Franca Rocco di Torrepadula

Advancing Edge AI Systems for Smart Cities

Tutor: Prof. Mazzocca

Cycle: XXXVII

co-Tutor: Prof. Di Martino

Year: Third

Candidate's information

- MSc degree in Computer Engineering (October 2021)
- DIETI Research group/laboratory: SECLAB
- PhD start date 01/11/2021 – end date 31/10/2024
- Scholarship type: UNINA
- Periods abroad:
  - University College of Dublin, Ireland. Under the supervision of Prof. Gavin McArdle (24/01/2023-04/02/2023)
  - L3S Research Center, Leibniz University, Hannover, Germany. Under the supervision of Prof. Wolfgang Nejdl (06/11/2023-06/03/2024)
- Scientific/Industrial Collaborations:
  - Hitachi Rail Company
  - ETH Zurich

Summary of study activities

- **Ad hoc PhD courses / schools:**
  - Virtualization technologies and their applications
  - Statistical data analysis for science and engineering research
  - Imprenditorialità Accademica
  - IoT Data Analysis
  - Semantic artifacts and multimedia knowledge graphs for bio-data integration
  - 2023 Spring School in Transferable Skills
  - Ethics and AI
  - Strategic Orientation for STEM Research & Writing
- **Conferences / events attended:**
  - International Conference on the Quality of Information and Communications Technology (QUATIC2022).
  - International Symposium on Web and Wireless Geographical Information Systems (W2GIS2022).
  - International Symposium on Web and Wireless Geographical Information Systems (W2GIS2023). **Winner of the Best Presentation Award** for the presentation of the paper *Bus Journey Time Prediction with Machine Learning: An Empirical Experience in Two Cities.*
  - International Symposium on Web and Wireless Geographical Information Systems (W2GIS2024).
  - ACM SIGSPATIAL Workshop on Sustainable Mobility.

Research area(s)

- The research primarily focuses on distributing artificial intelligence (AI) within smart city environments to overcome the **privacy** and **energy** limitations of traditional cloud-centric approaches.
- The contribution is positioned within the research fields of **Federated Learning** and **Edge AI**.

Research results

- A Systematic Literature Review on data-driven passenger flow prediciton.
- A visual-based toolkit for mobility analytics
- A methodology for generating synthetic mobility dataset, leveraging Eclipse SUMO.
- A reference architecture for Intelligent Public Transportation Systems.
- FedFlow: A personalized federated learning framework for passenger flow predictive systems
- X-PILOT: A framework for designing on-board and explainable passenger flow predictive systems based on XGBoost
- A workflow for distilling knowledge in low-carbon Edge AI applications, avoiding grid search.

PhD Thesis

Research Products

[J1]	S. Di Martino, E. Landolfi, N. Mazzocca, **F. Rocco di Torrepadula**, L. L. L. Starace, A visual-based toolkit to support mobility data analytics, *Expert Systems with Applications* [Published]
[J2]	A. Cilardo, V. Maisto, N. Mazzocca, **F. Rocco Di Torrepadula**, An approach to the systematic characterization of multitask accelerated CNN inference in edge MPSoCs, *ACM Transactions on Embedded Computing Systems* [Published]
[J3]	**F. Rocco di Torrepadula**, S. Di Martino, N. Mazzocca, P. Sannino, A Reference Architecture for Data-Driven Intelligent Public Transportation Systems, *IEEE Open Journal of Intelligent Transportation Systems* [Published]
[J4]	**F. Rocco di Torrepadula**, E. V. Napolitano, S. Di Martino, N. Mazzocca, Machine Learning for public transportation demand prediction: A Systematic Literature Review *Engineering Applications of Artificial Intelligence* [Published]
[J5]	L. L. L. Starace, **F. Rocco di Torrepadula**, S. Di Martino, N. Mazzocca, Vehicular Crowdsensing with High-Mileage Vehicles: Investigating Spatiotemporal Coverage Dynamics in Historical Cities with Complex Urban Road Networks, *Journal of Advanced Transportation* [Published]
[J6]	M. Barbareschi, A. Emmanuele, N. Mazzocca, **F. Rocco di Torrepadula**, Designing On-Board Explainable Passenger Flow Prediction, *Expert Systems with Applications* [Under the 2nd round of revision]
[J7]	**F. Rocco di Torrepadula**, A. Somma, A. De Benedictis, N. Mazzocca, Smart Ecosystems and Digital Twins: an architectural perspective and a FIWARE-based solution, *IEEE Software* [Under the 2nd round of revision]

