# Edge AI nel mondo reale: ricerca, sfide e applicazioni

**Laura Falaschetti**

Dipartimento di Ingegneria dell'Informazione · Università Politecnica delle Marche

l.falaschetti@staff.univpm.it

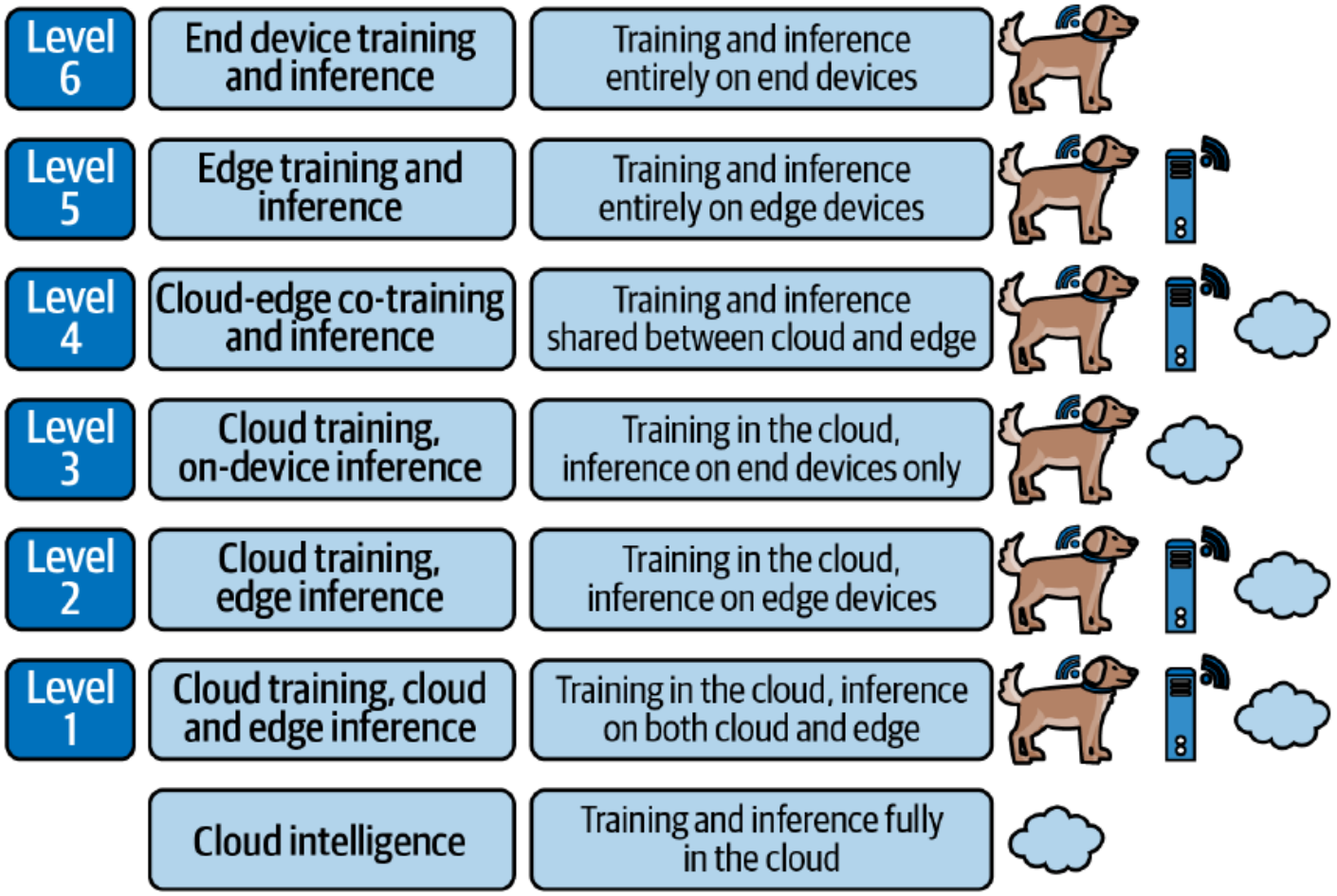
**AI@DII – Nuove Frontiere nell'Ingegneria dell'Informazione**

Ancona · 29 ottobre 2026

# Cos’è l’Edge AI e perché utilizzarla?

**DALL’ELABORAZIONE NEL CLOUD ALL’INTELLIGENZA DISTRIBUITA**

Training e inferenza possono essere distribuiti tra cloud, nodi edge e dispositivi.



