MISSIONE 4
ISTRUZIONE RICERCA

Future Artificial Intelligence Research

SPOKE ACTIVITIES

SPOKE:
RESILIENT AI
3

AFFILIATES:
CRN
UNICAMPUS BIOMEDICO

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Summary of the overall scientific contributions of the Spoke

WP 3.1: Creation and Annotation of massive datasets

Objectives: In this WP the focus will be on the creation of samples and labels for extremely large datasets. The focus will be on the massive acquisition of partially labeled data, using generative models for data imputation, label imputation, and missing modality imputation. Also, we aim at extracting a high-level semantic description of the data in natural language. All these generated data and annotations can be used by subsequent learning-based techniques to improve their performance.

Current Results: During the reporting period, a general-purpose dataset such as ImageNet was identified as the glue between the specific datasets. In particular, two areas of interest were identified, namely the biological field and the autonomous driving field, and different domain datasets were identified for the two application domains identified. The objective of the research was then to align these datasets through domain ontologies using semantification techniques on the instances of the identified datasets.
In order to make all data conform to an ontological standard, it was decided to select domain ontologies such as Toyota Ontology and MONDO Disease Ontology for the semantic process, extending their semantic capabilities by introducing ad-hoc classes for the semantification of the selected datasets.
To co-ordinate the inter-domain process, ImageNet and WordNet were chosen as interconnected graphs for extending the classes in the domain ontologies, creating a bridge between the various domains by exploiting a top-level ontology.
To increase ImageNet's capacity, an ImageNet graph extension approach was defined, creating an enriched and more populated version of the knowledge base, called ImageNet++, populating nodes of the graph with innovative generative techniques.
Aside from the data curation process, different research topics were addressed involving important features of the data, namely their privacy and the presence of bias. For what concerns privacy, the research focused on identifying those features of the datasets that can make the enforcement of differential privacy easier. The activity involved two clustering approaches, the former based on Gaussian mixtures builds a soft clustering solution, and the k-means, which returns a partition of the data set. The results of this analysis were presented at the Ital-IA 2024 Workshop “FAIR TP7—Data Centric AI and Infrastructures” with the talk “The Advantage of Size: Privacy-Preserving Clustering in Large Dataset.”
The fairness of data was investigated in relation to the presence of bias in pretrained large language models. The goal was to identify, if possible, where the bias resides using ablation techniques to get some understanding of which portions of the Transformers were involved. We started with a Large BERT pretrained model, in which we attempted to explore the presence of gender bias both in the whole attention heads and in smaller portions of the weights of the layers.

WP 3.2: Resilient AI in adversarial environments

Objectives: This WP aims at addressing AI resiliency in adversarial scenarios from different points of view, towards the design of approaches and methodologies intended to i) detect and recover from attacks, ii) increase the robustness of federated learning, iii) enforce privacy, iv) enforcing fairness. Moreover, in the knowledge representation area, we will develop inference-proof countermeasures against attacks to knowledge confidentiality, based on various kinds of background knowledge and meta-knowledge.

Current Results: Environmental problems are increasingly recognized, and technology is evolving to find suitable solutions. The ancestral technique of crop rotation has been identified as a solution to address the pollution problems

caused by intensive food production (i.e., using fertilizers and pesticides). To ensure this technique can genuinely improve food production, it is crucial to understand how modern technologies, particularly adversarial machine learning, can support it. Analyzing crop rotation with adversarial machine learning can assist farmers in the decision-making process and optimize farm management practices by providing robust models that account for potential adversarial threats. The aim of our research is to investigate how predictive process monitoring techniques, enhanced by adversarial machine learning, can improve crop rotation strategies by leveraging Agriculture 4.0 through real-time monitoring. This approach aims to develop more accurate and adaptive strategies that are resilient to adversarial attacks. Our work proposes research questions for further study, which may help advance this research area. Moreover, we highlight the potential of adversarial machine learning in improving the robustness and effectiveness of crop rotation strategies within the context of modern agriculture. This activity is summarized in a publication (Simona Fioretto, Dino Ienco, Roberto Interdonato, Elio Masciari: Integrating Predictive Process Monitoring Techniques in Smart Agriculture. ISMIS 2024: 306-313) and will be the basis for the research in next period.
In the last months, research has been also focused on two lines related to static knowledge confidentiality enforcement. The first line is about the anonymization of knowledge graphs. We have defined an exact analogue of k-anonymity for first-order RDF graphs and studied the computational complexity of three types of anonymization procedure characterized by increasing information loss.
The second line of research focusses on module extraction from OWL2 knowledge bases. One of the applications of module extraction is removing all the knowledge about a given set of predicates from a knowledge base (by means of so-called "depleting" modules).
"Locality based" module extraction techniques are very effective (they yield small modules) and fast on terminological knowledge; less effective on knowledge bases whose ABox (i.e., roughly speaking, the explicit knowledge graph of the knowledge base) is nonempty. We have introduced a novel method for improving the performance of all module extractors (not only locality based) over consistent knowledge bases with nonempty ABoxes. Then, we set up and run extensive experiments to assess the new module extraction method.

WP 3.3: Resilient multi-task learning on the edge from incomplete and/or noisy data

Objectives: The goal of this WP is the definition of novel learning procedures for multi-task neural networks applicable to incomplete (missing labels) or noisy (soft or wrong labels) data. We will also investigate training methods for dealing with unbalanced data and self-paced learning approaches. Finally, we will explore optimization techniques aimed at making the multi-task neural networks usable in real-time on the edge.