Research Products

[C1]	A. De Benedictis, **F. Rocco di Torrepadula**, A. Somma, A Digital Twin Architecture for Intelligent Public Transportation Systems: A FIWARE-Based Solution, *International Symposium on Web and Wireless Geographical Information Systems* [Published]
[C2]	F. Amato, S. Di Martino, N. Mazzocca, D. Nardone, **F. Rocco di Torrepadula**, P. Sannino, Bus Passenger Load Prediction: Challenges from an Industrial Experience, *International Symposium on Web and Wireless Geographical Information Systems* [Published]
[C3]	A. Cilardo, V. Maisto, N. Mazzocca, **F. Rocco di Torrepadula**, A Proposal for FPGA-Accelerated Deep Learning Ensembles in MPSoC Platforms Applied to Malware Detection, *International Conference on the Quality of Information and Communications Technology. QUATIC* [Published]
[C4]	L. Dunne, **F. Rocco Di Torrepadula**, S. Di Martino, G. McArdle, D. Nardone, Bus Journey Time Prediction with Machine Learning: An Empirical Experience in Two Cities, *International Symposium on Web and Wireless Geographical Information Systems* [Published]
[C5]	S. Di Martino, N. Mazzocca, **F. Rocco Di Torrepadula**, L. L. L. Starace, Mobility Data Analytics with KNOT: The KNime mObility Toolkit, *International Symposium on Web and Wireless Geographical Information Systems* [Published]
[C6]	**F. Rocco Di Torrepadula**, D. Russo, S. Di Martino, N. Mazzocca, P. Sannino, Using SUMO towards Proactive Public Mobility: Some Lessons Learned, *1st ACM SIGSPATIAL Workshop on Sustainable Mobility* [Published]

PhD thesis overview: Problem

Smart city services often rely on DNNs (Deep Neural Networks), being the state-of-the-art techniques in many domains. However, given their complexity, these models raise several challenges:

High energy consumptions

Privacy concerns (given their typical cloud-centric deployment)

Low interpretability

PhD thesis overview: Problem

Smart city services often rely on DNNs (Deep Neural Networks), being the state-of-the-art techniques in many domains. However, given their complexity, these models raise several challenges:

High energy consumptions

Privacy concerns (given their typical cloud-centric deployment)

GDPR

Low interpretability

PhD thesis overview: Objective

Distributing the intelligence within the smart city, to address the concerns of traditional cloud-based systems.

This involves developing energy-efficient models that can be effectively deployed at the edge, while also enhancing interpretability to ensure transparency and trust in AI-driven applications.

PhD thesis overview: Methodology

- **Federated Learning (FL)** to enable decentralized and collaborative model training across smart city entities, addressing privacy concerns by keeping data localized.

- **Edge AI** for enabling the execution of ML models at the edge, leveraging:
  - The definition of **lightweight models** (e.g. not based on DNNs).
  - Optimizing existing, pre-trained models through **Knowledge Distillation (KD).**

Contribution 1:

A Personalized FL Framework for Passenger Flow Prediction

Problem

- Passenger Flow (PF) prediction focuses on forecasting the number of people using a specific service based on historical data.