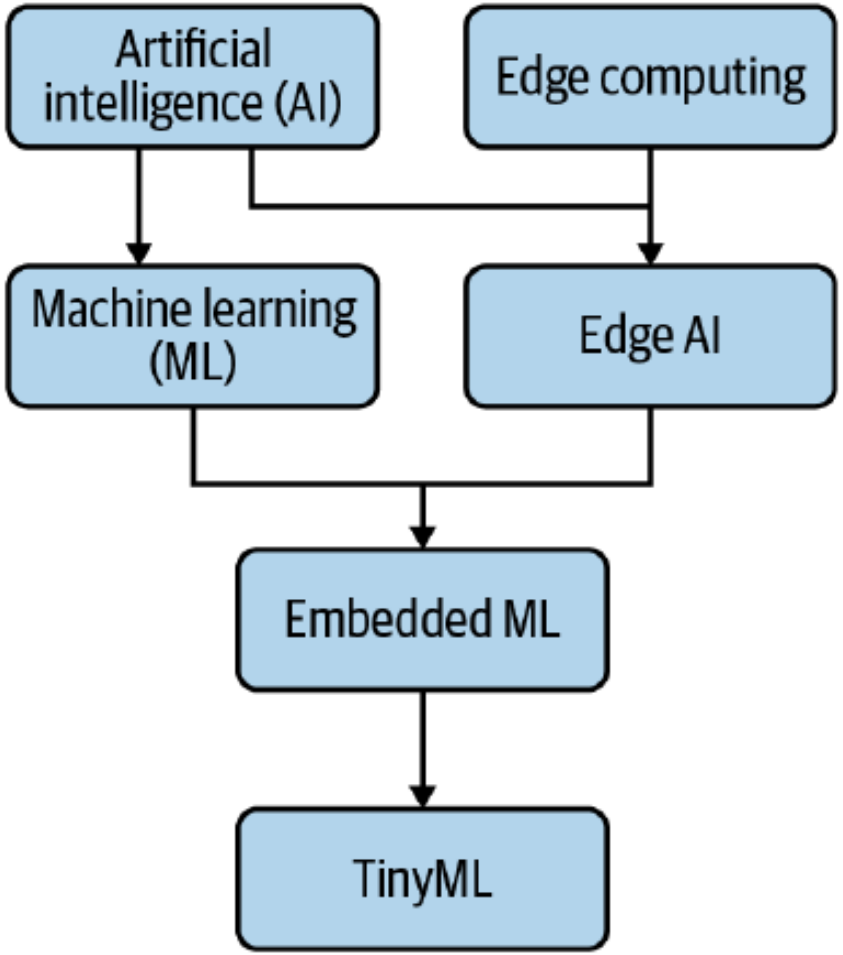

## Perché elaborare localmente?

- **Bandwidth —** meno dati trasmessi
- **Latency —** meno ritardi di comunicazione
- **Economics —** costi operativi da valutare
- **Reliability —** minore dipendenza dalla rete
- **Privacy —** dati trattati localmente

BLERP · Jeff Bier, Edge AI and Vision Alliance

# Edge AI, Embedded ML e TinyML

**DALL'INTELLIGENZA DISTRIBUITA AI DISPOSITIVI A RISORSE LIMITATE**




## Tre concetti con ambiti sovrapposti

### Edge AI

Elaborazione AI vicino alla sorgente dei dati, su dispositivi o infrastrutture edge.

### Embedded ML

Modelli ML integrati in sistemi embedded dedicati.

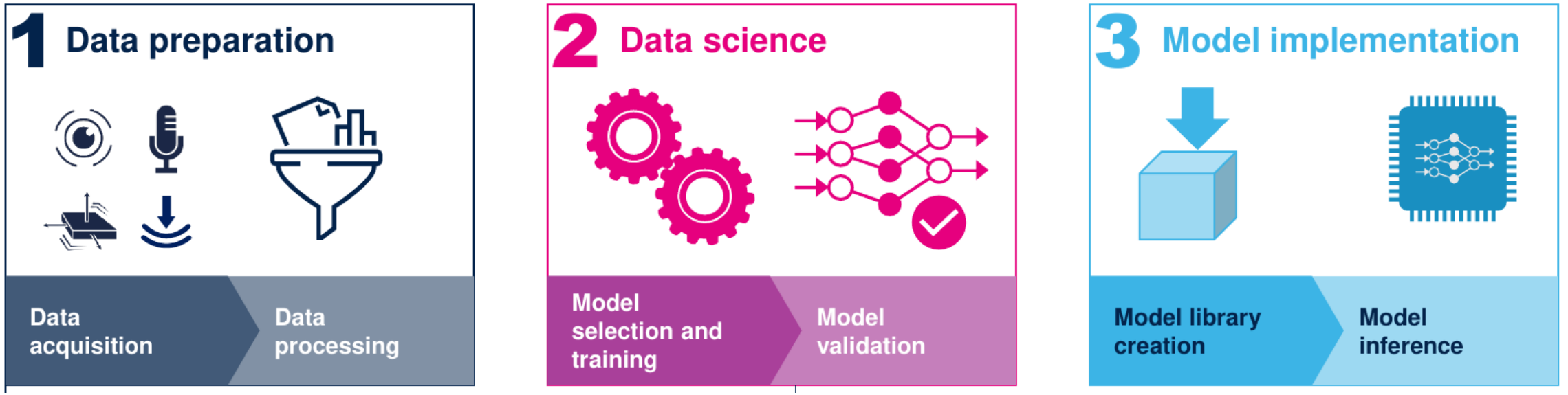
### TinyML

ML su dispositivi con forti vincoli di memoria, calcolo ed energia.

Le categorie non costituiscono una tassonomia rigida: descrivono collocazione, integrazione e vincoli diversi.

