#INNOVATION@DET

Deep learning for image processing with applications to remote sensing and industry

E. Magli, D. Valseaia

Politecnico di Torino

Contesto e competenze

Gruppo di ricerca «**Image Processing Lab**», focalizzato su deep learning con applicazioni a **spazio** e **computer vision**

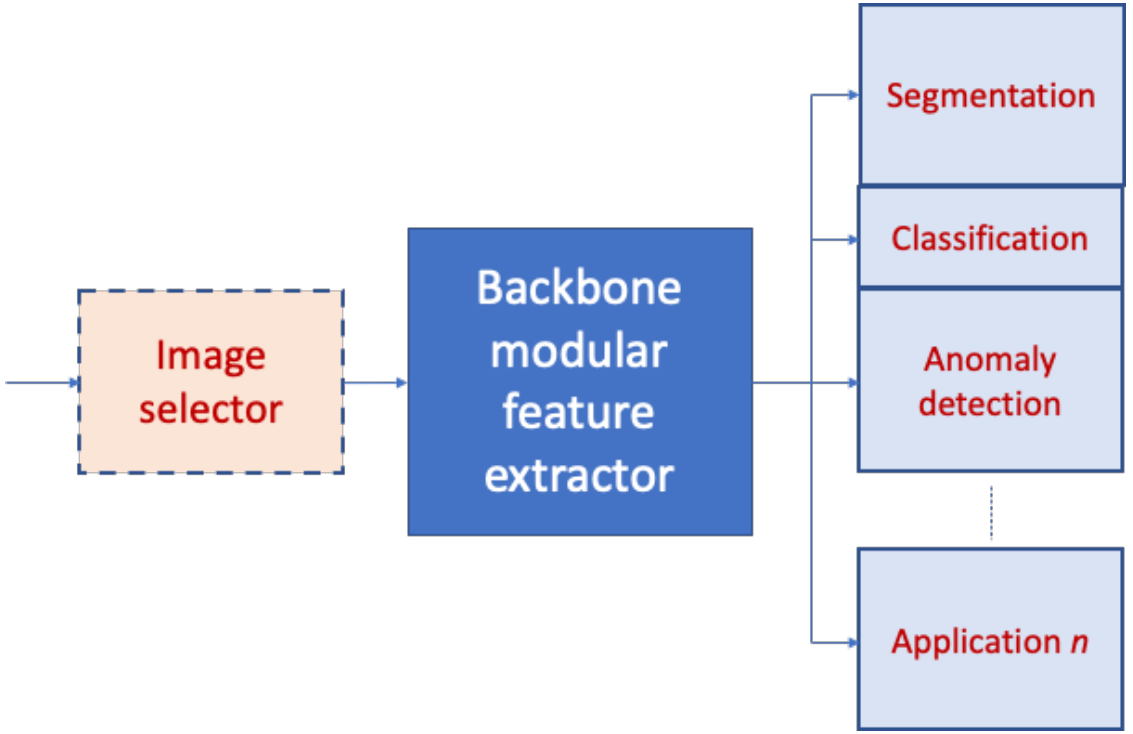
- *Reti neurali* avanzate: transformer, graph neural networks, GAN, reti ricorrenti, …
- *Training* avanzato: supervised, self-supervised, adversarial, …
- *Problemi* affrontati:
  - Elaborazione immagini: deconvoluzione, denoising, super-resolution
  - Analisi immagini: classificazione, segmentazione, object detection
  - Predizione temporale da sequenze di dati

Self-supervised learning per edge-AI

Problema: scarsità di dati etichettati, complessità

Soluzione: piattaforma con feature extractor condiviso

- Addestrato in modo auto-supervisionato
- Eseguito una sola volta
- «Teste» addestrate con pochi dati
- Adatto per piattaforme edge-AI (p.es. spazio)



Super-resolution di immagini

Obiettivo: da immagini multiple, generare una singola immagine a risoluzione più elevata

- Applicazione 1: **Spazio** (passaggi multipli sulla stessa area)
- Applicazione 2: **Burst di fotografie** su smartphone
- Applicazione 3: Lettura di **codici a barre** da immagini a bassa risoluzione

Soluzioni: Metodologie originali basate su:

- Reti neurali «permutation invariant» e «explainable»
- Modelli neurali continui di scene (NERF) per rendering

Contatti

E-mail: enrico.magli@polito.it

Web: www.ipl.polito.it

Politecnico
di Torino

#INNOVATION@DET

Machine Learning per interazioni sicure fra robot e umani in ambienti industriali

M. Indri, P.D. Cen Cheng

Politecnico di Torino

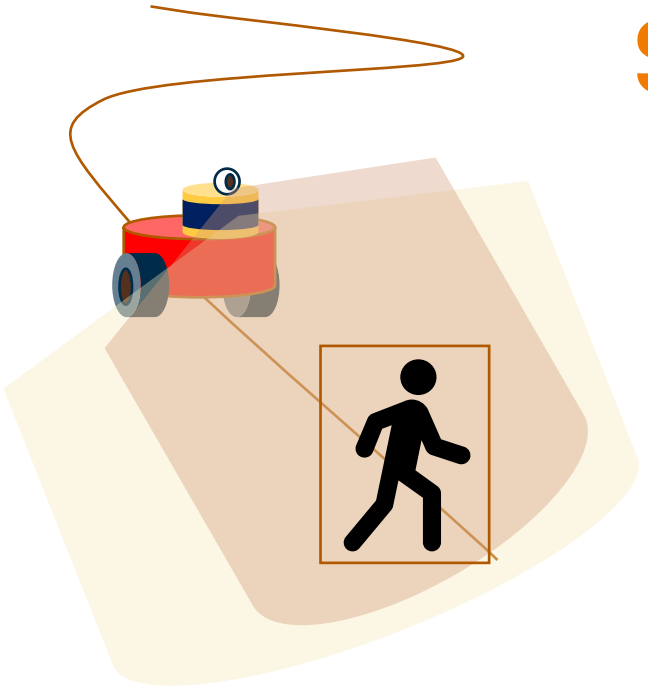
Contesto e competenze

Contesto: collaborazione fra operatori umani e robot in ambienti industriali in spazi lavorativi condivisi da umani e robot (o manipolatori) mobili, operanti come *assistenti*

Obiettivi: garanzia della sicurezza mediante riconoscimento dell'ostacolo umano ed eventuale ripianificazione dell'agente mobile, manipolazione intelligente di oggetti per la loro presa o consegna all'operatore con modalità user-friendly

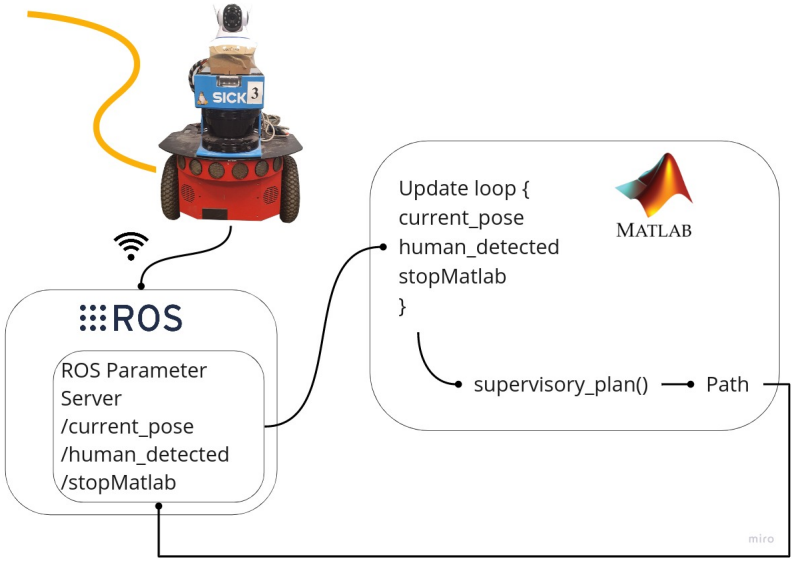
Metodologie e competenze: algoritmi di sensor fusion, tecniche di Machine Learning per la rilevazione di operatori umani, algoritmi di object detection, pianificazione del moto di manipolatori e agenti mobili

Soluzioni sviluppate / 1

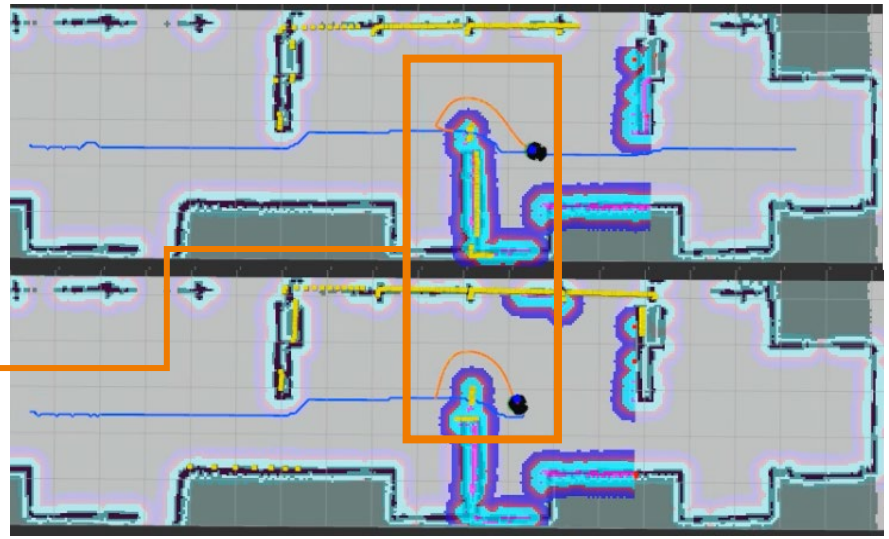


Algoritmo di sensor data fusion per sistemi con camera-LIDAR, in ambienti condivisi con persone:

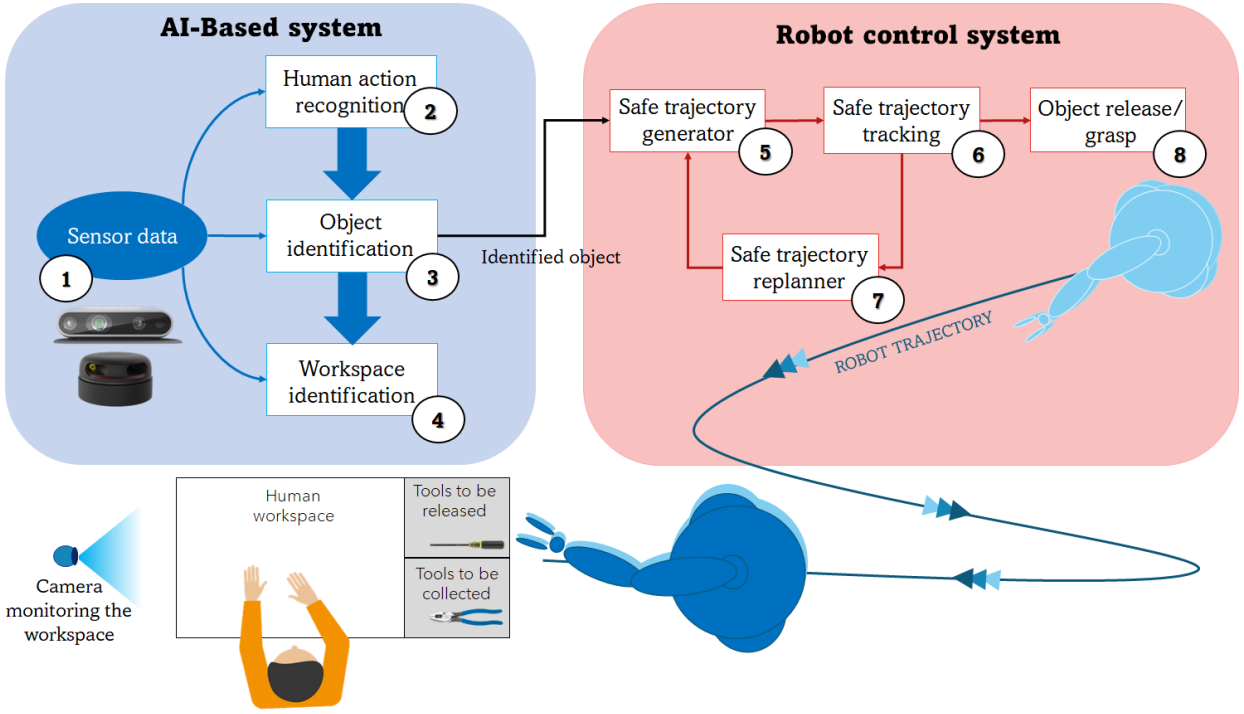
- Fusione dei dati della video camera con le misure del LIDAR
- Classificazione degli ostacoli generici e umani tramite la rete YOLO (You Only Look Once)



Ripianificazione conservativa della traiettoria dopo aver rilevato un ostacolo umano



Soluzioni sviluppate / 2



Work-in-progress: Framework per applicazioni di robot assistenti

- Utilizzo di tecniche di machine learning per migliorare le performance di un manipolatore mobile:
  - Monitoraggio dell'ambiente di lavoro del operatore umano e comprensione del contesto
  - Riconoscimento dell'operatore umano e richieste
  - Riconoscimento di tool/attrezzi da prendere, a richiesta dell'operatore umano

Contatti

Mail: marina.indri@polito.it

Web: https://staff.polito.it/marina.indri/

Politecnico
di Torino

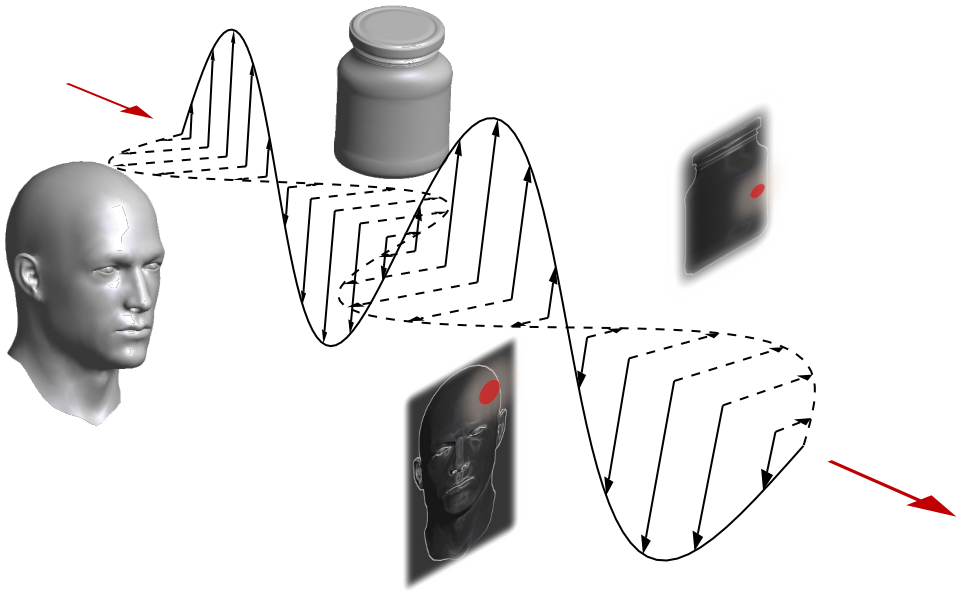
#INNOVATION@DET

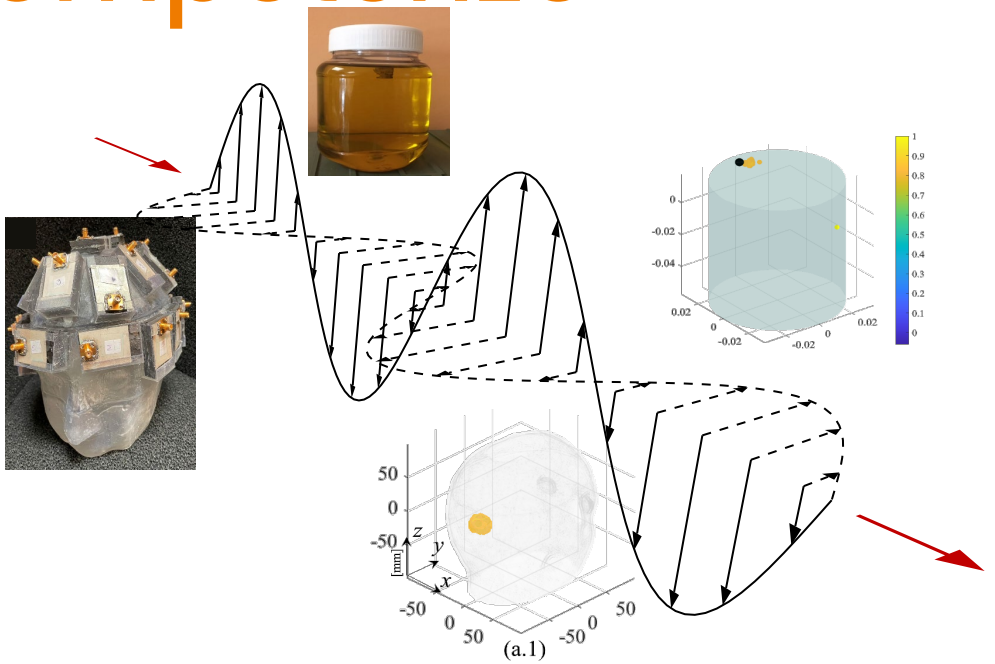
Microonde e Machine Learning per il monitoraggio di accidenti cerebrovascolari e l'ispezione di prodotti alimentari e bevande

J.A. Tobon, F. Vipiana, M. Casu, G. Turvani, M. Ricci, V. Mariano, A. Darwish, L. Cardinali

Politecnico di Torino

Contesto e competenze





IDEA

- Estrarre informazione dall'interno di un oggetto per costruire un'imagine o rilevare in corpo estraneo.
- Usare un fenomeno che possa penetrare e portare informazioni.
- Prove non destruttive e poco invadenti
- Basso costo dei dispositive a microonde

COSA ABBIAMO?