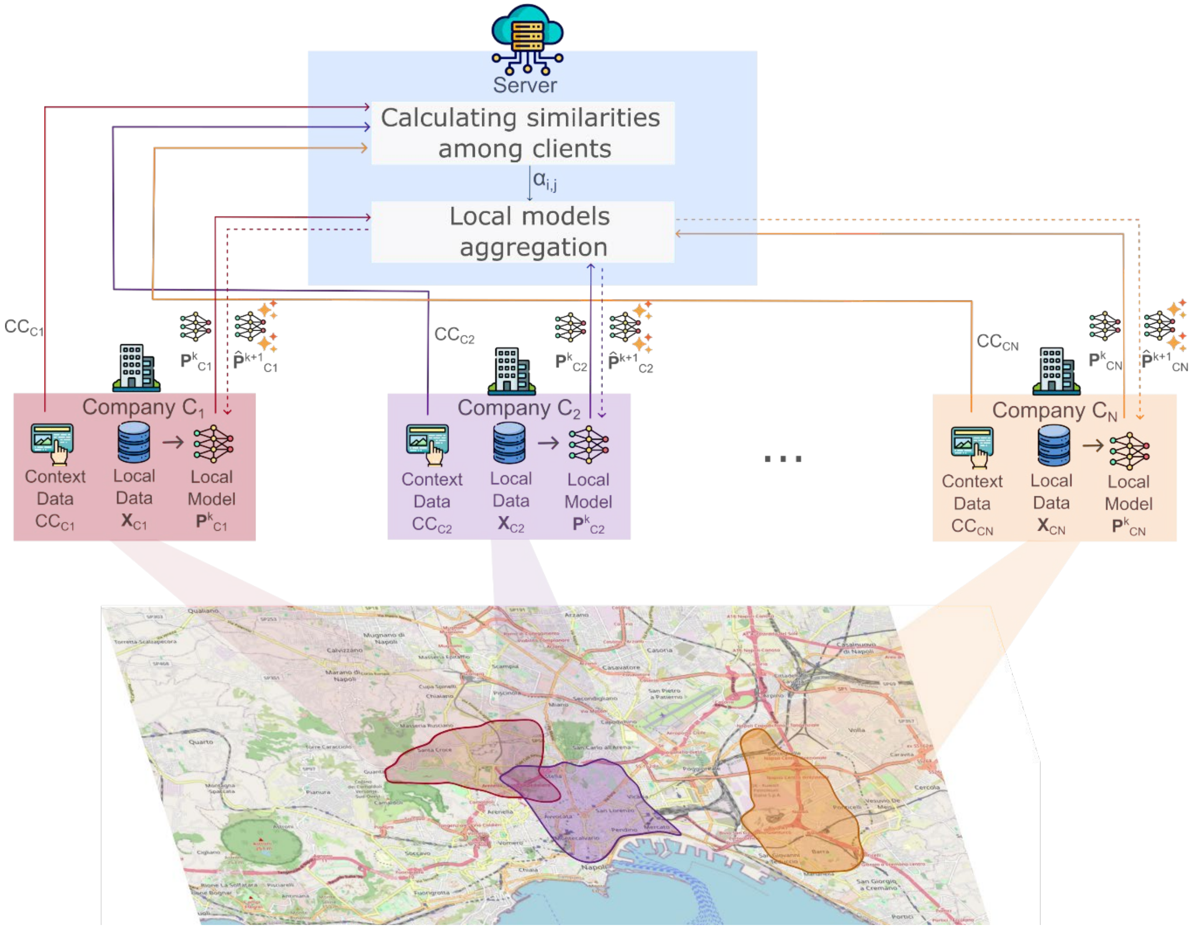
- **Privacy** is a key concern in PF prediction, as the data often includes sensitive information about passengers' routes, travel patterns, frequently visited locations, and routines.
- FL offers a promising solution to address these privacy concerns, but a significant challenge is handling **data heterogeneity**, which is common in smart city environments.

State of Art

- The typical PF prediction setting involves training deep learning models (typically RNNs or GNNs) on a central back-end, raising significant security/privacy concerns
- Despite being a valuable solution to privacy concerns in many domains, there is a lack of work aiming at exploiting FL for PF prediction.
- Several proposals have integrated FL into other mobility-related tasks, especially regarding traffic predictions, which typically do not address data heterogeneity.

Contribution

FedFlow: A personalized federated learning framework for passenger flow prediction.



To tackle data heterogeneity:

- A personalized model is realized for each client, giving more emphasis to the most similar clients.
- Similarities are calculated based on publicly available information about clients services.

Results

Metric	Horizon	Predictive Technique							
		NF	ARIMA	XGBoost	CNN	LSTM	Centr	FedAvg	**FedFlow**
MAE	1	3.58	2.67	2.83	2.80	2.31	2.87	3.67	**2.22**
	2	5.98	4.82	4.03	3.93	3.23	4.02	5.57	**3.11**
	3	7.96	6.87	4.80	4.75	3.88	4.83	6.99	**3.74**
RMSE	1	9.07	5.82	6.17	6.88	5.89	7.14	8.52	**5.75**
	2	12.53	8.93	7.82	8.64	7.35	8.83	11.17	**7.15**
	3	15.19	10.94	8.82	9.77	8.31	9.97	13.03	**8.08**
MAPE	1	326.98	61.98	64.93	49.56	36.32	48.24	75.29	**34.77**
	2	308.31	117.45	91.94	64.26	49.36	66.89	113.76	**47.96**
	3	291.69	170.32	109.21	76.82	58.46	80.58	147.35	**56.47**

Table 1 The average evaluation metrics, with predictive horizon ranging from 1 to 3. The best performance among all models is highlighted.