# Dal modello al dispositivo: workflow dell’Edge AI

**DALLA PREPARAZIONE DEI DATI ALL’INFERENZA EMBEDDED**




**Training e validazione del modello → conversione, integrazione e inferenza sul target embedded**

La validazione offline non sostituisce la misura di latenza, memoria e consumi sulla piattaforma target.

# Le sfide dell’Edge AI e le strategie di ottimizzazione

**DALLE LIMITAZIONI HARDWARE ALLE SOLUZIONI METODOLOGICHE**

## SFIDE TECNOLOGICHE

- **Accuratezza**
  Prestazioni e generalizzazione
- **Memoria**
  ROM del modello e RAM
- **Calcolo e latenza**
  Capacità e requisiti real-time
- **Energia**
  Inferenza, acquisizione, trasmissione
- **Robustezza**
  Rumore, variabilità e domain shift

## STRATEGIE DI OTTIMIZZAZIONE

- **Architetture efficienti**
  Progettazione hardware-aware
- **Decomposizione tensoriale**
  CP e Tucker, reti low-rank
- **Pruning**
  Rimozione di componenti ridondanti
- **Quantizzazione**
  Riduzione della precisione numerica
- **Fine-tuning e verifica**
  Recupero prestazioni e test su target

**Obiettivo: bilanciare accuratezza, risorse disponibili e vincoli applicativi.**

# Agricoltura: riconoscere le malattie della vite

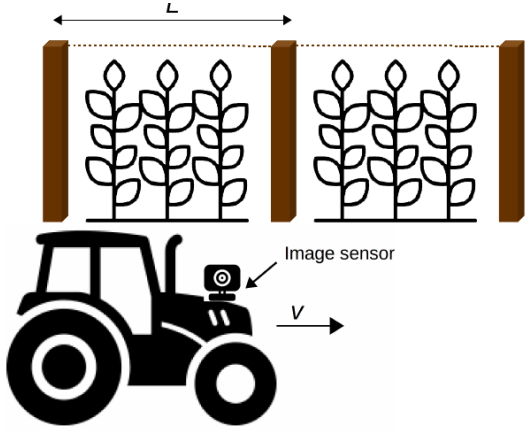
**CLASSIFICAZIONE CON VISIONE ARTIFICIALE EMBEDDED**

**Foglia sana**

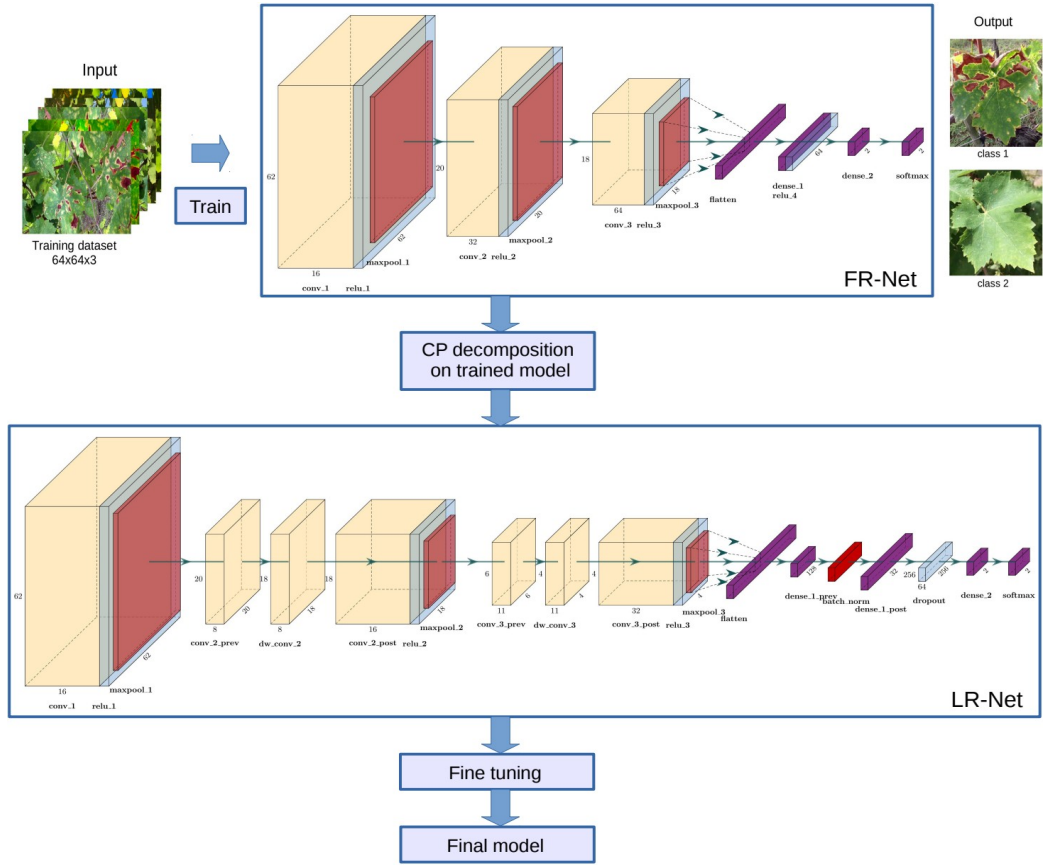


**Foglia con Esca**








Riconoscimento dei sintomi di Esca in immagini fogliari.

