM and flash memory.
3. **On-device Inference**: No need for internet/cloud connection, leading to lower latency and improved privacy.
4. **Low Latency**: Real-time decision-making, often in milliseconds.
5. **Efficient Deployment**: Models are deployed to microcontrollers using frameworks like TensorFlow Lite for Microcontrollers (TFLM), uTensor, etc.

3 Importance

TinyML has emerged as a transformative technology due to:

- **Edge AI Revolution**: Bringing intelligence to the edge reduces bandwidth, cost, and privacy concerns.
- **Ubiquitous Computing**: Enables AI in devices where it was previously impossible due to resource constraints.
- **Cost-effective AI**: Removes dependency on cloud infrastructure for inference.
- **Scalability**: Can be deployed in billions of devices (IoT sensors, wearables, etc.).
- **Environmental Impact**: Reduces energy usage and carbon footprint compared to cloud-based ML inference.

4 Applications

TinyML is being widely adopted in several domains due to its efficiency and feasibility:

1. Healthcare and Health Monitoring

- Wearable devices monitor heart rate, blood oxygen, and detect anomalies like arrhythmia.
- Sleep tracking and fall detection in elderly care.

Case Study: Edge Impulse + Arduino Nano 33 BLE Sense

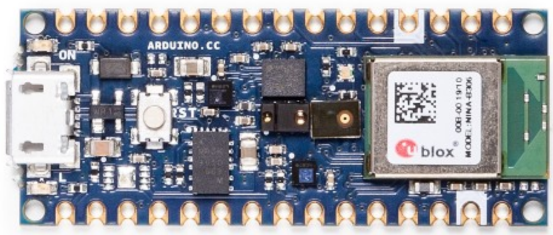

Figure 2: Arduino Nano 33 BLE Sense

Researchers built a cough detection system using an Arduino Nano with a microphone. The TinyML model runs locally and helps identify COVID-19 related symptoms without sending data to the cloud—preserving privacy and reducing costs.

2. Smart Agriculture

- Monitoring soil moisture, crop health using sound or vision-based ML models.
- Animal health monitoring using sensor data.

Case Study: Harvesting Efficiency in India

A TinyML-enabled soil moisture sensor system was deployed in rural farms to optimize irrigation. It used a microcontroller running a neural net model trained on environmental conditions and saved up to 30% water.

3. Industrial IoT (IIoT)

- Predictive maintenance of motors, compressors by analyzing vibration or sound data.
- Anomaly detection in machine behavior in real-time.

4. Consumer Electronics

- Always-on voice assistants (e.g., keyword spotting like "Hey Siri").
- Gesture recognition for smart devices.

5. Smart Cities

- Noise pollution detection and classification.
- Smart street lighting systems based on activity detection.

5 Popular Frameworks and Tools

- **TensorFlow Lite for Microcontrollers (TFLM)** – Lightweight framework for deploying models on microcontrollers.
- **Edge Impulse** – End-to-end TinyML pipeline with data collection, model training, and deployment.
- **CMSIS-NN** – ARM's optimized neural network kernels for Cortex-M processors.

6 Challenges

- **Model Compression**: Pruning, quantization and knowledge distillation are often required.
- **Limited Hardware Resources**: Needs careful hardware-aware optimization.
- **Deployment Complexity**: Requires deep understanding of both ML and embedded systems.

7 Key Takeaways

1. **TinyML enables machine learning on ultra-low-power devices**, making AI accessible at the edge without relying on the cloud.
2. It is characterized by extremely low memory and power usage, making it suitable for real-time, privacy-preserving applications.
3. **TinyML is not just a technical breakthrough—it is a catalyst for social impact**, enabling applications in healthcare, agriculture, industry, and smart infrastructure.
4. With tools like TensorFlow Lite for Microcontrollers and Edge Impulse, the TinyML ecosystem is becoming increasingly developer-friendly.
5. **The future of AI is distributed and embedded—TinyML is leading that revolution**.