- Compared to LSTM, the centralized and the FedAvg approaches present a performance drop: a single global model struggles to accommodate the diverse local client data distributions.
- FedFlow addresses the heterogeneity challenge, significantly outperforming the FedAvg and centralized frameworks across all considered metrics.
- It also surpasses the performance of locally trained LSTM models. This improvement is attributed to the collaborative nature of the framework.

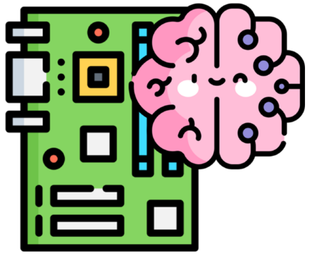
Contribution 2:

Designing Lightweight and Explainable PF Predictive Models

Problem

- Being a distribute learning paradigm, FL *enables* the execution of ML/DL models at the edge.

- However, running ML, especially DL, models at the edge introduces significant challenges due to the strict **limitations of edge devices**, such as constrained computational and storage capabilities.



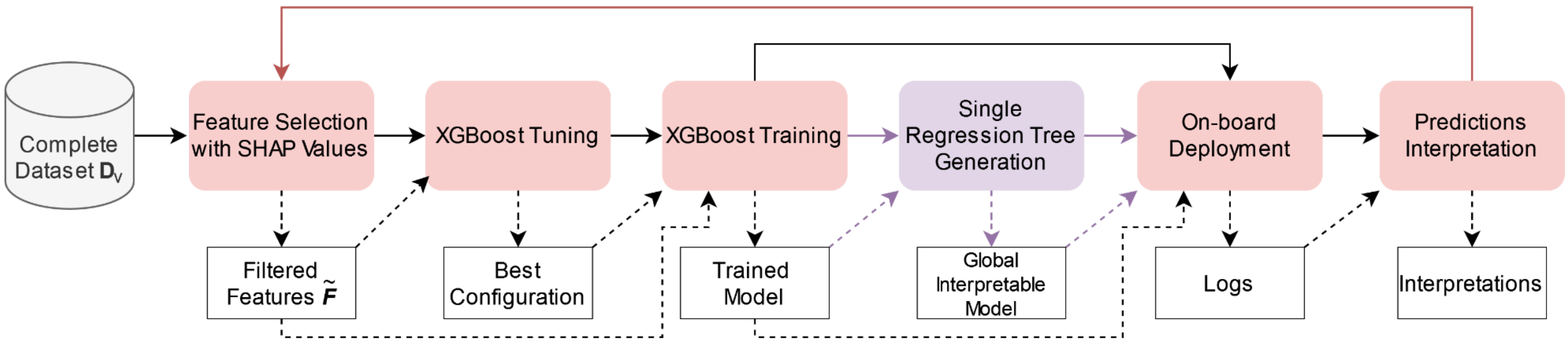
itee PhD
information technology
electrical engineering

State of Art

- On the algorithm side, the proposed Edge AI solutions can be broadly classified into two main approaches:
  - Creating smaller and more efficient networks from the outset
  - Compressing existing, pre-trained models through a process known as *model compression.*

Contribution

X-PILOT: a framework for designing onboard and explainable PF prediction based on XGBoost



- *XGBoost is a tree-ensemble model that offers high accuracy while reducing computational costs w.r.t DL models.*
- *Being based on simple trees, many solutions for explaining its predictions are already proposed in the scientific community, e.g. SHAP values.*

Results

Table 5.2. MAPE and MAE for each model, averaged across all the considered buses.

	LSTM	CNN	XGBoost	RF	ARIMA
MAPE	**23.288**	24.003	24.057	24.562	33.113
MAE	**3.448**	3.525	3.49	3.53	4.55

Table 5.3. Mean time and energy required to complete an in inference in microseconds and milli-Joule respectively.