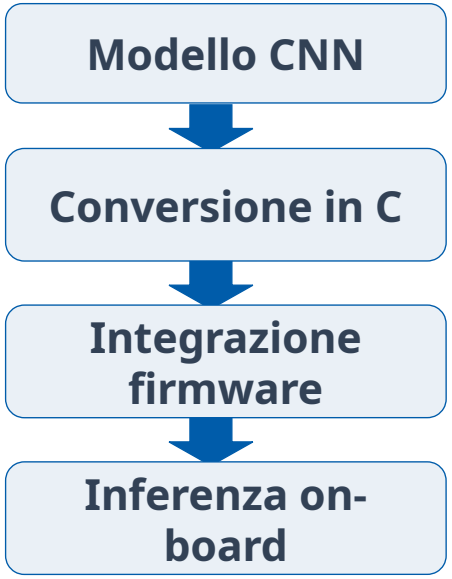
- Compressione della CNN tramite decomposizione CP.
- Deployment su OpenMV Cam basata su STM32.

Falaschetti et al., IEEE JETCAS, 2021 · Collaborazione scientifica: STMicroelectronics

# Agricoltura: compressione e deployment

FR-NET E LR-NET SULLA STESSA APPLICAZIONE

## Deployment su OpenMV Cam




**Implementazione completa su sistema embedded**

| Indicatore | FR-Net | LR-Net |
|---|---|---|
| Parametri | 40162 | **12204** |
| Memoria modello | 48.0 KB | **25.1 KB** |
| Accuratezza test | **99.0%** | 98.4% |

**Inferenza su OpenMV Cam STM32H7 Plus: LR-Net 15.57 FPS (64.21 ms/immagine).**

**Il modello presenta una notevole riduzione in termini di numero di parametri e occupazione di memoria, con un lieve calo nell'accuratezza rispetto al modello di partenza.**

Falaschetti et al., IEEE JETCAS, 2021 · Collaborazione scientifica: STMicroelectronics

# Verso la potatura automatizzata della vite

## SEGMENTAZIONE DELLE STRUTTURE DELLA PIANTA



Esempio di immagine reale del vigneto

### Potatura manuale
La potatura della vite è un'attività ad alta intensità di lavoro che richiede operatori esperti.

### Complessità visiva
Le immagini del vigneto presentano sfondi complessi, occlusioni, germogli sovrapposti e morfologie variabili.

### Obiettivo dell'automazione
Identificare e segmentare automaticamente le strutture rilevanti della pianta in immagini reali.

Manoni, Falaschetti et al., IEEE COINS 2026

# Potatura automatizzata della vite: risultati visivi

**RISULTATI QUALITATIVI DELLA SEGMENTAZIONE**

Confronto qualitativo tra ground truth (giallo) e predizioni del modello (rosso).



Esempio 1



Esempio 2

**Giallo: Ground truth Rosso: Model prediction**

Buona corrispondenza visiva tra annotazioni e predizioni nei due esempi.

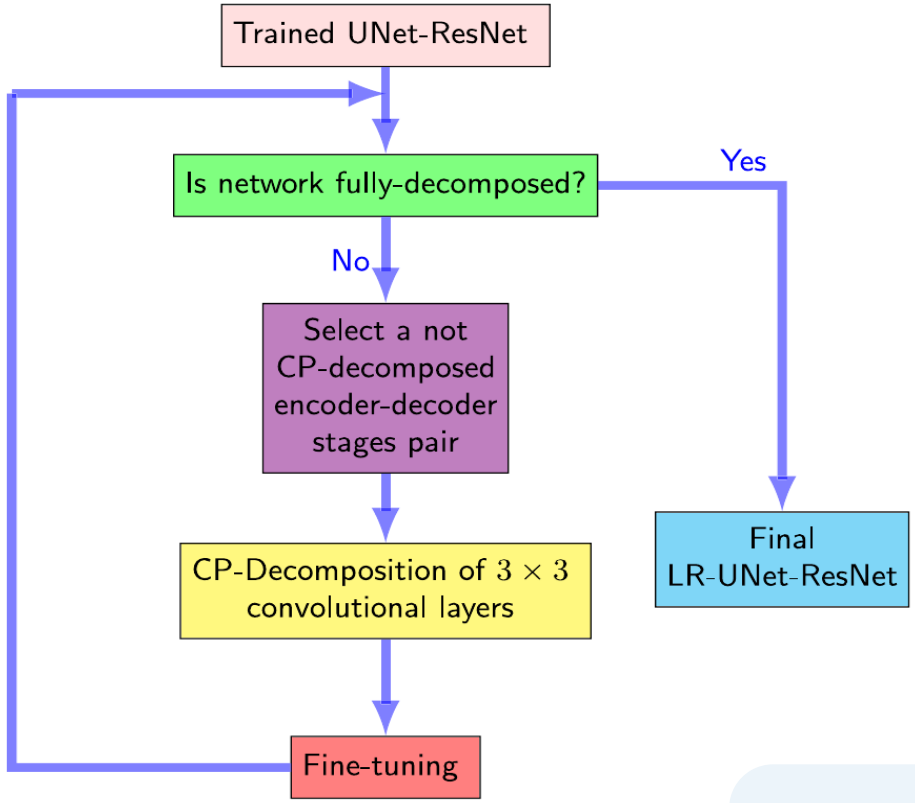
**Sviluppi futuri: ottimizzazione della rete per il deployment su sistema embedded.**