- Progettazione e sviluppo di antenne e sistemi Tx/Rx e misure a RF (0.5 – 12 GHz)
- Sviluppo e implementazione di algoritmi di imaging e di inversion (lineari e non)
- Validazione dei prototipi in laboratorio e in industria
- Accelerazione dei tempi di calcolo

Soluzioni sviluppate / 1

S^{tot}

DETECTION

TYPOLOGY

1° STEP

ML Algorithm

HEALTHY

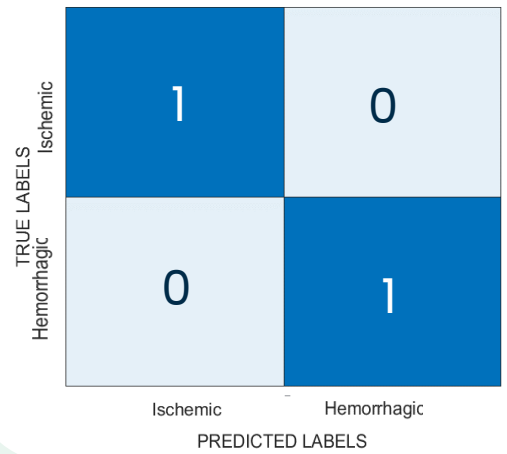
STROKE

2° STEP

ML Algorithm

ISCHEMIC

HEMORRHAGIC



Il **training** richiede un numero alto di dati, non sempre disponibili. In questo caso complementiamo il nostro data-set con delle simulazioni ottimizzate.

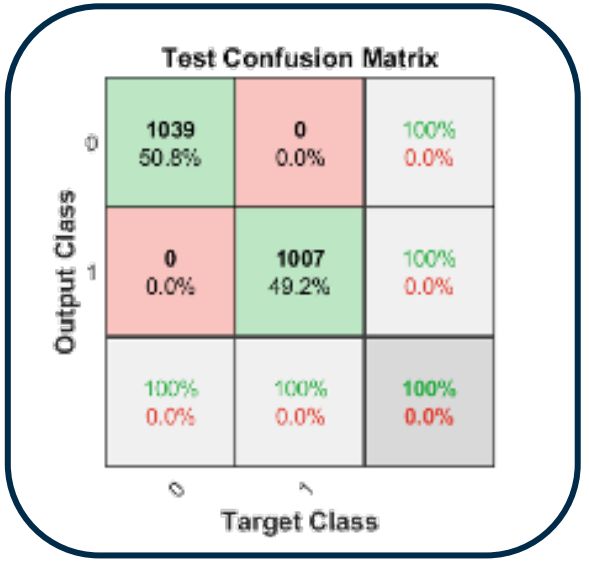
Possiamo anche definire altre classi, tipo posizione della lesione (quadranti) oppure altre tipologie.

Altre patologie possono variare le proprietà elettromagnetiche di zone più ampie, non localizzate. Una metodologia simile può essere usata per diagnosi precoce.

Soluzioni sviluppate / 2

Prodotto contaminato

Acquisizione dati in 11 punti in frequenza[9-11 GHz]
Prodotto in movimento (multivista)



Riusciamo a riconoscere gli elementi contaminati tramite Support Vector Machine (SVM) e Multi-Layer Perceptron (MLP)

Contatti

Mail: jorge.tobon@polito.it

Politecnico
di Torino

#INNOVATION @DET

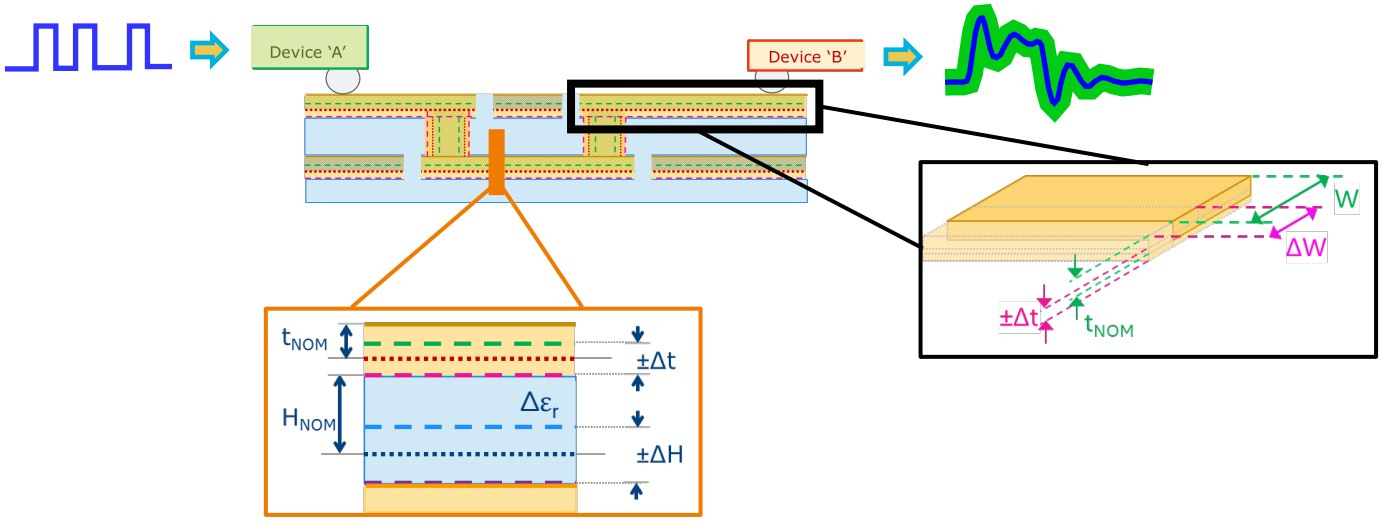
Machine Learning Based Surrogate Models for the Statical Analysis and Optimization of Electronic Circuits and Devices

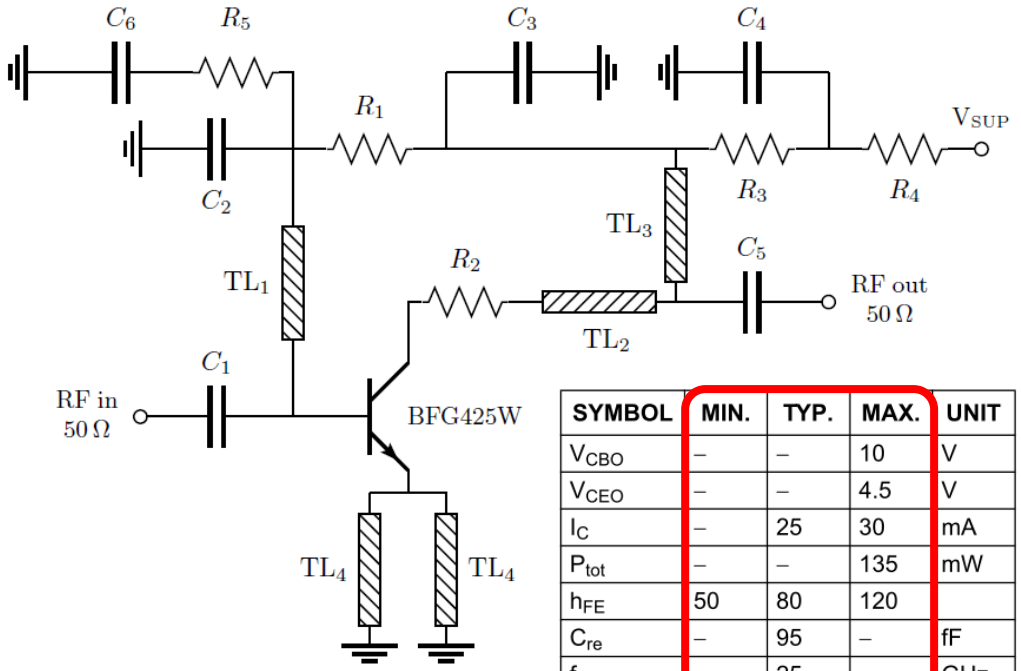
P. Manfredi, R. Trinchero

Politecnico di Torino

Contesto e motivazioni

Le variabilità di fabbricazione sono uno dei maggiori fattori limitanti nel progetto di circuiti ad elevate prestazioni



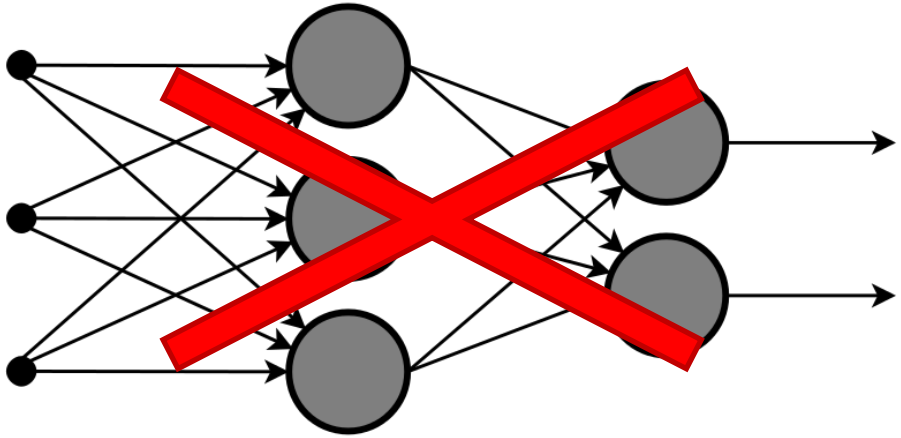


SYMBOL	MIN.	TYP.	MAX.	UNIT
V_{CBO}	–	–	10	V
V_{CEO}	–	–	4.5	V
I_C	–	25	30	mA
P_{tot}	–	–	135	mW
h_{FE}	50	80	120	
C_{re}	–	95	–	fF
f_T	–	25	–	GHz
G_{max}	–	20	–	dB
F	–	1.2	–	dB

Le simulazioni di tipo Monte Carlo richiedono migliaia di run e quindi spesso giorni o settimane

Obiettivo e metodi

Intelligenza artificiale = reti neurali?



Le reti neurali svolgono tipicamente funzioni iper-specifiche dopo un addestramento su grandi quantità di dati

Obiettivo opposto: analizzare design eterogenei utilizzando ogni volta il minor numero di dati possibile!

Soluzione: machine learning basato su kernel (LSSVM/GPR)

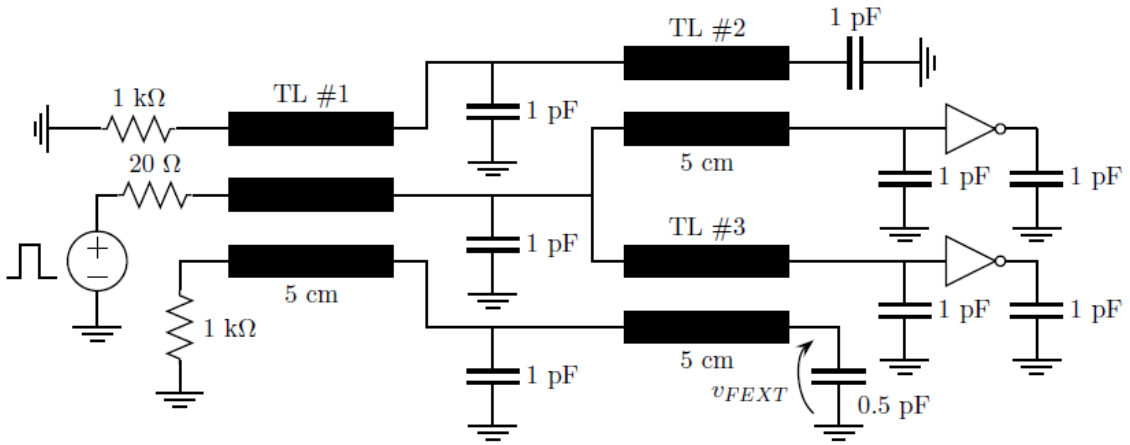
Caratteristiche

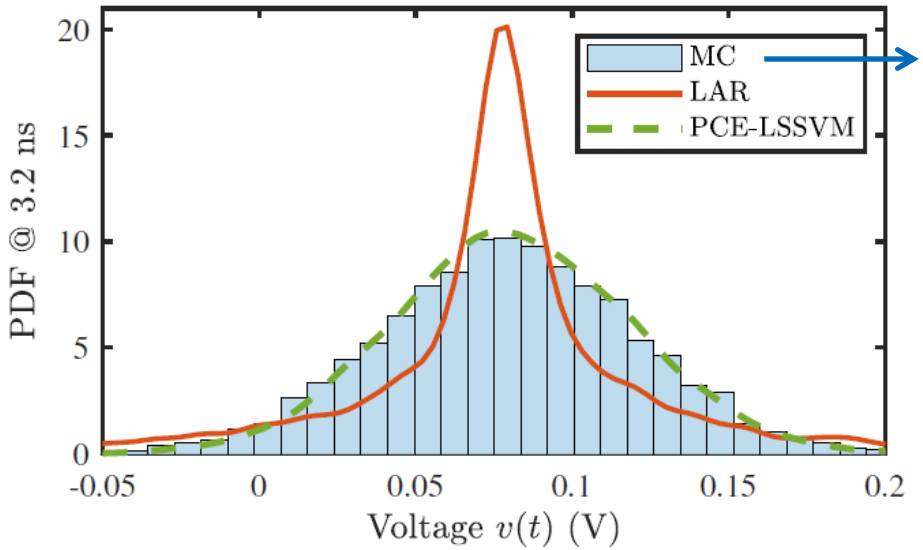
- **Flessibilità**: ottime capacità di adattamento e generalizzazione a problemi diversi
- **Model free**: non utilizzano un modello fisso e predeterminato
- **Elevata dimensionalità**: ottima scalabilità rispetto al numero di parametri indipendenti
- **Efficienza di apprendimento**: ottima accuratezza anche con un numero molto ridotto di dati di addestramento

Esempio

11 parametri incerti,
15 campioni di addestramento

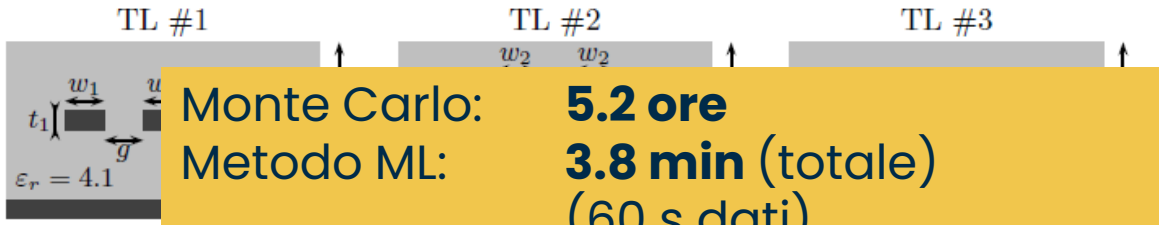
Crosstalk in interconnessione PCB



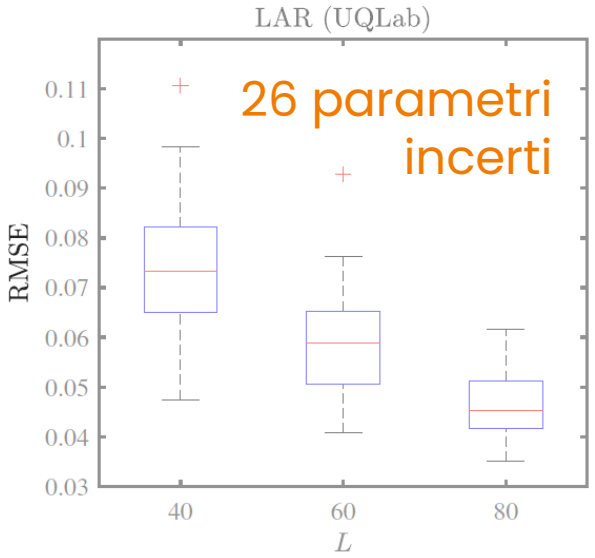


5k runs

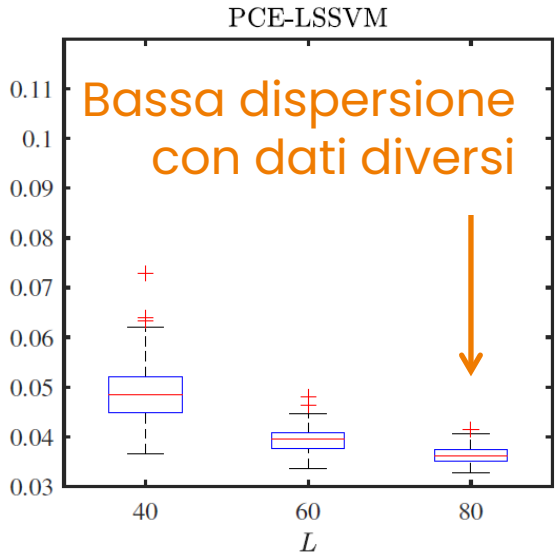
Da **11** a **26** parametri incerti



Monte Carlo: **5.2 ore**
Metodo ML: **3.8 min** (totale)
(60 s dati)
(157 s addestramento)
(9 s predizione)



26 parametri incerti



Bassa dispersione con dati diversi

Contatti

Mail: paolo.manfredi@polito.it

Web: https://emc.polito.it/

Politecnico
di Torino

#INNOVATION@DET

Detection and classification of jamming signals

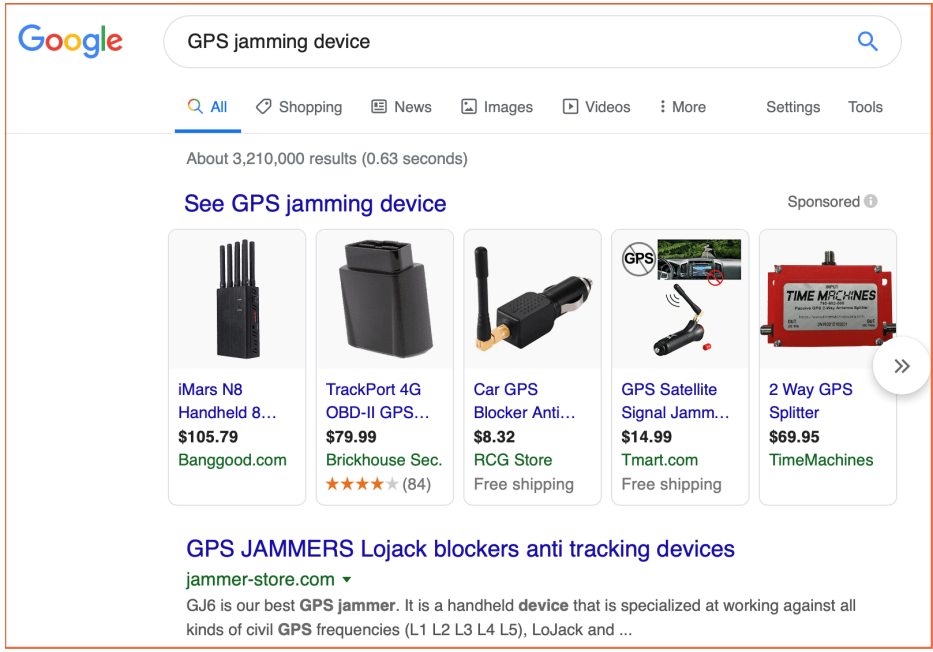
F. Dovis, A. Minetto, A. Nardin, I. E. Mehr, R. Garello, C. Chiasserini