	ARIMA	XGBoost	CNN	CNN-TL	LSTM	LSTM TL	RF
Time(us)	**10**	55	249	147	16614	n.s.	n.s.
Energy(mJ)	**0.014**	0.06	0.278	0.166	18.4	n.s.	n.s.

- Energy vs Accuracy trade-off (from LSTM to XGBoost):
  - Energy saving: 97.8%.
  - Accuracy reduction: 3.3%.

itee PhD
information technology
electrical engineering

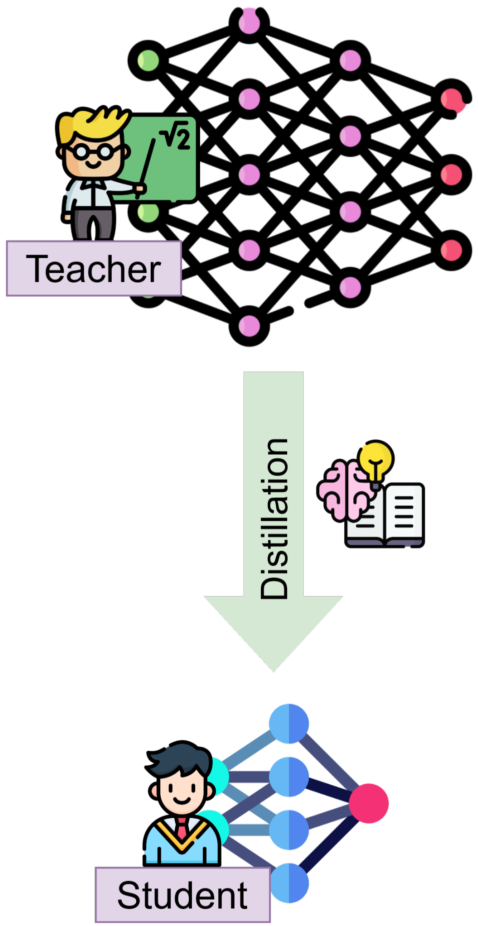
Contribution 3:

Distilling Knowledge for Low-Carbon ML Models

Problem

- Knowledge Distillation (KD) has emerged as a promising technique for ML model compression, however it further complicates the training process as it involves tuning additional hyperparameters.

- Hence, while KD is commonly employed to develop energy-efficient models, the tuning process can be inefficient, typically relying on **expensive grid-search** methods.



State of Art

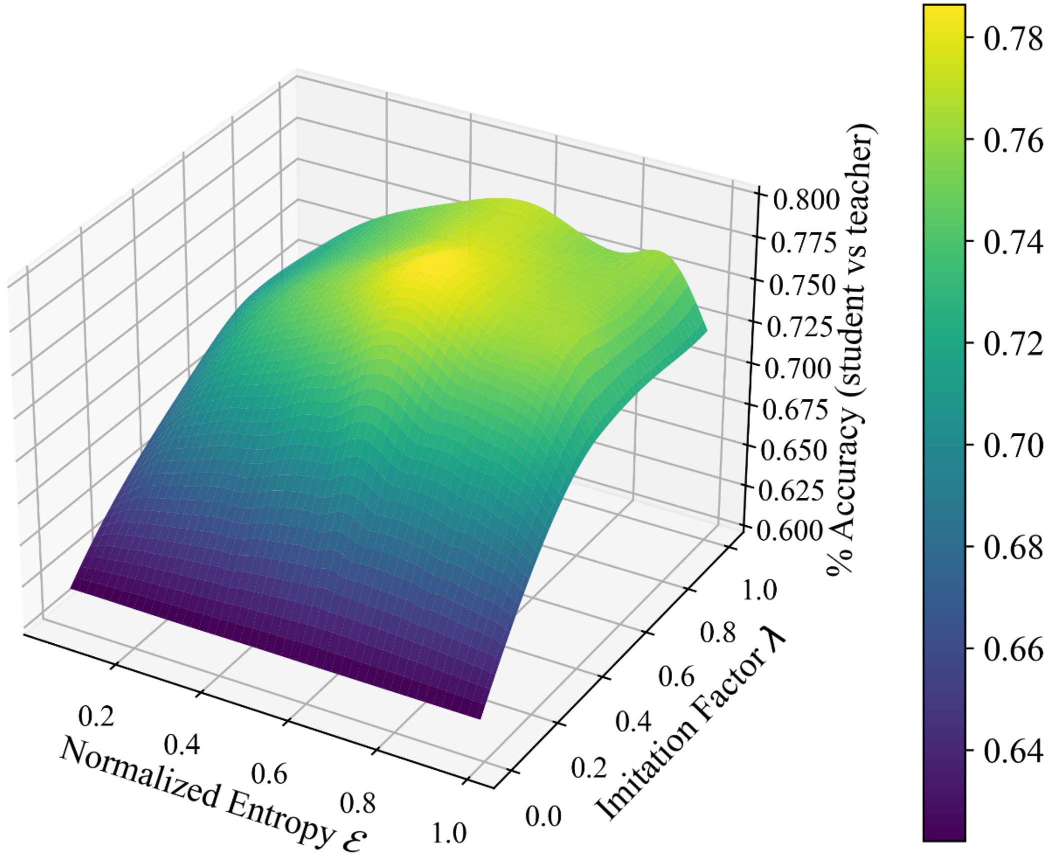
- Several works demonstrated the effectiveness of the KD approach across different domains.
- However, there is a lack of theoretical work meant to understand why and how KD works.
- An unsolved issue is how to optimally configure the additional hyperparameters, avoiding expensive grid search.