Manoni, Falaschetti et al., IEEE COINS 2026

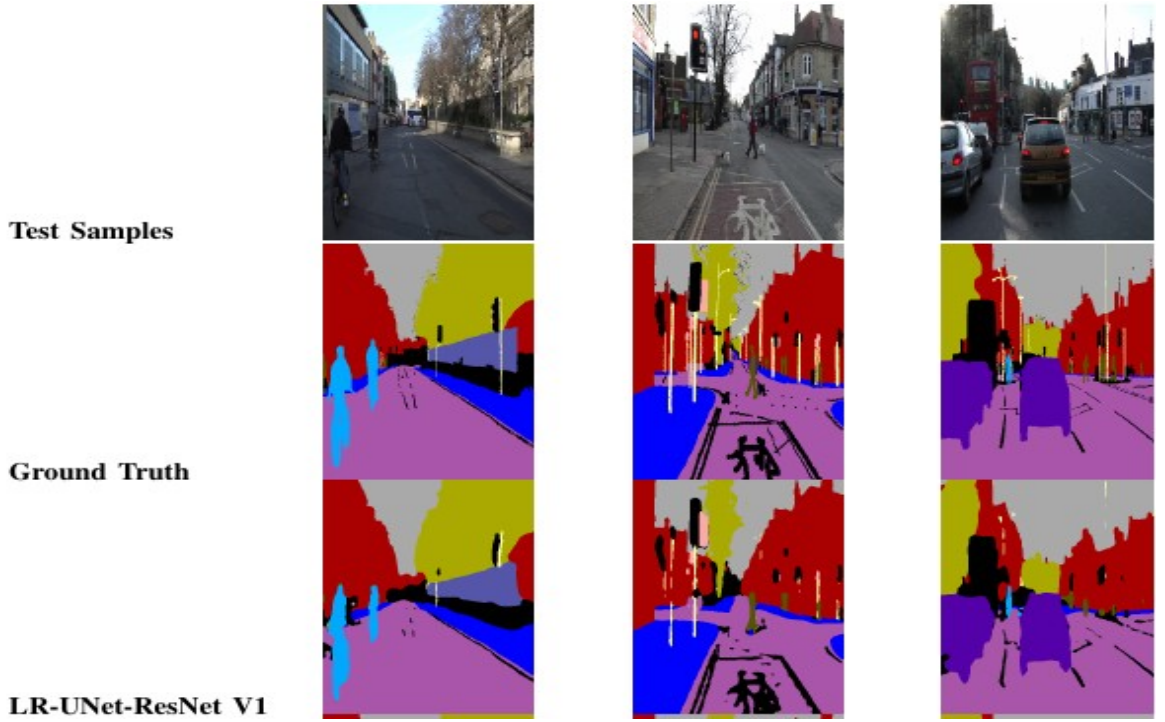
# Guida assistita: segmentazione semantica

**ARCHITETTURE LOW-RANK PER SCENE STRADALI**

## Compressione iterativa CP + fine-tuning




## Segmentazione semantica: risultati qualitativi




### LR-UNet-ResNet

Tre configurazioni low-rank (V1–V3). Valutazione su Raspberry Pi 4 e Jetson Nano

### Risultati: LR-UNet-ResNet V1

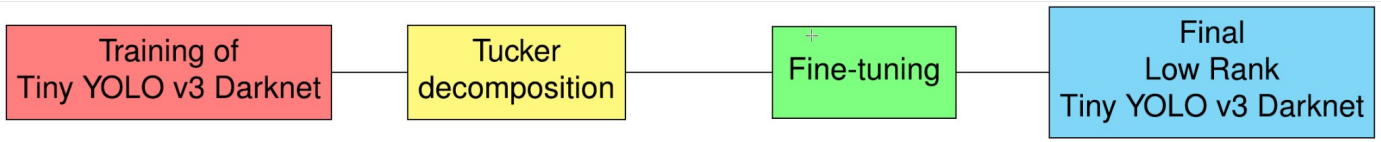
| | |
|---|---|
| Parametri | **130531** |
| Accuratezza test | **85.46%** |
| Inferenza CPU | **0.091 s** |