Politecnico di Torino

Contesto e competenze

L'attività di ricerca riguarda l'uso di tecniche di machine learning per l'identificazione di **disturbi intenzionali** e non intenzionali su sistemi di **comunicazione** e di **navigazione satellitare** (GPS e Galileo)

- Fenomeni di interferenza atmosferica naturali (scintillazioni ionosferiche)
- Jamming
- Spoofing di sistemi di navigazione



Obiettivi della ricerca:

- Rilevamento automatico e classificazione
- Attivazione delle contromisure

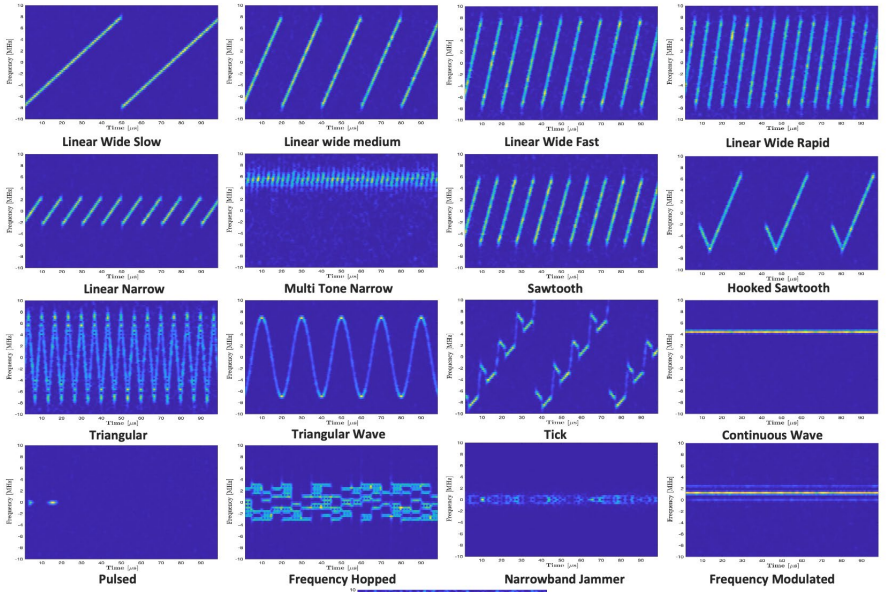
Soluzioni sviluppate / 1

Jamming detection and classification

-Soluzioni basate su supervised ML e Convolutional Neural Networks

- Test su segnali reali e in contesti applicativi

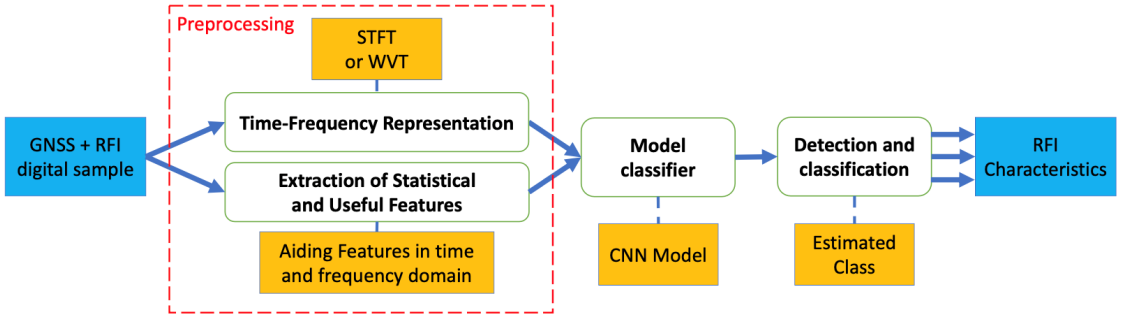
- Implementazioni per monitoraggio da stazioni terrestri e su satelliti LEO



Andamento tempo-frequenza di jammer “commerciali”

Jamming localisation

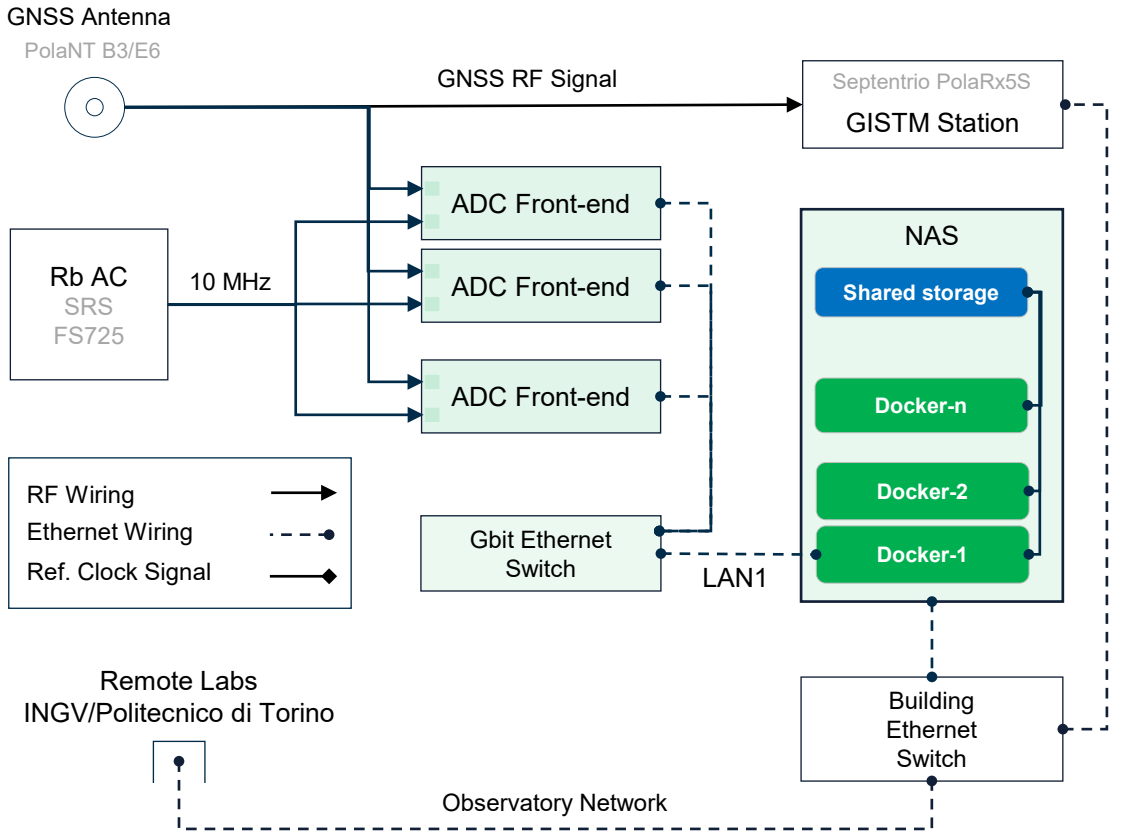
- Soluzioni basate su misure multiple raccolte da reti di ricevitori interconnessi



Soluzioni sviluppate / 2

Applicazione in stazioni di monitoraggio del segnale GNSS

- Implementazione Software Defined Radio in stazioni installate sul campo con gestione remota
- Signal acquisition and digital recording
- Real time operation
- Ottimizzazione low-complexity degli algoritmi ML e CNN implementati in architetture docker



Contatti

Mail: fabio.dovis@polito.it

Web: www.navsas.eu

Navigazione satellitare

Prof. Fabio Dovis

Dr. Alex Minetto

Dr. Andrea Nardin

Ing. Iman Ebrahimi Mehr

Sistemi di comunicazione

Prof. Roberto Garello

Prof. Carla Chiasserini

Politecnico
di Torino

#INNOVATION@DET

Reti Neurali per l'ottimizzazione di comunicazioni in fibra ottica a basso consumo ed altissimo bit rate

L. Minelli, R. Gaudino

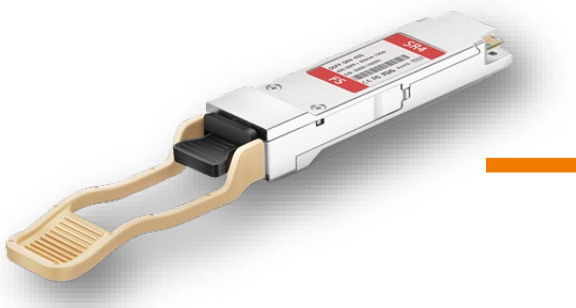
PHOTONEXT

OPTCOM

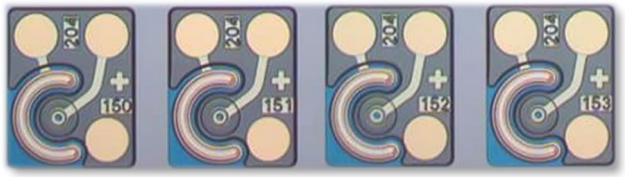
Politecnico di Torino

1859

Digital Signal Processing nelle comunicazioni ottiche a basso consumo

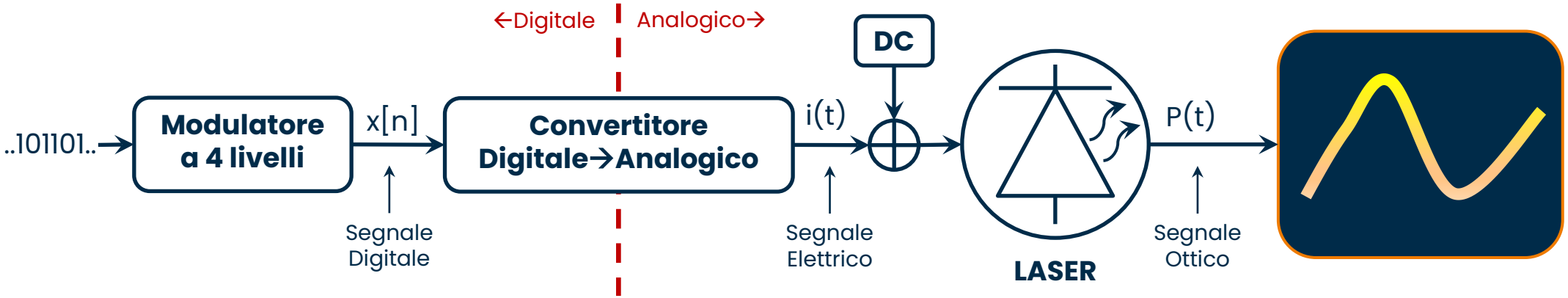


Ricetrasmettitore ottico commerciale @4x25 Gigabit/s



Array di 4 LASER VCSEL

I VCSEL sono una tipologia di laser particolarmente a basso costo rispetto ai laser tradizionali (edge-emitting)



Che cosa è la modulazione PAM-4 ?

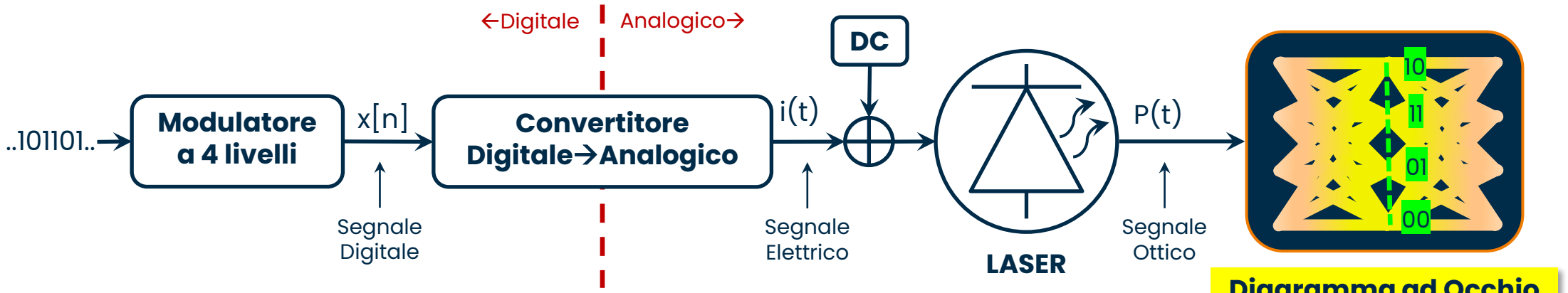
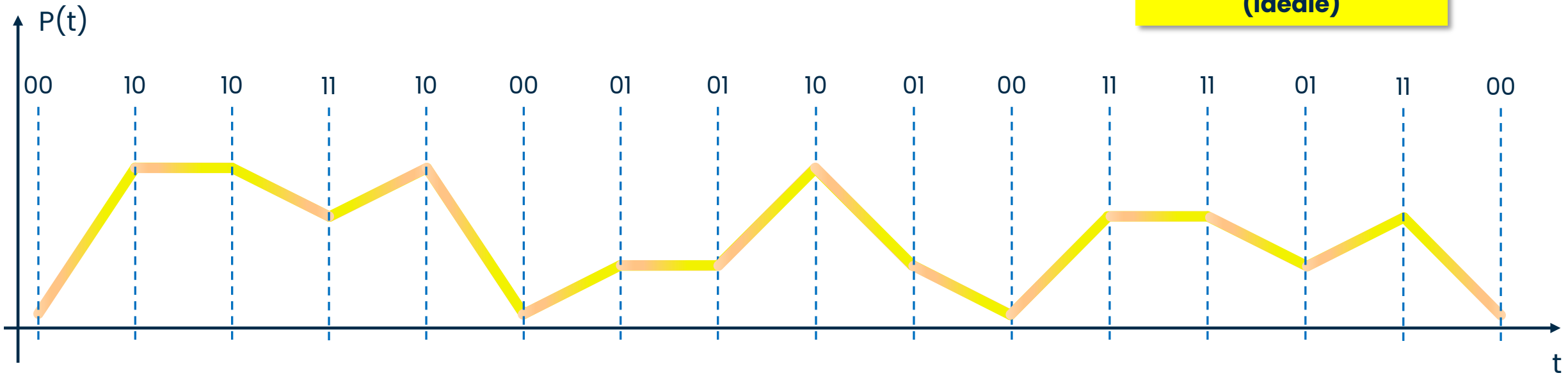
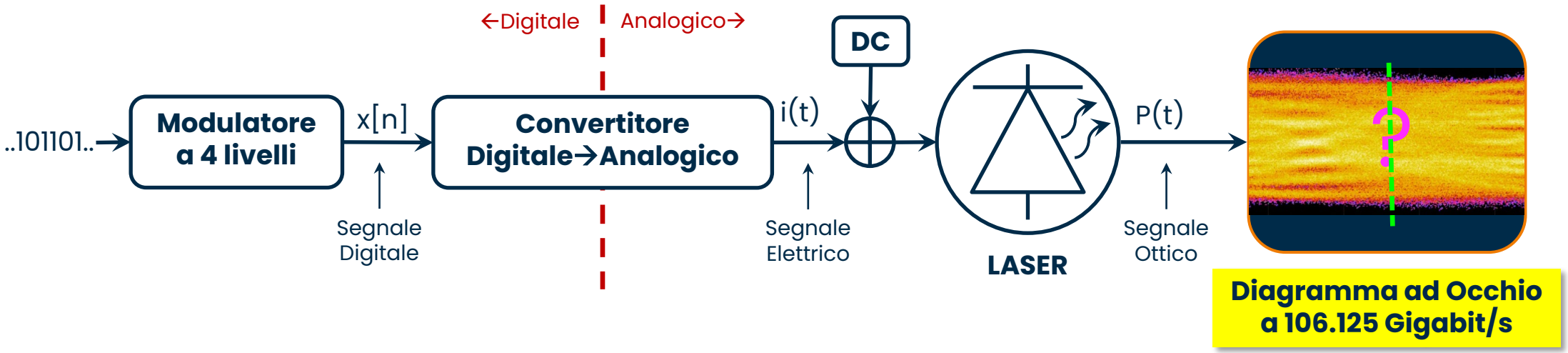


Diagramma ad Occhio (ideale)



Il problema da risolvere: distorsioni introdotte dal laser



- Ad elevatissimi Bit rate (>100 Gbps), la risposta degli attuali componenti opto-elettronici è troppo lenta