Contribution

A Geometrical Interpretation of KD leveraging the Shannon Entropy

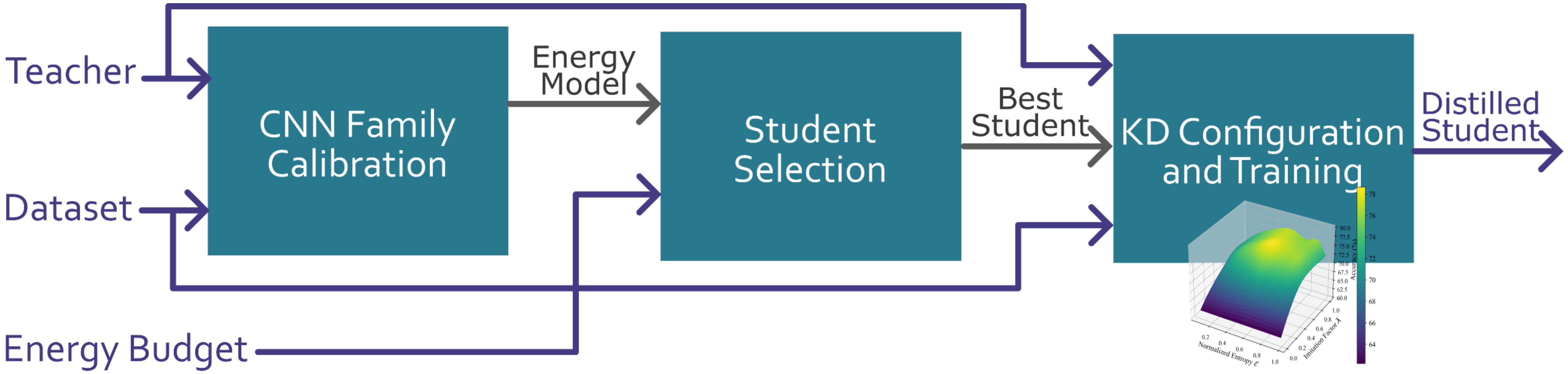
The best hyperparameters configuration mainly depends on:

- The accuracy of the teacher
- The accuracy of the student
- The model capacity gap between the teacher and the student