Falaschetti et al., IEEE Open Journal of Circuits and Systems, 2022

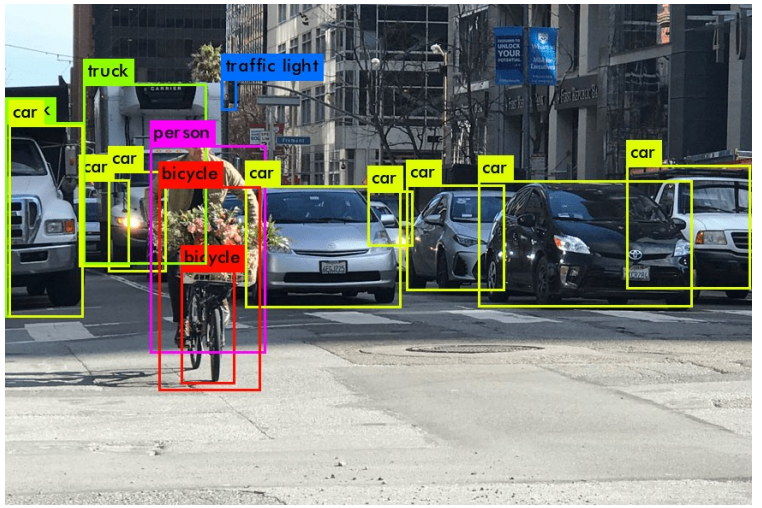
# Guida assistita: object detection

**VEICOLI E PEDONI IN TEMPO REALE**

## Compressione del detector




## Rilevamento di oggetti in scena stradale



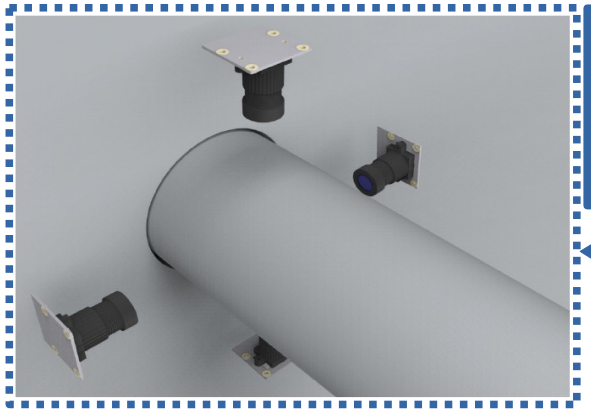

## Confronto su KITTI: Tiny YOLO v3 vs LR Tiny YOLO v3

| Indicatore | Tiny YOLO v3 | LR Tiny YOLO v3 |
|---|---|---|
| Parametri | 8.685 M | **2.815 M** |
| mAP@0.5 | **62.20%** | 60.02% |
| FPS — Raspberry Pi 4 + Coral TPU | 2.40 | **2.74** |
| FPS — Jetson Nano 2 GB | 27.53 | **34.72** |

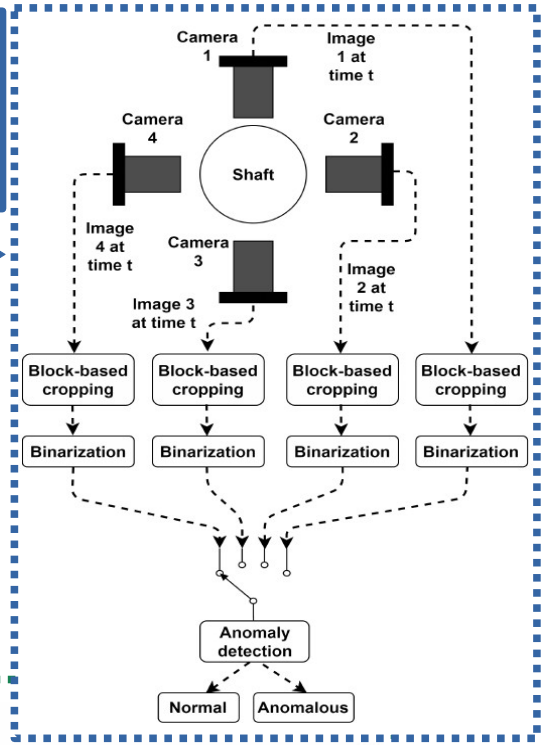
Falaschetti et al., IEEE Transactions on Intelligent Transportation Systems, 2024

# Industria: manutenzione predittiva

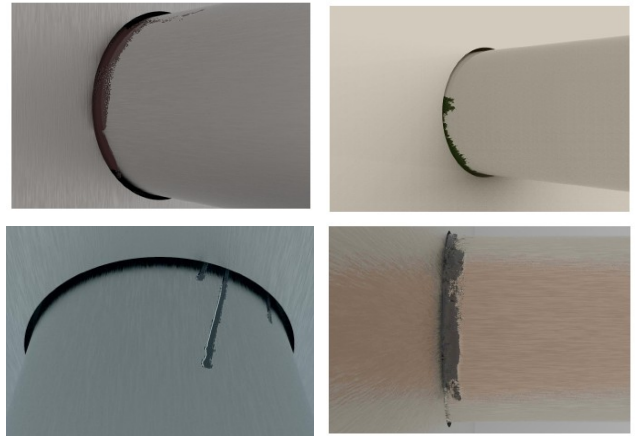
**ANOMALY DETECTION IN TURBINE EOLICHE**

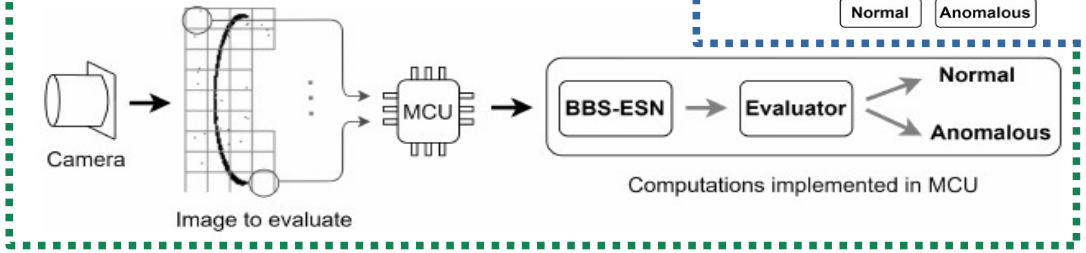


Installation with 4 cameras framing the junction where anomalies can happen.
(a) The cameras framing the junction. (b) Block diagram of the cameras framing the junction, being synchronized in time.