- Il Diagramma ad occhio reale è completamente chiuso

La soluzione: tecniche avanzate di pre-distorsione

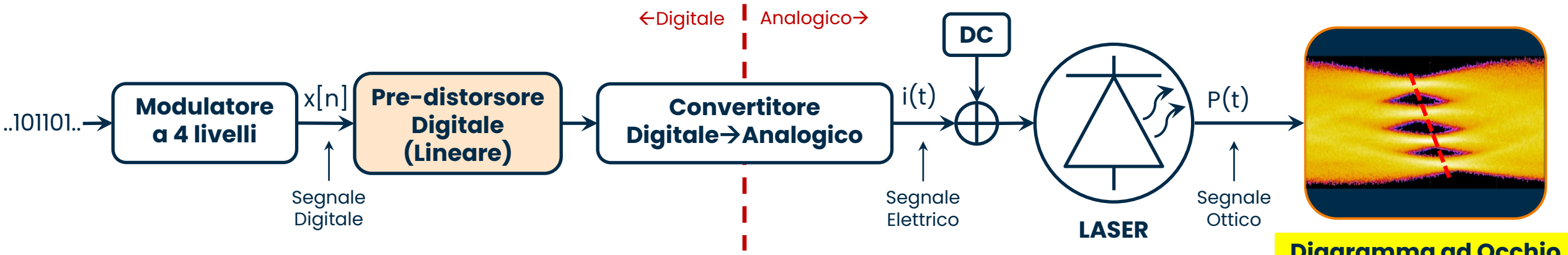
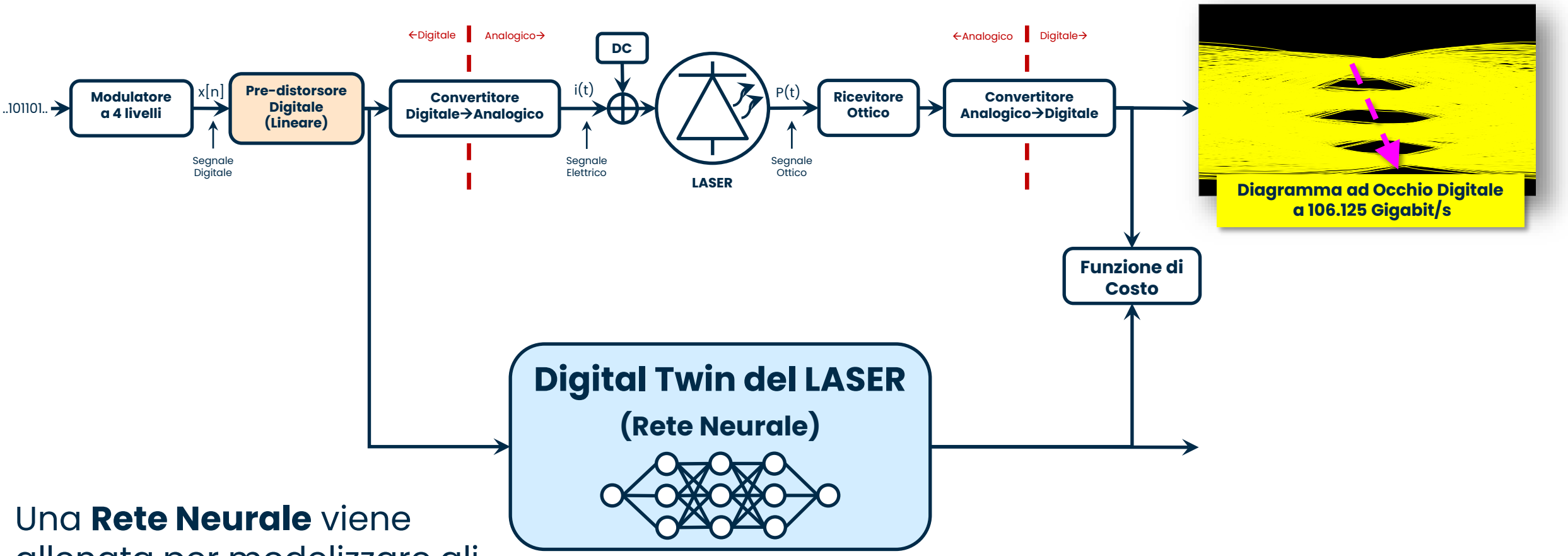


Diagramma ad Occhio a 106.125 Gigabit/s Predistorto (linearmente)

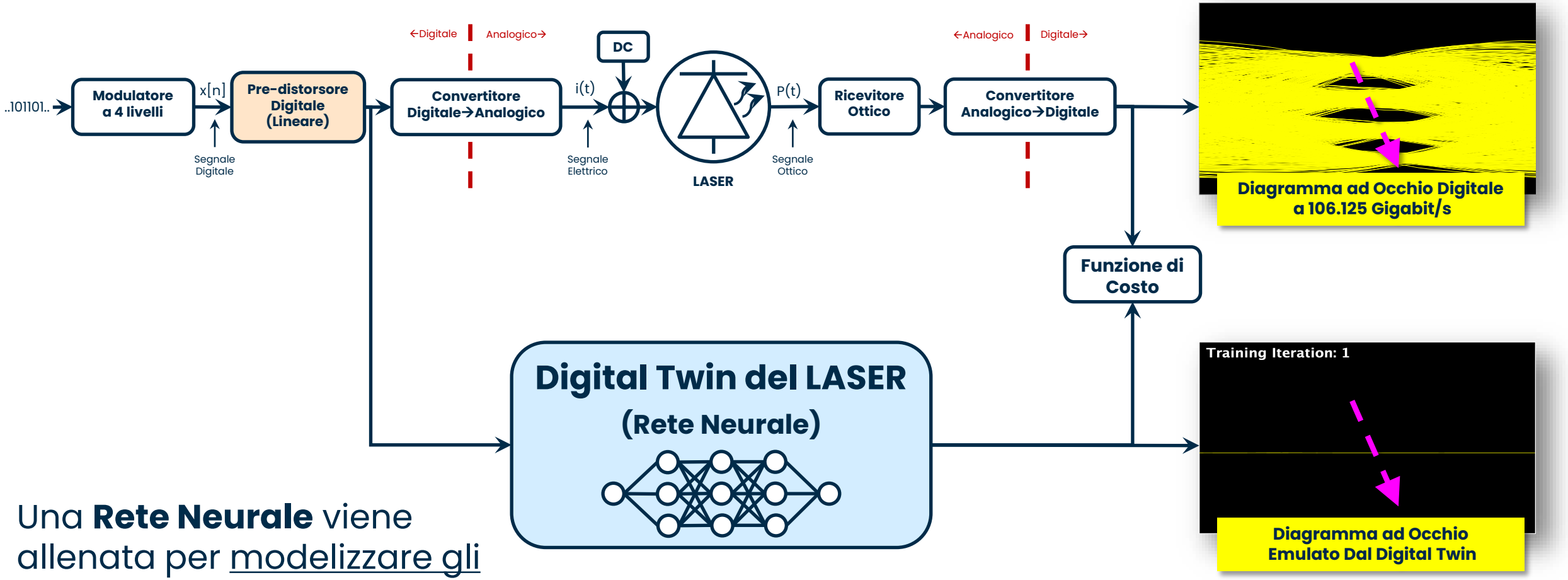
- Tramite la predistorsione digitale convenzionale, è possibile pre-compensare le limitatezze in banda del segnale…
- … <u>ma non gli effetti nonlineari causati dal LASER</u>

Step #1: Modelizzazione del LASER tramite Rete Neurale «Digital Twin»



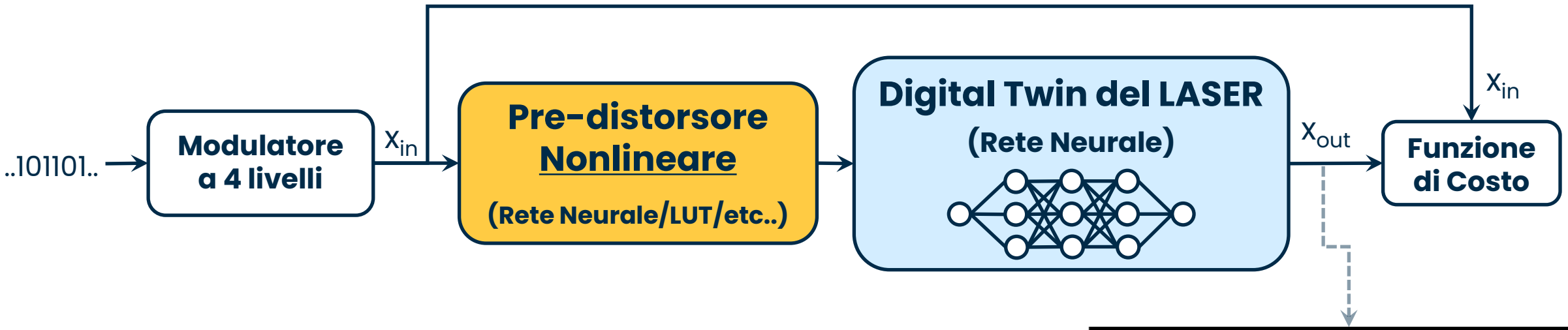
Una **Rete Neurale** viene allenata per modelizzare gli effetti nonlineari del LASER

Step #1: Modelizzazione del LASER tramite Rete Neurale «Digital Twin»

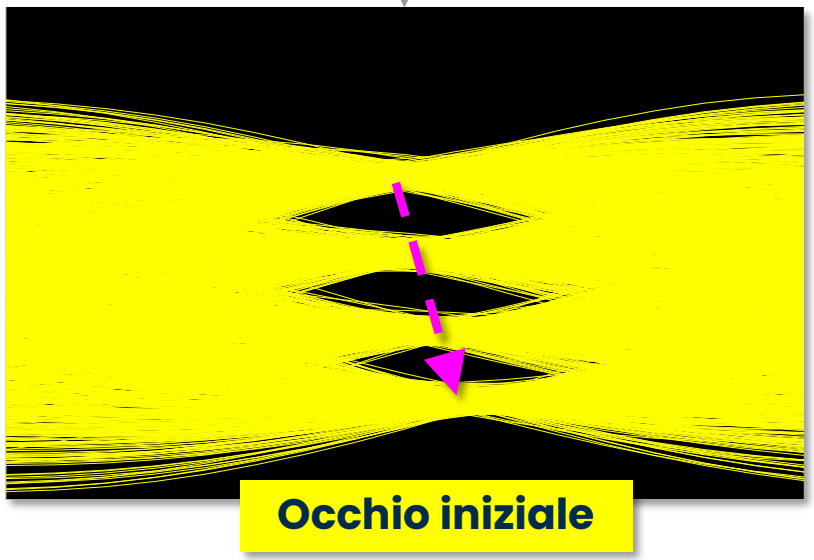


Una **Rete Neurale** viene allenata per modelizzare gli effetti nonlineari del LASER

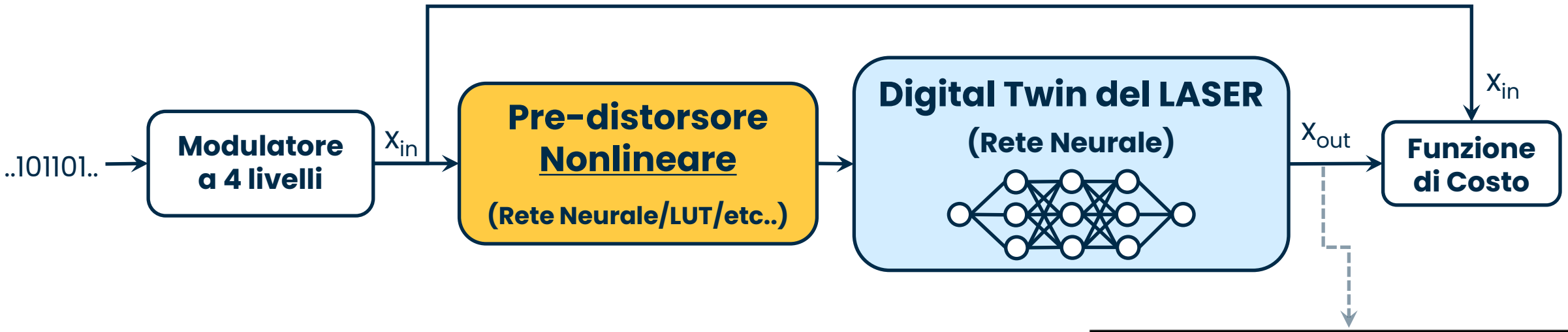
Step #2: Precompensazione del LASER tramite utilizzo del Digital Twin



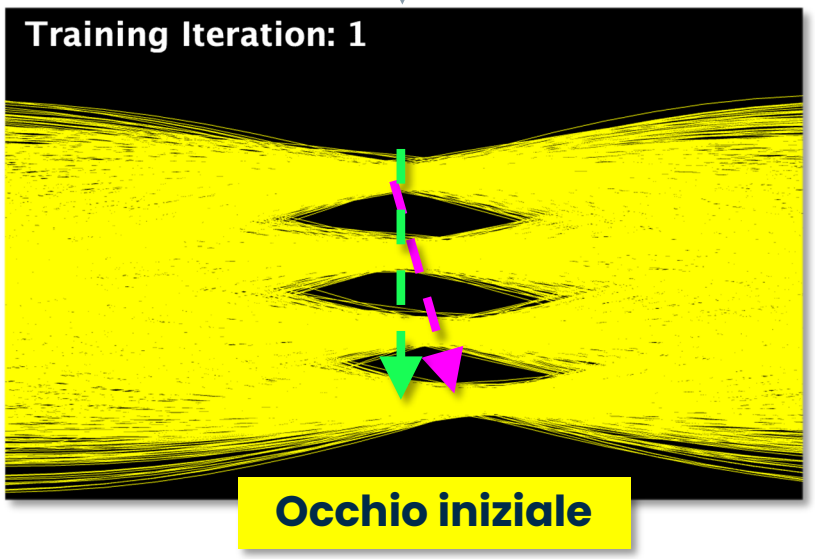
- Il Digital Twin e il Pre-distorsore Nonlineare vengono messi in cascata, formando una singola Rete Neurale



Step #2: Precompensazione del LASER tramite utilizzo del Digital Twin



- Il Digital Twin e il Pre-distorsore Nonlineare vengono messi in cascata, formando una singola Rete Neurale
- Durante l'allenamento, il Pre-distorsore impara a precompensare le nonlinearità del LASER



Acknowledgments:

This work was carried out under a research contract with Cisco Photonics.

We also acknowledge the PhotoNext Center at Politecnico di Torino, Cisco Optical GmbH and Links Foundation

Politecnico di Torino

CISCO

OPTCOM

FONDAZIONE links PASSION FOR INNOVATION

Contatti

Email:

leonardo.minelli@polito.it

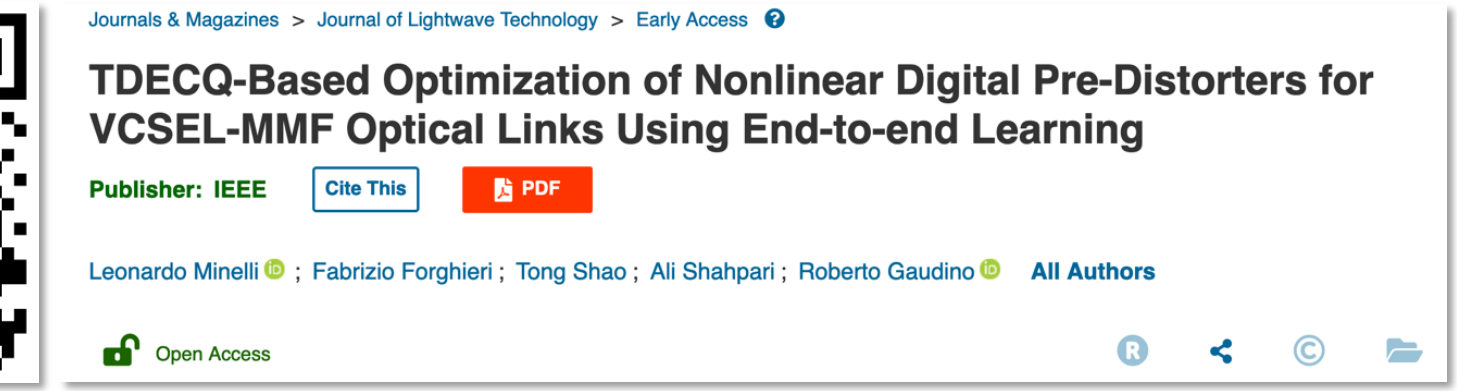
Web:

Per ulteriori informazioni, potete contattarci direttamente a questi link:

www.optcom.polito.it

www.photonext.polito.it

Per dettagli tecnici sul lavoro di ricerca presentato oggi, si consiglia il seguente articolo:



#INNOVATION@DET

ML assisted simulation and design of photonic devices and systems

P. Bardella, A. Carena, V. Curri

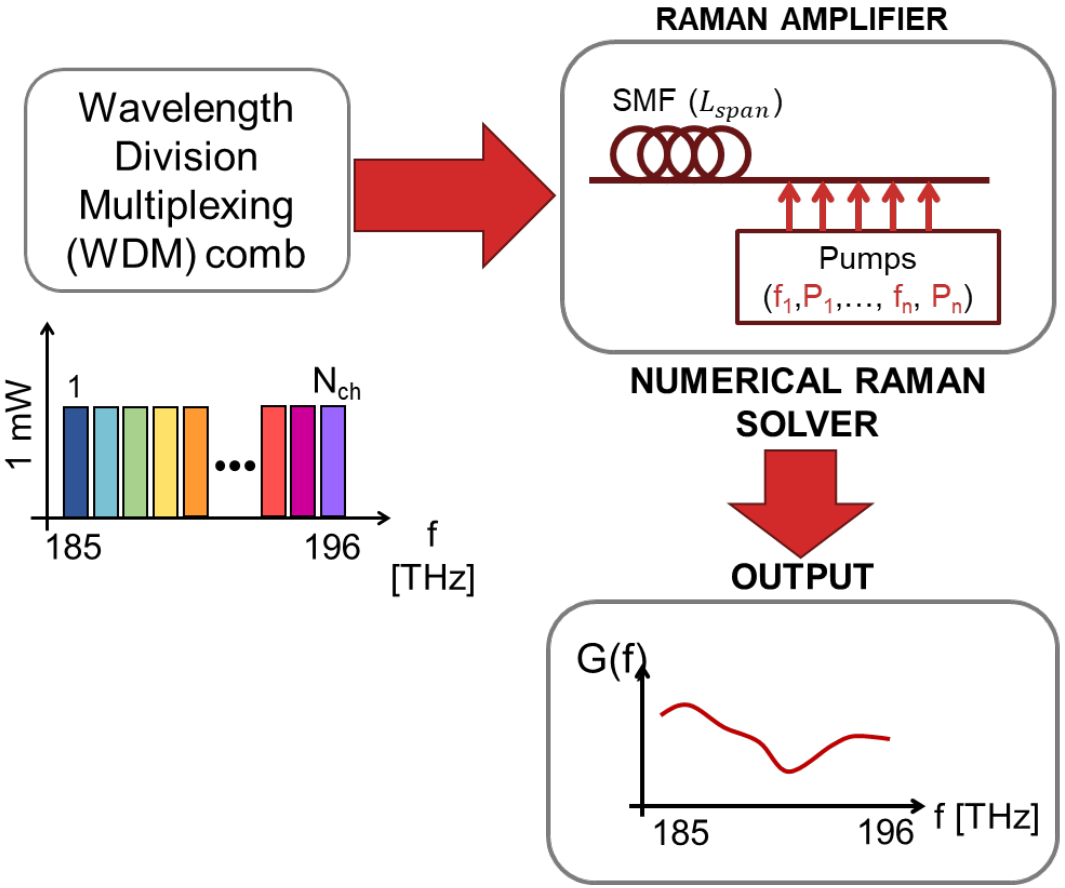
Politecnico di Torino

ML applications to Photonics

- Raman amplifiers design
- Spectral-Spatial power profile prediction in presence of Raman scattering
- Assessment of QoT impairments due to PICs
- Autonomous control of photonic switches for SDN networks
- Parameter extraction for VCSEL modeling