Contribution

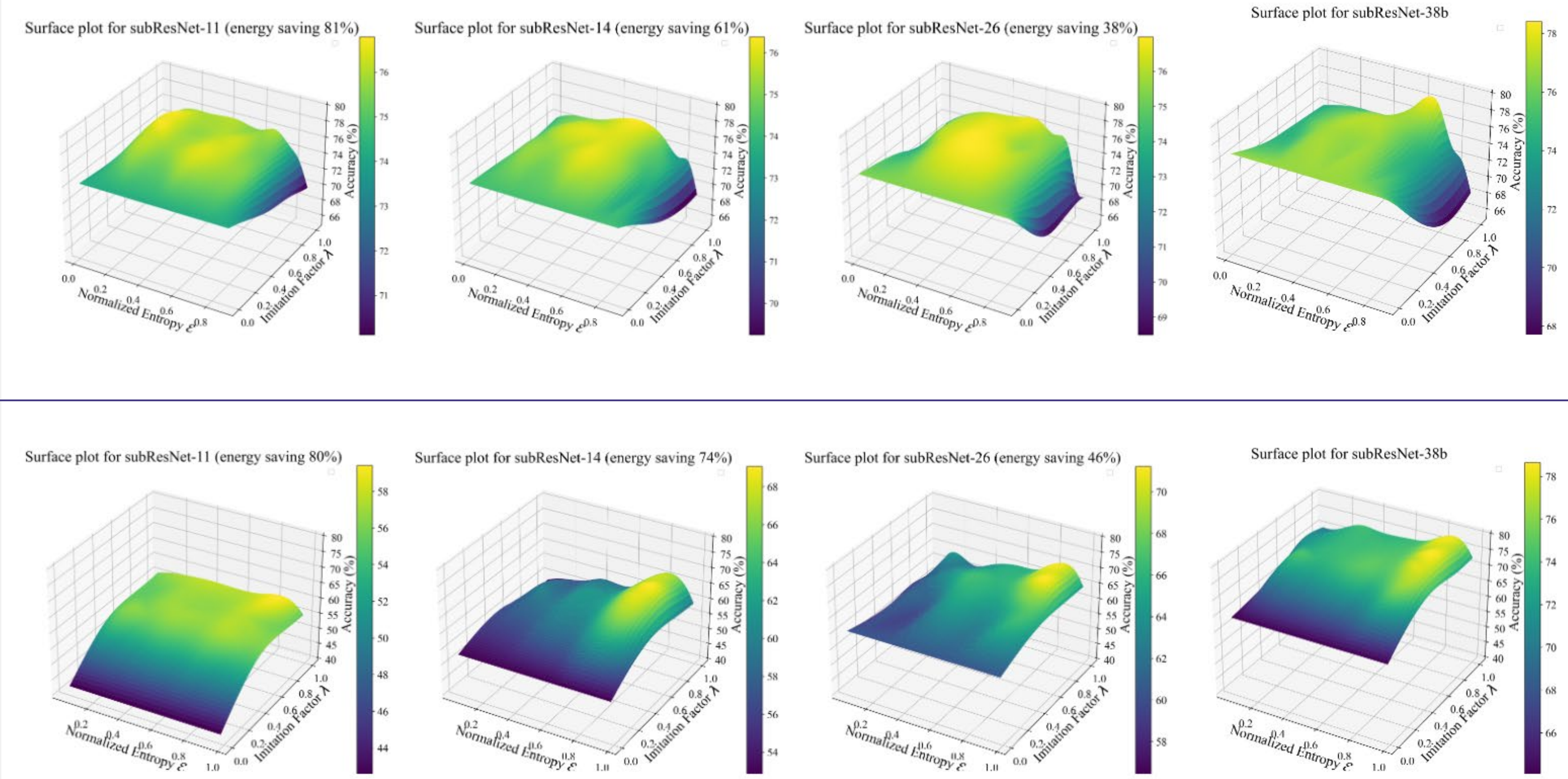
A workflow for designing low-carbon ML models,
leveraging knowledge distillation



Results

CIFAR-10

CIFAR-100



Increasing number of layers
(decreasing MCG)

Discussion

- Energy vs Accuracy trade-off (from ResNet-50 to ResNet-14 on CIFAR-100):
  - Energy saving: 74%.
  - Accuracy reduction: 19%.
  - Accuracy improvement (w.r.t ResNet-14 from scratch): 24%.
- Reduction of training time compared to traditional grid search: 80%.

Conclusions

- This thesis tackled key challenges associated with the employment of deep learning models within smart city environments, with a particular focus on privacy, interpretability, and energy efficiency.
- The goal was to create distributed and privacy-preserving AI ecosystems for smart cities.
- To this purpose three contributions were proposed, positioned in the field of Federated Learning (FL) and Edge AI.

Thank you for your attention

Contact:
franca.roccoditorrepadula@unina.it
Room 4.03 – building 3/A – via Claudio 21