Block diagram of the system architecture: from the image acquisition to their classification as normal or anomalous. The outputs of the neural network are evaluated with a decision process, called Evaluator, which works on the reconstruction error between the BBS-ESN input blocks and output blocks, mediated over all the blocks in an image. Specifically, the error metric is the number of pixels that are reproduced wrongly.






## BBS-ESN

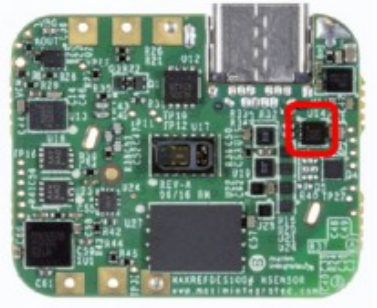
Rilevazione di perdite d'olio mediante immagini.

- Elaborazione a blocchi, rappresentazioni binarie e apprendimento online.
- Progettazione orientata a MCU.

Cardoni et al., Electronics, 2021 · Collaborazione scientifica: STMicroelectronics

# Salute: riconoscimento dell’attività umana

**RNN E SENSING INDOSSABILE SU STM32**



PPG and accelerometer signals were recorded from the wrist during some activities (stepper, squat, rest), using the Maxim Integrated MAXREFDES100 sensor while the network was incorporated into the Cloud Jam L4 board, based on a STM32 MCU.





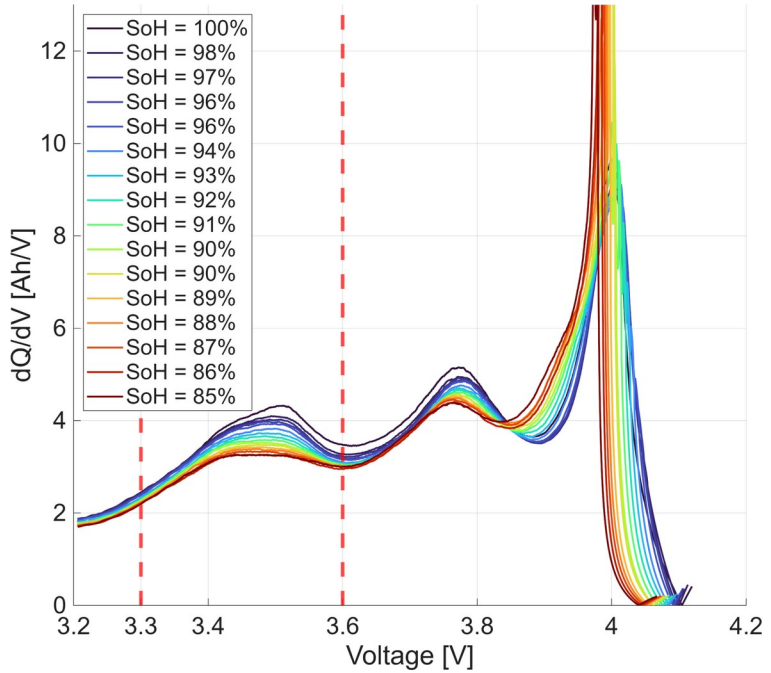

## Sensori + inferenza locale

- PPG e accelerometro triassiale acquisiti al polso.
- Rete ricorrente implementata su piattaforma Cloud-JAM L4 (STM32).
- Studio con accuratezza riportata >95%.