Raman amplifiers design

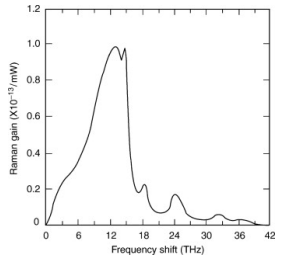
NN can be applied to both direct and inverse problem

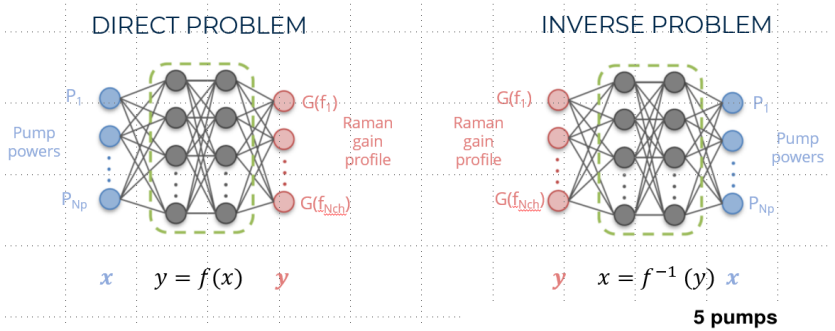


$$\frac{dP_s}{dz} = -\alpha_s P_s + C_R(\lambda_s,\lambda_p)\left[P_p^+ + P_p^-\right] P_s \qquad (1)$$

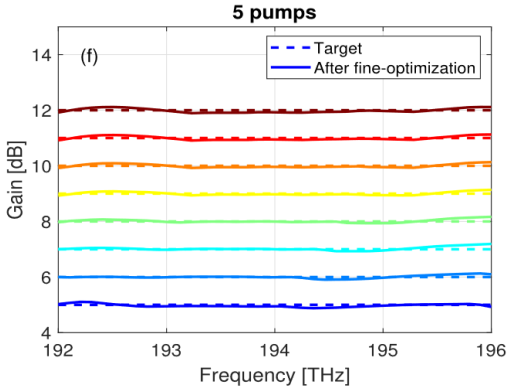
$$\pm\frac{dP_p^{\pm}}{dz} = -\alpha_p P_p^{\pm} - \left(\frac{\lambda_s}{\lambda_p}\right) C_R(\lambda_s,\lambda_p)\, P_s\, P_p^{\pm} \qquad (2)$$

$$\pm\frac{dP_A^{\pm}}{dz} = -\alpha_A P_A^{\pm} + C_R(\lambda_A,\lambda_p) P_p P_A^{\pm} + C_R(\lambda_A,\lambda_p)[1+\eta(T)] h\nu_A B_{ref} P_p \qquad (3)$$



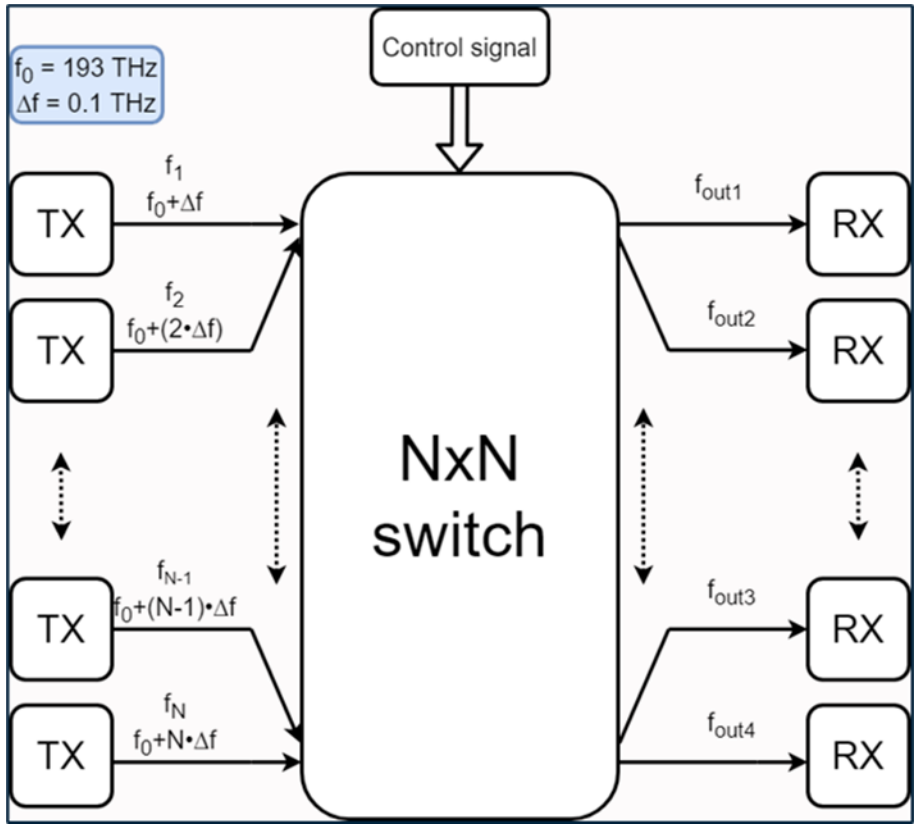


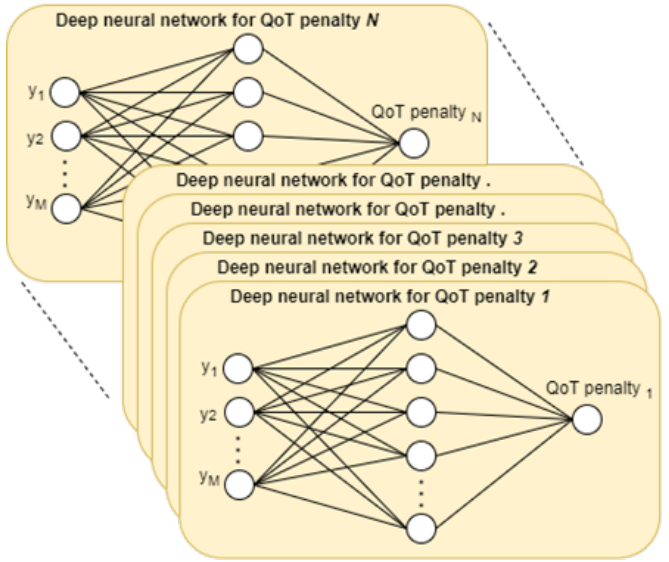
MULTI-PUMP DEFINITION
FLAT GAIN DESIGN:
RESULTS



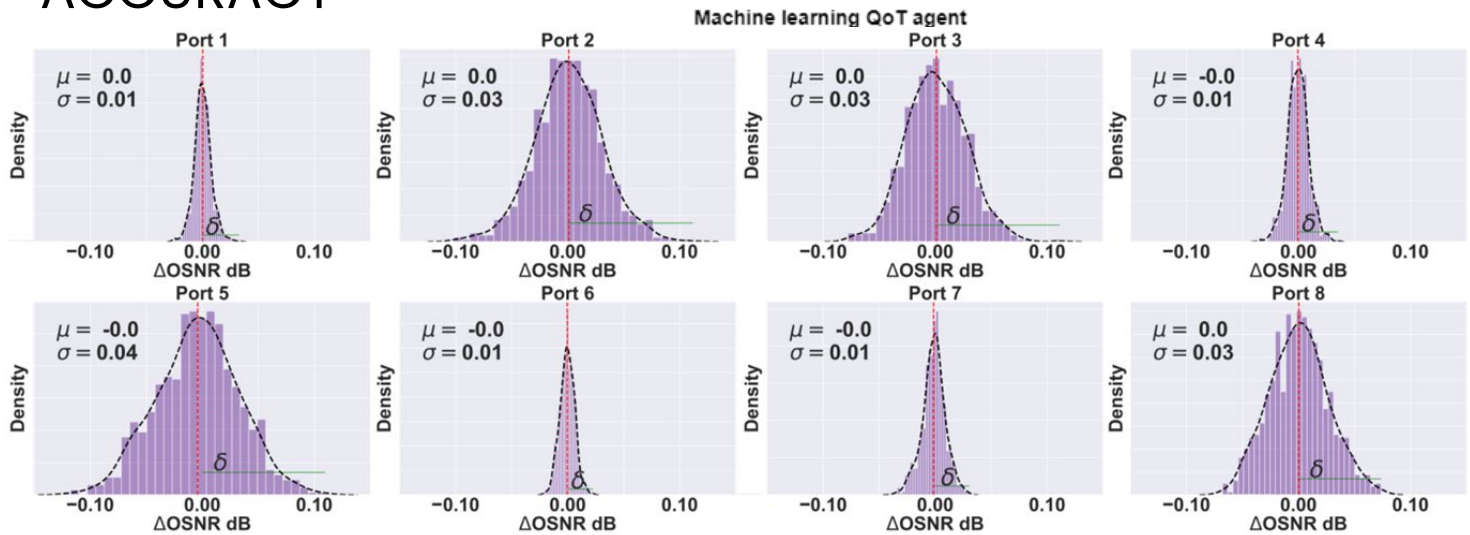
Assessment of QoT impairments due to PICs

400ZR transmission and switching





QoT PREDICTION ACCURACY



Parameter extraction for VCSEL modeling

From measurements to rate equations: a two steps approach

$$\frac{\partial N_0}{\partial t} = \frac{\eta_i I}{q} - \frac{N_0}{\tau_n} - \frac{G\left[\gamma_{00}(N_0 - N_t) - \gamma_{01} N_1\right]}{1 + \varepsilon S} S - \frac{I_l}{q}$$

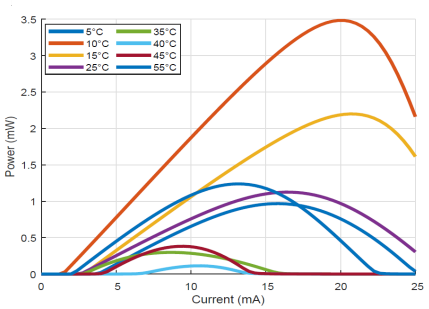
$$\frac{\partial N_1}{\partial t} = -\frac{N_1}{\tau_n}(1 + h_{diff}) + \frac{G\left[\phi_{100}(N_0 - N_t) - \phi_{101} N_1\right]}{1 + \varepsilon S}$$

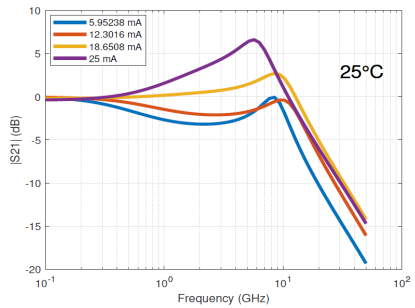
$$\frac{\partial S}{\partial t} = -\frac{S}{\tau_p} + \frac{\beta N_0}{\tau_n} + \frac{G\left[\gamma_{00}(N_0 - N_t) - \gamma_{01} N_1\right]}{1 + \varepsilon S}$$

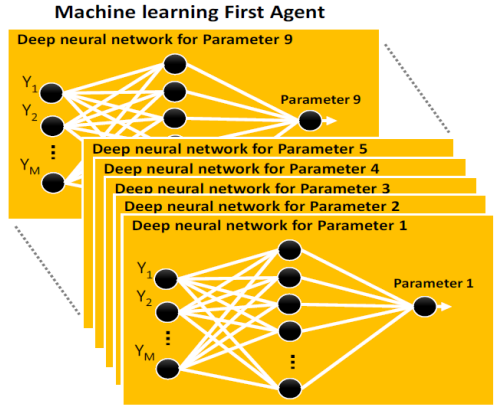
$$\frac{\partial \phi}{\partial t} = \frac{\alpha}{2} \frac{G\left[\gamma_{00}(N_0 - N_{tr}) - \gamma_{01} N_1\right]}{1 + \varepsilon S}$$

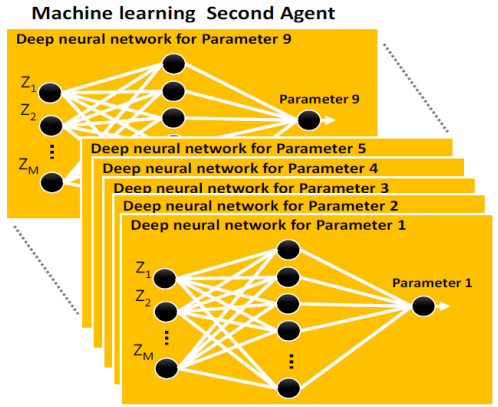
$$G(T) = G_0 \left(a_{g0} + a_{g1} T + a_{g2} T^2\right) / \left(b_{g0} + b_{g1} T + b_{g2} T^2\right)$$

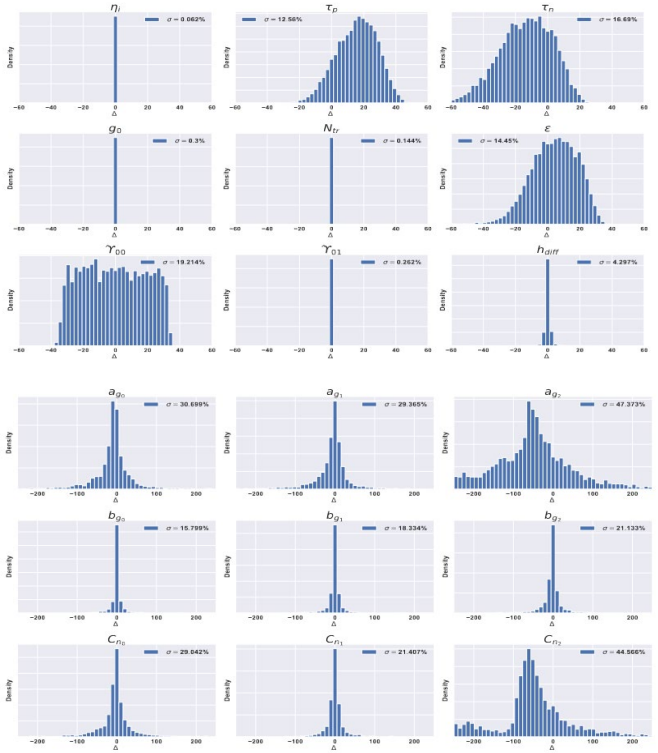
$$N_t(T) = N_{tr} \left(c_{n0} + c_{n1} T + c_{n2} T^2\right)$$











Contatti

Mail: paolo.bardella@polito.it

Web: OPTCOM group

MOG group

Politecnico
di Torino

#INNOVATION@DET

Support and design of ML tasks at the edge of the network

C. Puligheddu, C.F. Chiasserini

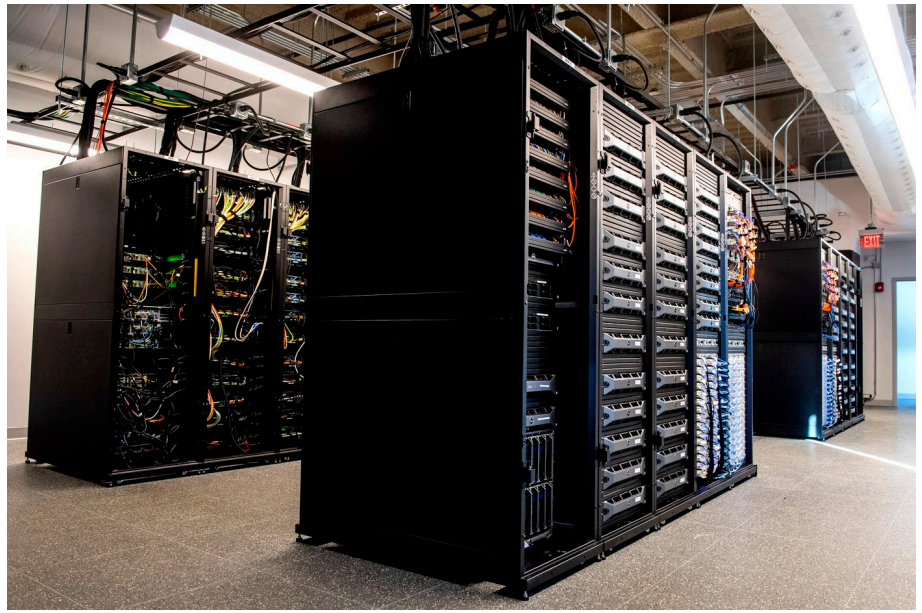
Politecnico di Torino

Machine learning & edge computing for mobile autonomous systems

Investigated topics:

- 5G and next-gen mobile networks
- Distributed computing
- Federated learning
- Task offloading
- Split computing

Development and validation on experimental equipment

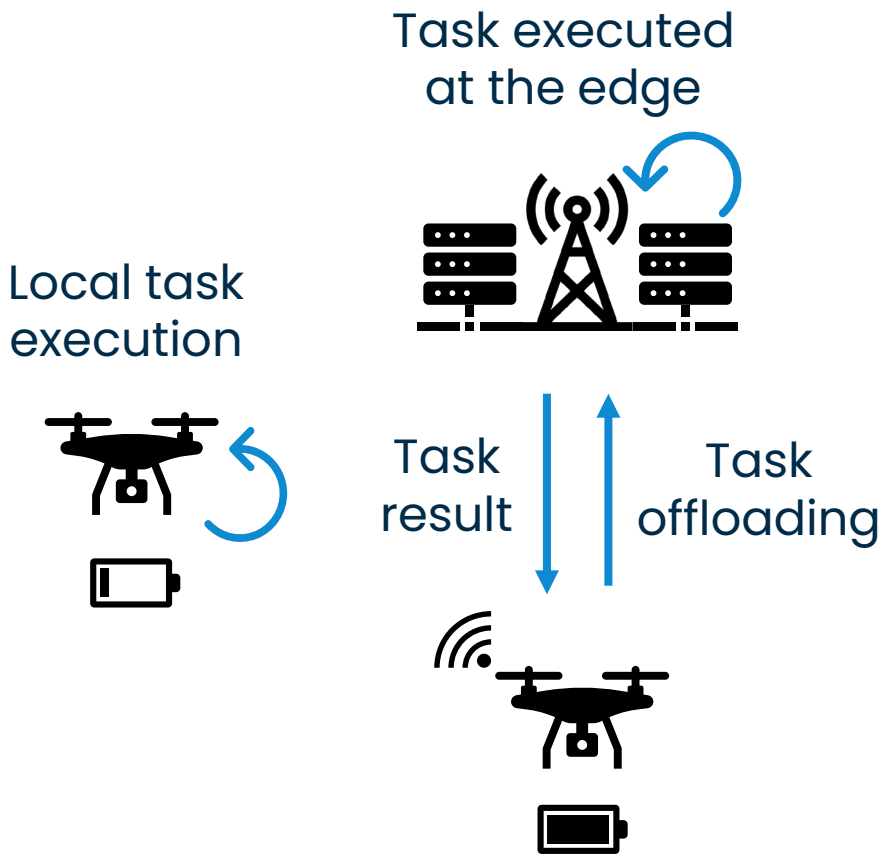


Offloading of semantic computer vision task

- Accuracy and latency-aware configuration of task offloading
- Flexible resource allocation
- Semantic image compression

Lightly compressed frame

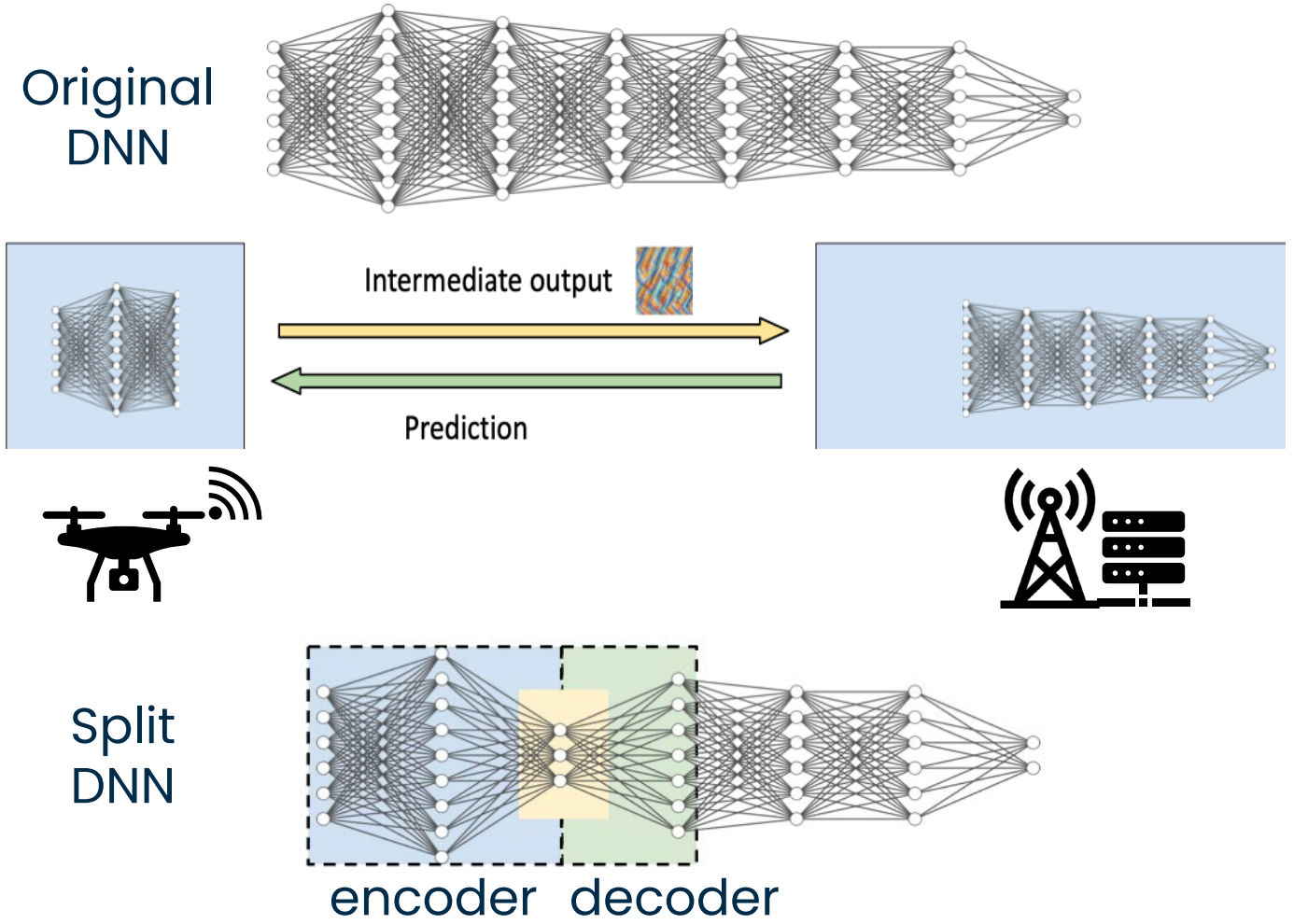
Highly compressed frame



C. Puligheddu, C.F. Chiasserini, et al., "SEM-O-RAN: Semantic and Flexible O-RAN Slicing for NextG Edge-Assisted Mobile Systems," IEEE INFOCOM 2023, New York City, NY, USA, 2023, doi: 10.1109/INFOCOM53939.2023.10228870.

Split computing for efficient inference

- Inference is not entirely at the edge
- Encoder and decoder are a bottleneck designed to:
  - Minimize network load by compression
  - Maximize inference accuracy
  - Preserve user privacy



Contatti

Mail: corrado.puligheddu@polito.it

Web: https://www.telematica.polito.it/

Politecnico
di Torino

#INNOVATION@DET

Machine Learning Inference Acceleration Using Embedded and Datacenter-Class FPGAs

M.R. Casu, L. Lavagno, M.T. Lazarescu,
F. Minnella, T. Urso, L. Urbinati, E. Manca

Politecnico
di Torino

Context & Competence

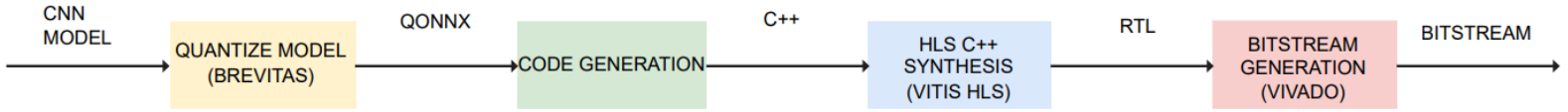
Context:

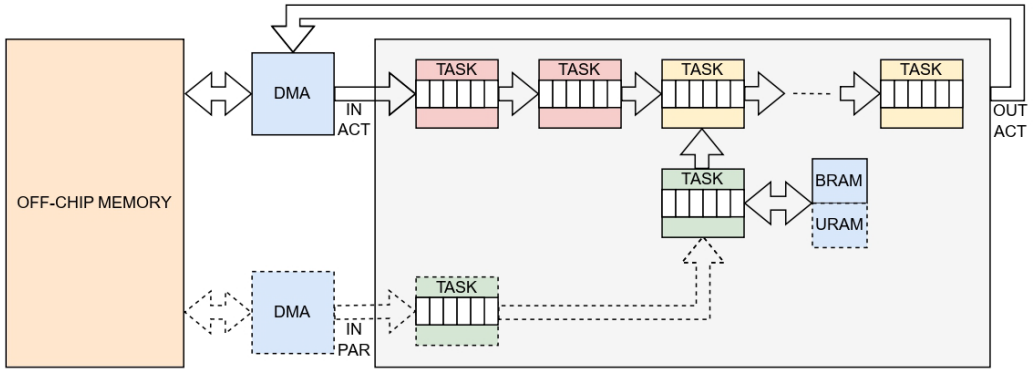
- Energy-efficient high-performance ML/DNNs algorithms targeting FPGAs to go beyond what CPUs/GPUs can get, while preserving flexibility to keep up with a continuous algorithmic innovation
- FPGAs can efficiently implement custom memory and computation architectures, with any precision level, beyond standard 8, 16 and 32 bits

Competence:

- Fine tuning of the implementation micro-architecture and quantization
- Broad experience using high-level synthesis to efficiently map algorithms modeled in C++ on various FPGA platforms
- Training with small INT & Fixed Point, to strike the perfect balance of accuracy and computation/memory resources

NN2FPGA : Neural Network Accelerators for Low-Power FPGAs Using High-Level Synthesis



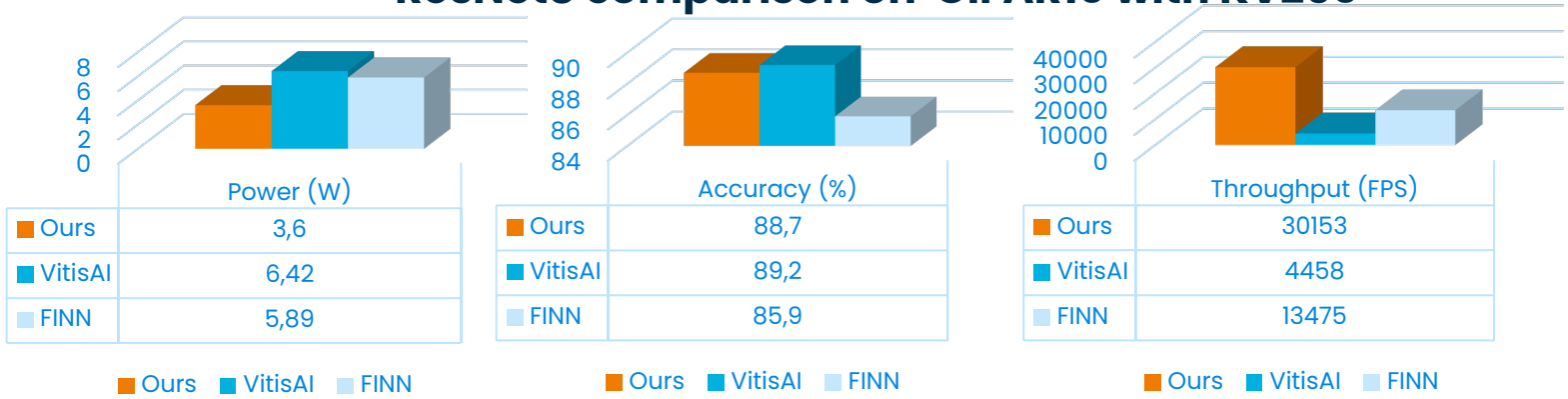


- **Dataflow** architecture
- **Fixed-point** quantization
- Compatible with **AMD-Xilinx boards**
- **High throughput/low power** tasks
- Optimized design for **skip connections**
  - **e.g., ResNets**

APPLICATIONS:

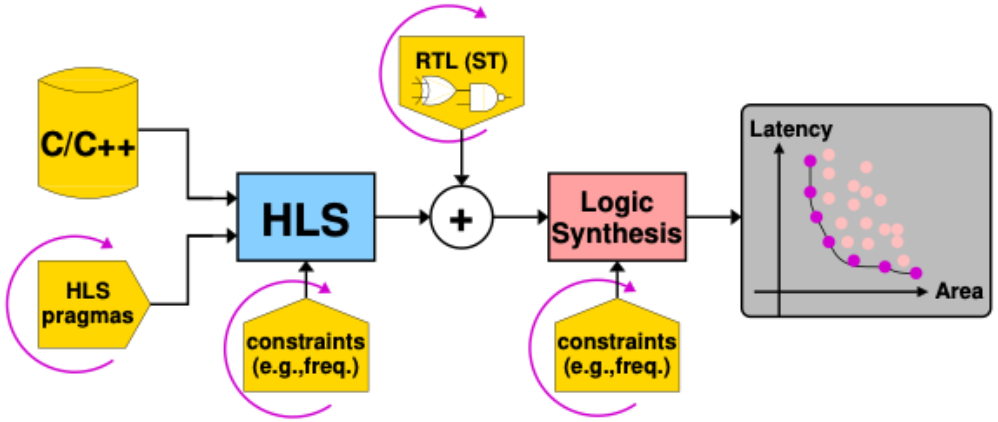
- Image classification
- Object detection
- NLP (in progress)
- Segmentation (in progress)

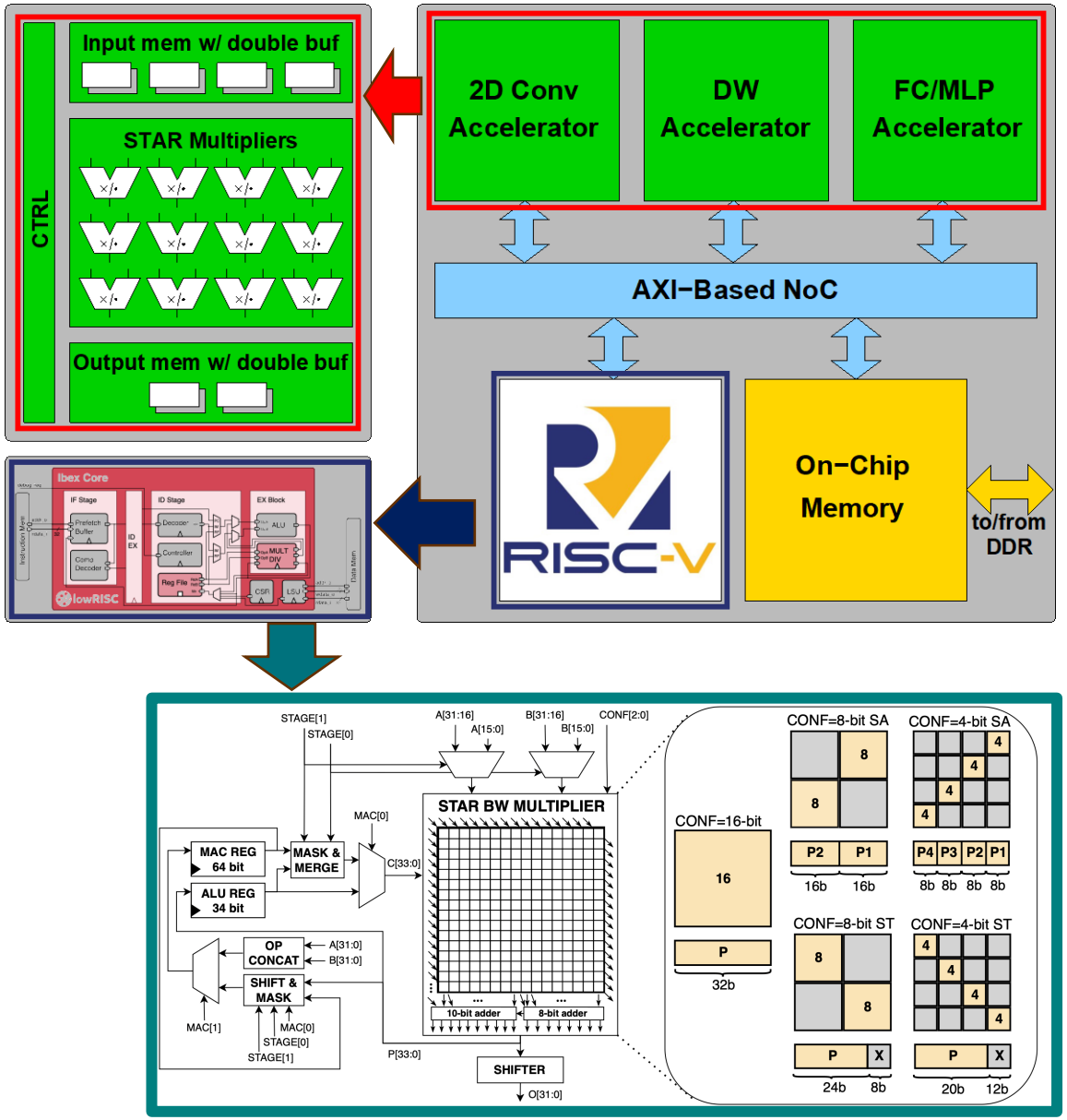
ResNet8 comparison on CIFAR10 with KV260



SoC for Embedded Machine Learning

- Specialized energy-efficient Accelerators and RISC-V based on **Sum-Together/Apart (STAR)** reconfigurable multipliers
  - **Mixed-Precision operation 4/8/16/32 bit**
- Accelerators obtained with **High-Level Synthesis (HLS)** and **Design Space Exploration**





Contacts

mario.casu@polito.it

Circ. & Sys. for ML@DET: https://tinyurl.com/DET-CAS-4-ML

MS Course: https://tinyurl.com/EDGE-COMP-SYS-FOR-ML

Politecnico
di Torino

#INNOVATION@DET

Application and Performance Oriented NN Optimizations for Embedded Systems

M.T. Lazarescu
G. Subbicini
L. Lavagno

Politecnico di Torino

Contesto e competenze

NN a volte *imprevedibili* (overfitting, adversarial, bias, rumore, giustificabili) e spesso *dispendiose* (addestramento, inferenza)
→ utilizzo limitato per sensori IoT, edge, mobile

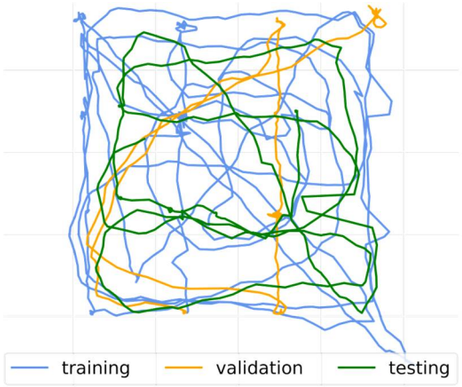
➔ Compattazione: *training* (multilivello, focus su caratteristiche essenziali), *architetturale* (NAS), *dati* (quantizzazione, ...),

➔ Energia: *acceleratori* (HW), *memorie* (accessi e architettura)

➔ Trasversale: per applicazione (conoscenza, strumenti): meno training, risorse, bias, più resilienti, giustificabili, generalizzabili