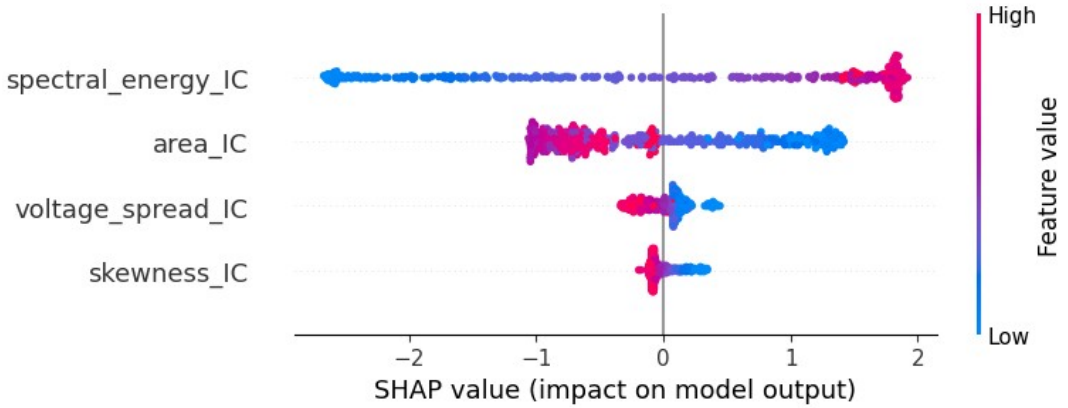
Alessandrini et al., Electronics, 2021

# Batterie: meno indicatori per stimare lo SoH

## DALLO STUDIO DEGLI INDICATORI ALLA RIDUZIONE DEGLI INPUT PER IL BMS




**SIE/Springer · Quali indicatori descrivono il degrado?**




**ApplePies · Quali feature servono davvero al GPR?**

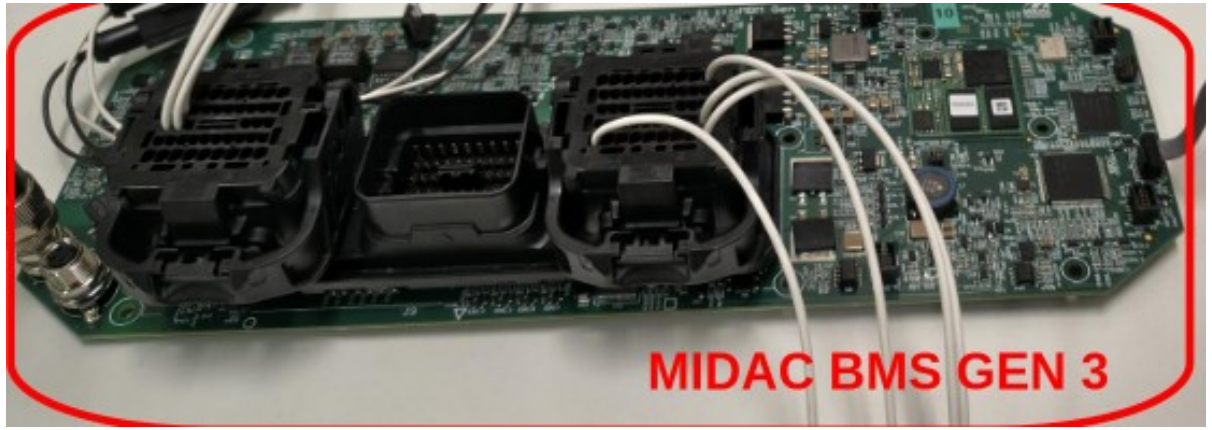
**17 feature → 4 feature | RMSE real-driving: 1.128 → 0.941 pp**

**Obiettivo embedded: minimizzare gli indicatori da comunicare o elaborare nel BMS, preservando la qualità della stima.**

Rapaccini et al., Springer LNEE, SIE 2026 (under review)
Rapaccini et al., Springer LNEE, ApplePies 2026 Proc. (accepted, in press) · Collaborazione: Infineon Technologies

# Batterie: dalla selezione delle feature al BMS

## DUE POSSIBILI SCELTE ARCHITETTURALI PER UN'INTEGRAZIONE EDGE AI




**BMS embedded · collaborazione industriale: MIDAC**

## Dove calcolare gli indicatori?

### A · Elaborazione esterna

Feature calcolate esternamente e comunicate al microcontrollore del BMS.

### B · Elaborazione locale

Feature estratte dalle misure disponibili e processate direttamente sulla MCU.

**Da verificare: disponibilità degli indicatori, costo di comunicazione e calcolo, memoria, latenza e accuratezza.**

Architetture possibili, non implementazione già validata del modello SoH sulla board mostrata.

Collaborazione industriale: MIDAC · sviluppo hardware e BMS (distinto dai risultati ApplePies)

# Dalle applicazioni alle sfide comuni

**CHE COSA ACCOMUNA QUESTI LAVORI?**

## Modelli più efficienti

Compressione e selezione degli indicatori

## Co-design hardware–software

Ottimizzazione congiunta di dati, modello, firmware e hardware

## Validazione hardware

Memoria, inferenza, capacità della board

## Prestazioni robuste

Dati e condizioni non presenti nel training

**L’ottimizzazione Edge AI riguarda l’intero sistema, non soltanto il modello.**

# Sfide aperte e prospettive

DALLA VALIDAZIONE ALGORITMICA AL CAMPO

## Robustezza

Domain shift, rumore, condizioni d’uso

## Energia

Consumi di sensori, elaborazione e radio

## Deployment

Modelli e pipeline su hardware reale
Misure realistiche, nel tempo e sul campo

## Agentic AI e Physical AI: prospettiva

Verso sistemi intelligenti autonomi, capaci di integrare percezione, pianificazione, decisione e azione nell'ambiente reale.

**Dall’Edge AI alla Physical AI: direzione futura.**

CONCLUSIONI

# Portare l’AI vicino ai dati è una sfida di sistema.

**Il task applicativo determina i vincoli del dispositivo.**

**L’ottimizzazione deve preservare prestazioni misurabili.**

**Il deployment e la robustezza richiedono verifiche dedicate.**

## Grazie per l’attenzione