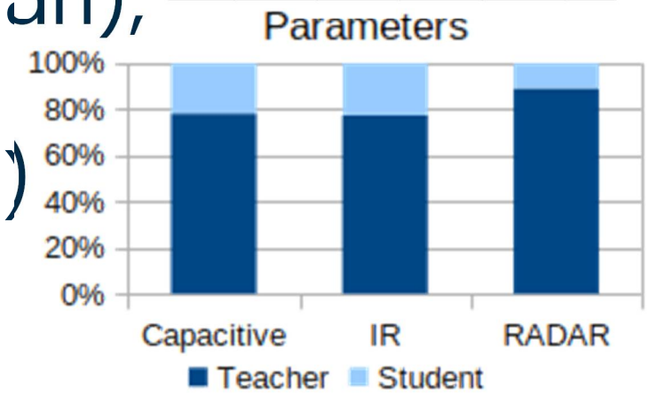
➔ Robustezza: stochastic computing, quantizzazione variabile

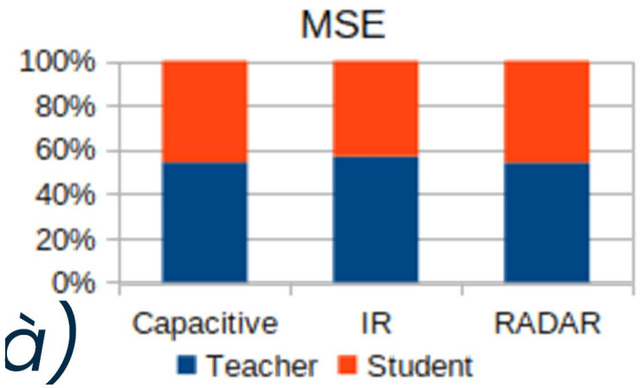
Monitoraggio multimodale attività umane



Localizzazione/tracciamento affidabile: *multimodale*
→ sensori *capacitivi* (molto rumorosi e non-lineari),
→ sensore *radar* (meno rumoroso, lineare),
→ sensore *IR 4x4 pixel* (poco rumoroso, molti dati)

NN ottimizzate per sistemi IoT edge/leaf immersi
→ *distillazione* 2 livelli, fonti diverse
→ *ottimizzazione* iperparametri
➔ *Risorse*: ridotte di **4–10 volte**
➔ *Accuratezza*: > 5 % **migliore**
Dipendente dai dati (dimensione, rumore, linearità)



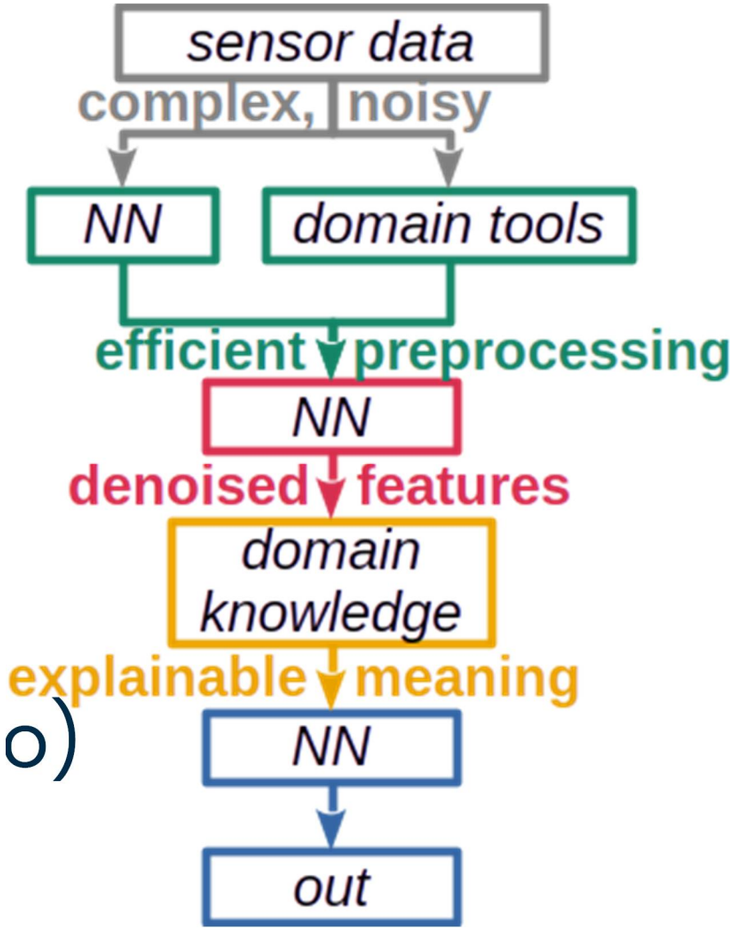


NN con strumenti e conoscenze formali

Conoscenza formale: efficiente, spiegabile,
ma «fragile» con *dati complessi e rumorosi*

NN: estrazione robusta di caratteristiche,
ma *inefficienti e opache sui formalismi*

NN allenate per usare *strumenti* e *conoscenze*
→ *minimizza* le NN (risorse, tempi, energia)
→ aumenta *affidabilità* e *generalizzazione*
(beneficio *dati sintetici* od off-site/laboratorio)
→ interazione *bidirezionale con esperti umani*



Contatti

Email: mihai.lazarescu@polito.it

Web: https://www.polito.it/en/staff?p=mihai.lazarescu

Politecnico
di Torino

#INNOVATION@DET

Deep learning inferencing on edge devices

M. Caon, F. Guella, M. Martina, G. Masera, E. Valpreda

Politecnico di Torino

Contesto e competenze

Inference acceleration:

- **Hardware platform** (FPGA, ASIC, MCU, etc)
- **Constraints** (energy, area, timing)
- **Performance** (accuracy, SNR, etc.)

Genetic algorithm

Reinforcement learning

ACTOR

CRITIC

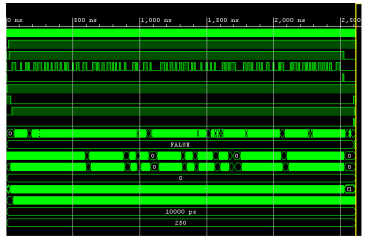
Problem specification

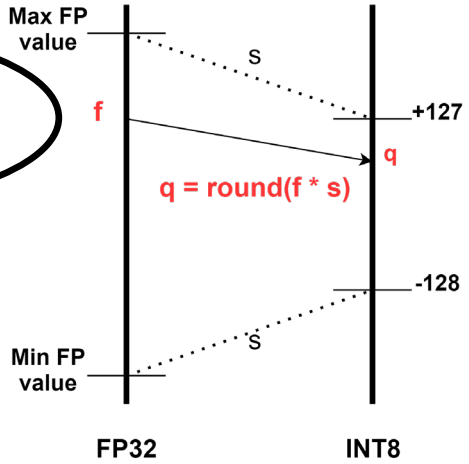
Model design and model compression

Mapping, HW acceleration, optimization

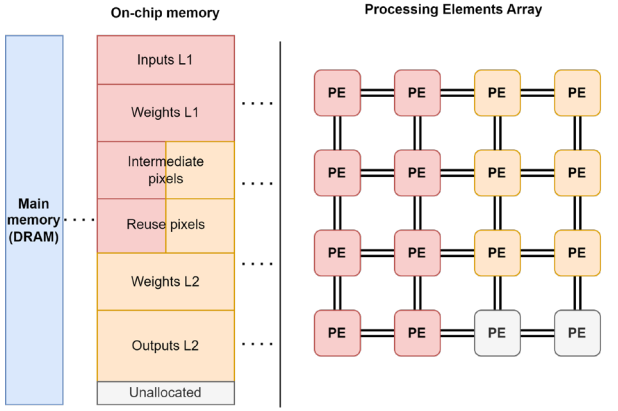
Simulation or execution

Performance evaluation



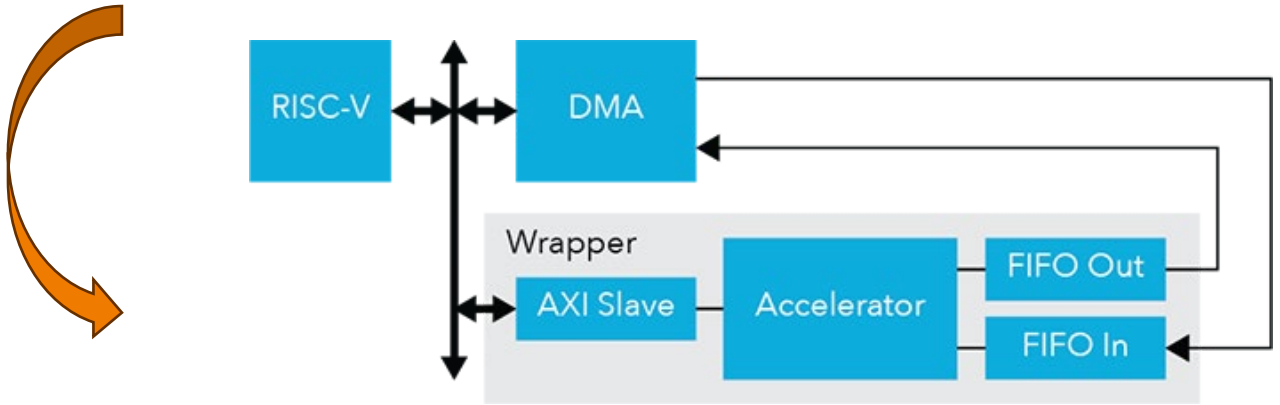


- Pruning
- Knowledge Distillation
- Quantization
- Weight sharing
- Arithmetic approximation
- Efficient Memory Access



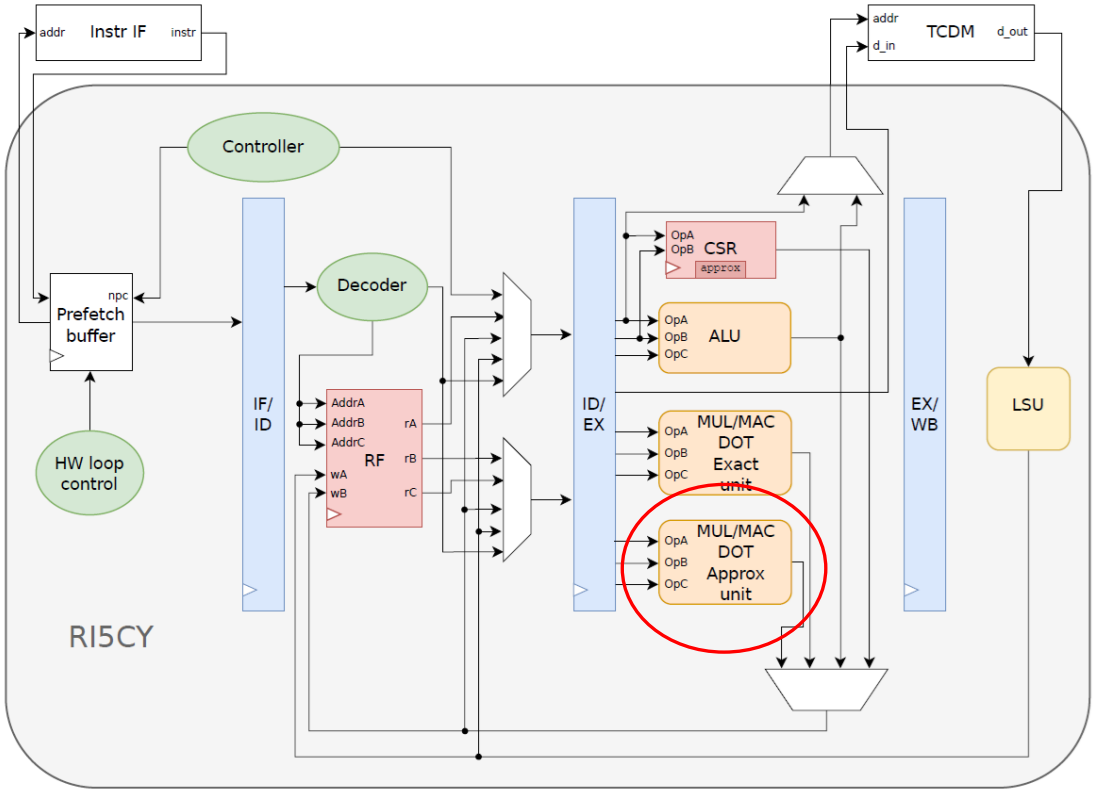
Soluzioni sviluppate / 1

EI platform	flexibility	Energy efficiency (Gop/s/W)	Peak performance (Gop/s)
CPU	High	*EI not feasible*	
GPU	High	*EI not feasible*	
ASIC/SoC	Low	10k – 100k	1k – 30k
FPGA	Medium	5 - 150	10 - 400
MCU + acceleration	High	100 - 500	1 - 5



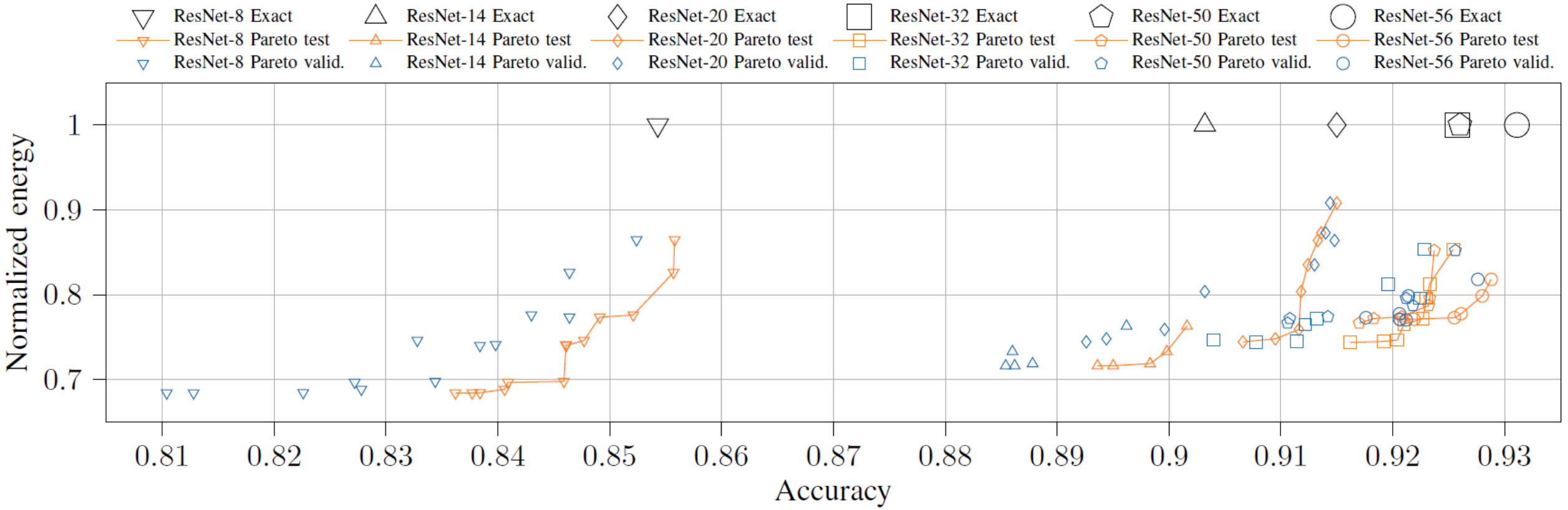
Loosely-coupled accelerators

RI5CY core with extensions



Tightly-coupled accelerators

Soluzioni sviluppate / 2



- Ottimizzazione di diverse reti allenate su specifici dataset e eseguite su RISC-V
- Allocazione di diversi moltiplicatori approssimati layer per layer
- Scelta soluzioni migliori rispetto a energia e precisione

Contatti

Mail: guido.masera@polito.it

www.det.polito.it/research/research_groups/electronics/vlsilab_group

Politecnico
di Torino

#INNOVATION@DET

Compression of Neural Networks for Tiny Machine Learning

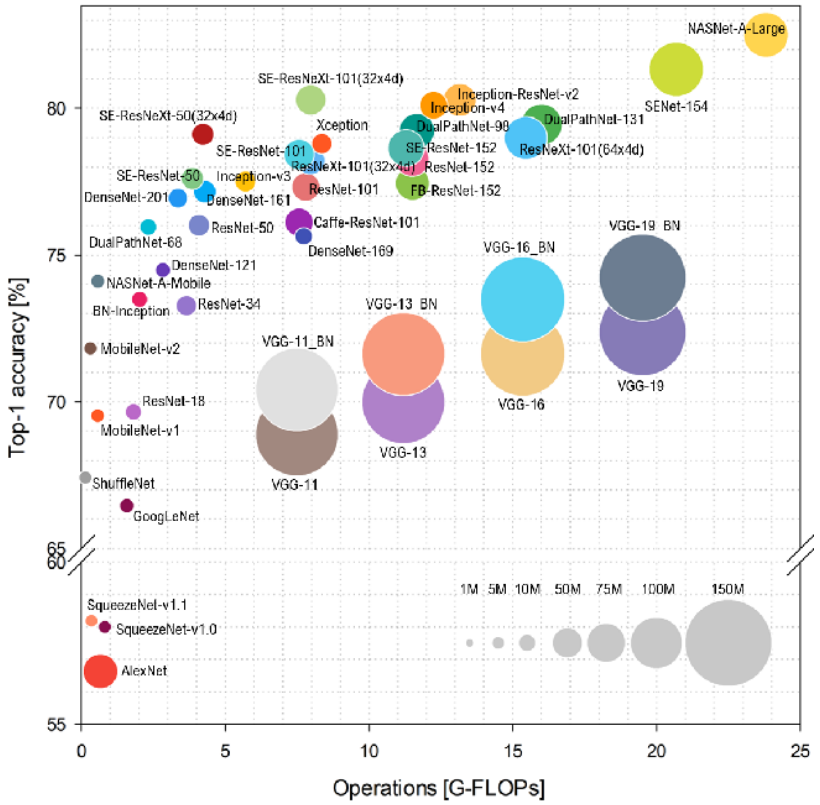
L. Prono, F. Pareschi, G. Setti, P. Bich, C. Boretti

Politecnico di Torino

Contesto e competenze

Tiny Machine Learning: fit algorithms on devices with ~100 kB memory

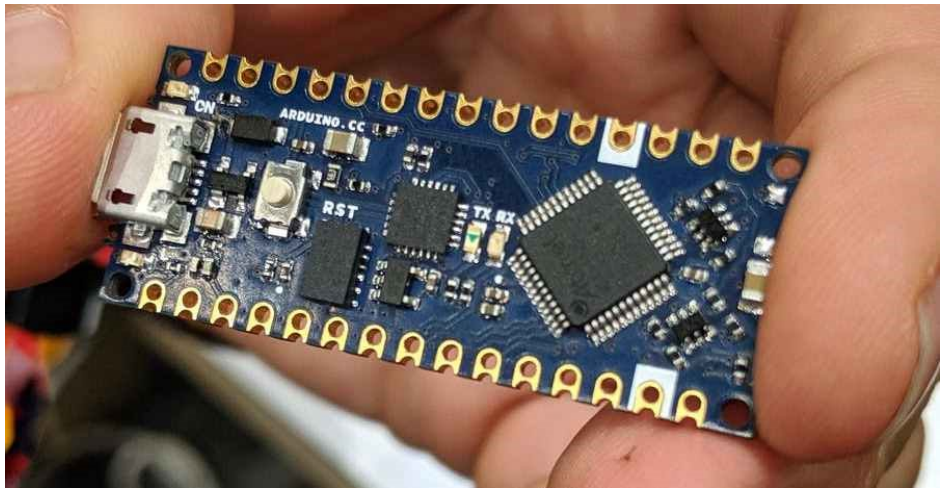
Today's Neural Network models are getting bigger and bigger…



With model compression, we can use them on small devices which run on batteries!

- Small memory footprint
- Small power consumption

Model Compression



Soluzioni sviluppate / 1

Anomaly Detection algorithms for on-board Satellite fault detection

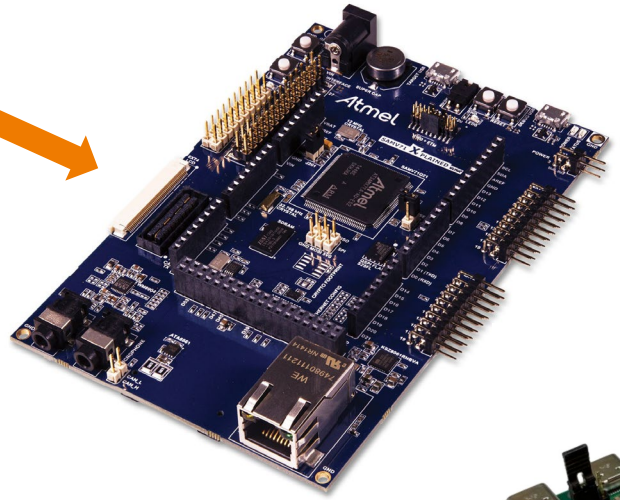
Fully Portable Application Software

Low-power/fast-inference

Anomaly Detection Algorithms:
- Principal Component Analysis
- Autoregressive Model
- Convolutional autoencoder
- Long-short Term Memory DNN

Model compression:
- Quantization
- Pruning
- Knowledge distillation

ATSAMV71 (ARM Cortex-M7)



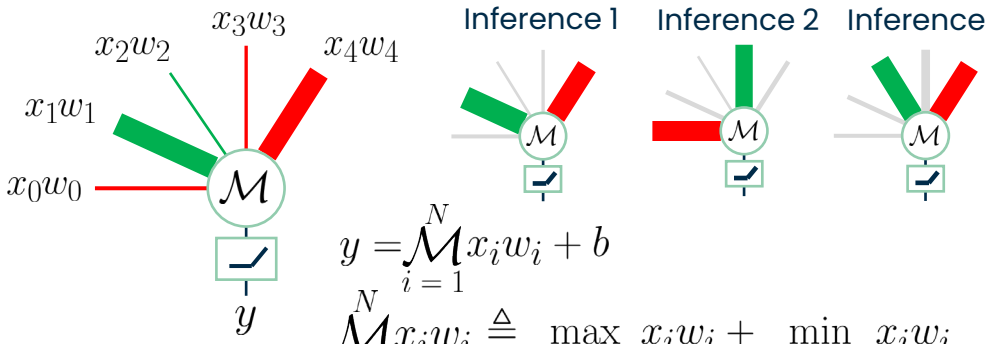
GR740 (Rad-hard device)

In collaboration with Thales Alenia Space

Soluzioni sviluppate / 2

Novel neuron paradigm for Deep Neural Network model compression

Multiply and Max/min (MAM)-based neurons **sum together only the maximum and minimum contributes.**



$$y = \overset{N}{\underset{i=1}{\mathcal{M}}} x_i w_i + b$$

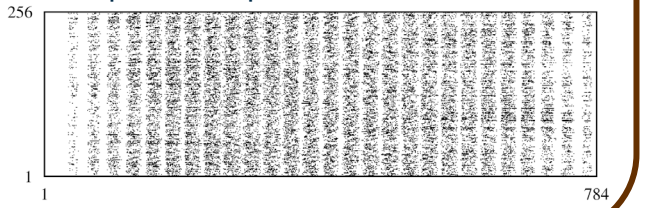
$$\overset{N}{\underset{i=1}{\mathcal{M}}} x_i w_i \triangleq \max_{i\in\{1,..,N\}} x_i w_i + \min_{i\in\{1,..,N\}} x_i w_i$$

MAM neural networks use **only a small subset of interconnections during inference.**

Highly prunable DNNs

Low parameters count

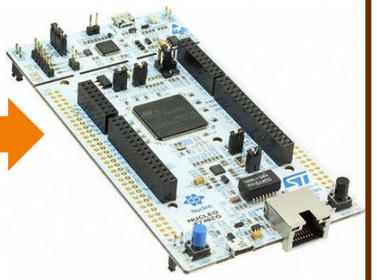
Example of map of the interconnections



ECG Autoencoder (Deep Neural Network) **~2.190.000 parameters**

Pruning and quantization (model compression)

STM32F767ZI µC unit (512 kB RAM, 216 MHz Clock)

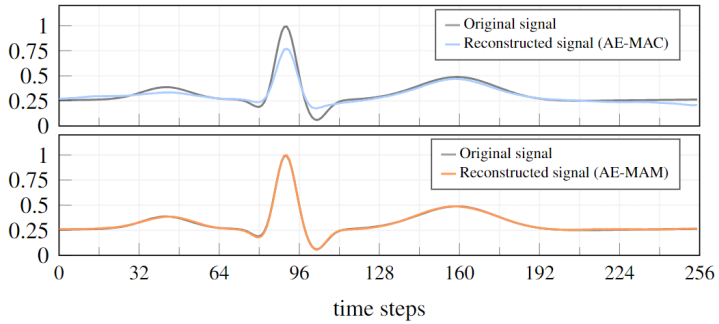


The pruned model contains 220.000 8-bit parameters (220 kB)

MAM DNN: 10x memory reduction with **no accuracy loss.**

- Standard DNN: 18 dB reconstruction
- **MAM DNN: 33 dB reconstruction**

Reconstructed ECG (standard DNN vs MAM)



Prono, Luciano; Bich, Philippe; Mangia, Mauro; Pareschi, Fabio; Rovatti, Riccardo; Setti, Gianluca (2023). A Multiply-And-Max/min Neuron Paradigm for Aggressively Prunable Deep Neural Networks. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22561567.v1

Contatti

Mail: luciano.prono@polito.it

Politecnico
di Torino

#INNOVATION@DET

Il ruolo dell'incertezza nelle tecniche di machine learning

A. Carullo, S. Corbellini,
L. Lombardo, M. Parvis, A. Vallan

Politecnico di Torino

Contesto …

- Valutazione del carico vocale dei professionisti della voce al fine di prevenire l'insorgenza di patologie a carico dell'apparato fonatorio e qualificare l'acustica architettonica
- Diagnosi precoce di disfonie e patologie neurologiche e valutazione dell'efficacia di tecniche chirurgiche e di riabilitazione

… e competenze

- ✓ Progetto, sviluppo e caratterizzazione metrologica di dosimetri vocali basati su microfono a contatto
- ✓ Acquisizione di materiale vocale da diverse categorie di soggetti
- ✓ Sviluppo di algoritmi per l'estrazione di parametri del segnale vocale nel dominio del tempo, della frequenza e della «quefrenza»
- ✓ Classificazione delle diverse categorie di soggetti tramite tecniche di machine learning

Classificazione di pazienti soggetti a laringectomia
(collaborazione con Ospedale San Giovanni Bosco – TO)

Tre classi corrispondenti al tipo di intervento

- ↳ OPHL-I (meno invasivo), OPHL-II, OPHL-III (più invasivo)

Estrazione di parametri vocali dai *frames* armonici di voce da 3 ripetizioni della vocale /a/ sostenuta e dalla lettura di un brano

- ↳ Estratte **189**(**198**) *features* dalla lettura(vocale)

- ➢ **Feature selection?**
- ➢ **Modello di classificazione?**

Classificazione di pazienti soggetti a laringectomia
(collaborazione con Ospedale San Giovanni Bosco – TO)

- **Modello di classificazione?**

  Regressione logistica → restituisce la probabilità di appartenere a una classe (utile per valutare l'efficacia di tecniche di riabilitazione)

- **Feature selection?**

  Guidata dal modello di classificazione → ricerca di gruppi di *features* (massimo 4) a bassa correlazione che permettono di identificare i parametri del modello in modo significativo ($p < 0.05$)

  ↳ Scelta dei gruppi che forniscono le migliori prestazioni

Classificazione di pazienti soggetti a laringectomia
(collaborazione con Ospedale San Giovanni Bosco – TO)

Data set

OPHL-I
(negative class)

OPHL-II
OPHL-III
(positive class)

Feature extraction and selection

Model identification

$$p = \frac{e^{\theta \cdot F}}{1 + e^{\theta \cdot F}}$$

$$\theta \cdot F = \beta_0 + \beta_1 \cdot F_1 + \cdots + \beta_N \cdot F_N$$

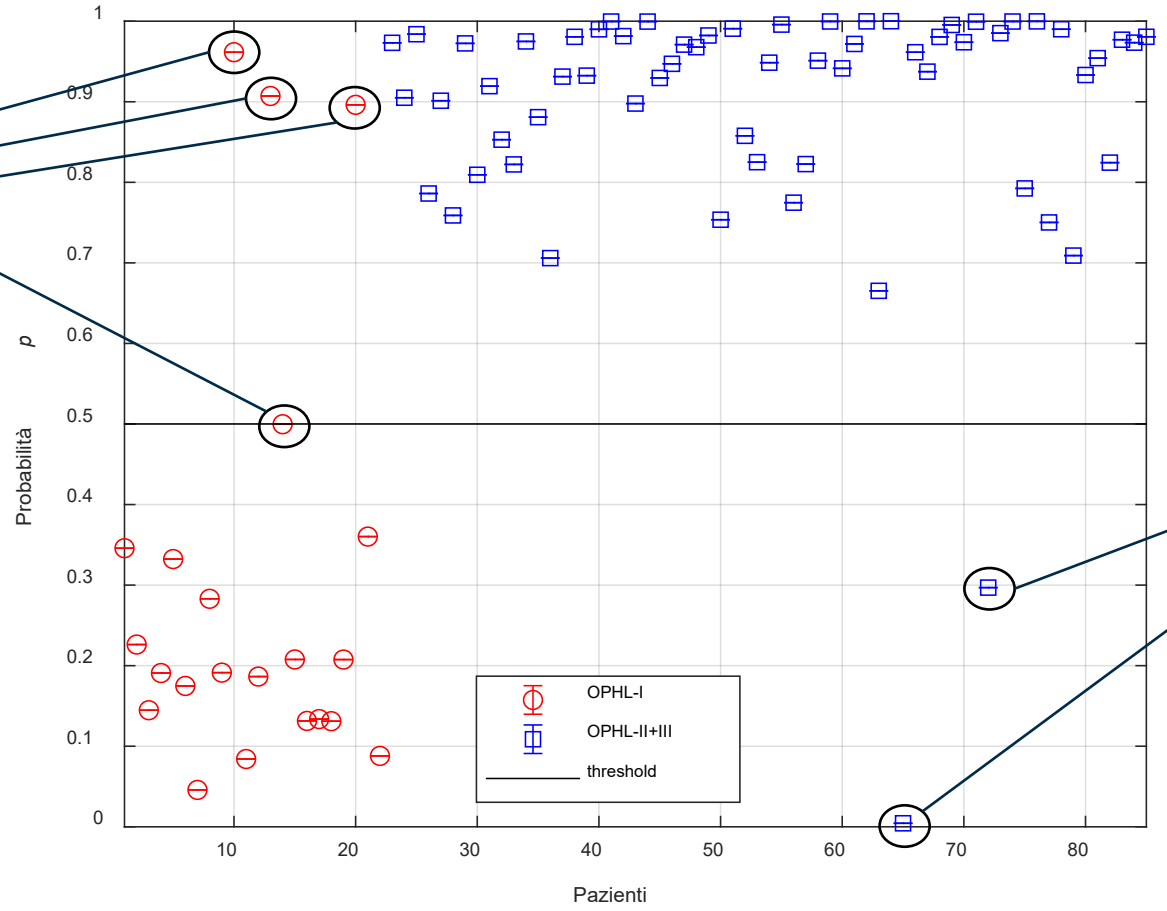
F_i sono le *features* (ingressi) del modello

β_i sono i parametri (identificati) del modello

Classificazione di pazienti soggetti a laringectomia
(collaborazione con Ospedale San Giovanni Bosco – TO)

Esempio di risultati: vocale /a/, 2 *features* (parametri β_0, β_1, β_2)

Falsi positivi



Falsi negativi

Classificazione di pazienti soggetti a laringectomia (collaborazione con Ospedale San Giovanni Bosco – TO)

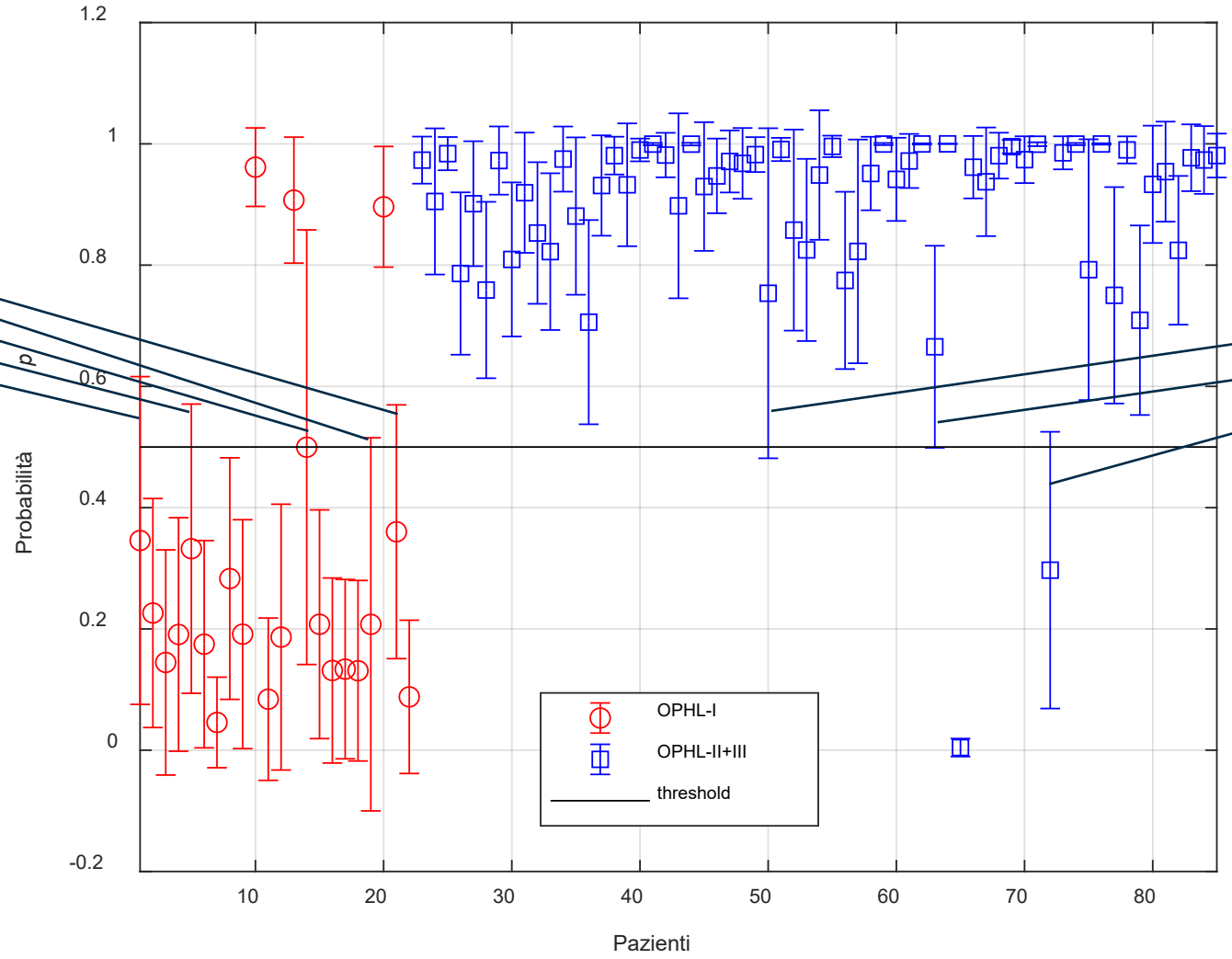
Esempio di risultati: vocale /a/, 2 *features* (parametri β_0, β_1, β_2)

...ma

- le 2 *features* di ingresso sono affette dall'incertezza con cui sono state misurate/estratte
- i 3 parametri del modello sono affetti da incertezza, che dipende dalla fase di identificazione del modello, quindi anche dall'incertezza delle *features* utilizzate nella fase di *training*
- la correlazione tra i parametri β_0, β_1, β_2 può essere non trascurabile

Classificazione di pazienti soggetti a laringectomia
(collaborazione con Ospedale San Giovanni Bosco – TO)

Soggetti NON classificabili



Soggetti NON classificabili

Eliminazione dei soggetti «dubbi» e valutazione di metriche realistiche

Conclusioni e consigli

➢ Fase di training

- *Features* di ingresso misurate con «bassa» incertezza
- *Features* di ingresso pesate in modo inverso rispetto all'incertezza
  - ↳ Parametri β_i del modello identificati con minore incertezza

➢ Fase di validazione/previsione

- Tenere sempre conto dell'incertezza di modello e delle incertezze delle *features* di ingresso
- Possibile utilizzo di *features* di ingresso misurate con incertezza più elevata
  - ↳ Rischio di aumento dei soggetti non classificabili

Contatti

Mail: alessio.carullo@polito.it

Politecnico
di Torino