Current Results: During the reported period, Work Package 3.3 focused on developing innovative Deep Learning techniques aimed at increasing the resilience of AI systems, particularly with noisy/missing and unbalanced data and in federated learning environments.
In the context of noisy and unbalanced data, we conducted an experimental campaign over the last six months to collect thermal udder images and their respective biological factors (like SCC, milk production, etc.), where, at times, some data for these features were missing. We processed those thermal images employing the neural network we developed during the WP and extracted the udder maximum temperature. We observed that the maximum temperature is one of the key predictors of mastitis condition, for which we analyzed the evolution of the temperature throughout each buffalo milking session. We release the results and the dataset as current results for that, which can be found at the following GitHub repository: https://github.com/Uzarel/udder-seg-pytorch.
In the context of Federated Learning (FL), recent advancements have tackled the significant challenge of non-IID (non-independent and identically distributed) data, which can cause divergence and poor generalization in global models. Building on previous work, our recent efforts have introduced the KAFÈ (Kernel Aggregation for Federated Learning) framework. This novel approach addresses non-IID data by leveraging Kernel Density Estimation (KDE) to aggregate classification layers, mitigating biases inherent in local updates. Through KDE, KAFÈ constructs a probability density function for the classification layers of client models, generating a new global classification layer that better captures diverse decision boundaries. Our experiments on image and text datasets have shown KAFÈ to consistently outperform

state-of-the-art FL methods, achieving higher accuracy with reduced computational overhead. This innovative aggregation strategy ensures more resilient multi-task learning on the edge, handling incomplete and noisy data more effectively. The source code is available here: https://github.com/MODAL-UNINA/KAFE.

WP 3.4: Enhanced Resilient AI through novel data pre-processing and human-interaction techniques

Objectives: The goal of this WP is the investigation of novel data pre-processing techniques for the enhancement of the machine learning process and the investigation of both natural interaction paradigms and intelligent human-interaction techniques to enhance user experience.

Current Results: Activities were targeted at: designing and developing a user interface for a module completely dedicated to the analysis and selection of data for training in view of CNN classification. This would mean that workflows for researchers and clinicians are optimized, thereby leading to effective curation and selection for relevant datasets, which is beneficial for the better fine-tuning of the CNN and will really make a noticeable difference in model accuracy and efficiency. Specifically, deep learning approaches were utilized in the analysis of 12-lead ECG data for SARS-CoV-2 infection identification, toward the development of a new fast, noninvasive, and reliable tool for COVID-19 diagnosis. In addition, a protocol for designing and implementing parallelization of entropy-based clustering algorithms was one of the significant works. This protocol allowed many nodes to run unsupervised learning jointly, enabling scalability and efficiency improvements in massive research scenarios with great protection of data privacy. Moreover, the extension of entropy-based clustering into adaptive label editing was applied for improving the classification performance of CNNs applied to histopathological medical images. It also outlined the application of diffusion probabilistic models such as DALL-E 2, Imagen, and Stable Diffusion to create high-fidelity medical imaging data. This work was majorly focused on Digital Breast Tomosynthesis (DBT) images for enriching the processes of training and validation of deep learning models for medical diagnostics. All these activities collectively will be aiding in the advancement of research capabilities and contributions toward medical imaging and diagnostics using novel AI methodologies.
During the reporting period, the legibility and the impact of robot's actions were investigated through a series of studies examining the role of explanations and their connection to the human user's beliefs. Specifically, the studies conducted in this Task focused on:

- Robot's persuasion and deception to increase human's trust in robots, submitted to IEEE International Conference on Robot and Human Interactive Communication.
- The assessment of distraction and disengagement caused by robots while monitoring older adults, submitted to IEEE Transaction on Affective Computing.
- The identification of physical and operational failures during interactions for robot interacting with humans, submitted to a challenge at the ACM International Conference on Multimodal Interaction.
- Currently, investigating robotic architectures to model first order and second order theory of mind, taking into account multi-agent tasks and methodologies

Thus, the progress of the research was characterized by means of neuro-evolution techniques applied to embodied agents situated in an external environment. In this case, we were interested to study the rules underlying the emergence of flocking behavior in a swarm of embodied agents and effects of the environments on the agent's behavior. More recently, we are investigating how to ground low-level features of large language models into external embodied perception derived from an agent perceptive field. Moreover, we are using large language models to retrieve and analyze structures and similarities among psychological test items.

WP 3.5: Resilient multimodal systems

Objectives: The goal of this WP is to improve the resilience of multimodal systems by addressing, both in the training phase and in the inference phase, the incompleteness, inconsistency, and great heterogeneity of the data. We will also

develop techniques for improving the robustness to adversarial attacks as well as the fairness with respect to possible multimodal biases.

Current Results: During the reporting period, the activities have been directed towards several goals, now described.

To advance multimodal systems in medical imaging, and to ensure resilience to missing modalities or in case where modalities are hard to collect for technical and medical reasons (e.g., shortage of resources, patient's medical situation that does not allow to carry out the scan, etc), our team focused on developing a virtual contrast enhancement technique using deep generative models for Contrast Enhanced Spectral Mammography (CESM). Employing an array of models like an autoencoder, Pix2Pix, and CycleGAN, we targeted the generation of synthetic recombined images from low-energy scans, where CycleGAN was highlighted as the most effective, supported by quantitative analysis and radiologist assessments of 1138 images. This methodology underscores a significant reduction in adverse effects and radiation exposure.

Still focusing on missing data, and how to make AI model resilient by design to such a case, we have directed our research efforts to engineer a transformer-based model specifically designed for handling missing data, without the need for imputation. We first investigated its use for AI-driven lung cancer prognosis. This model effectively manages both censored and uncensored data without the need for prior data imputation, setting a new benchmark in the accuracy of overall survival predictions. Our innovative approach, which excels beyond current state-of-the-art models, emphasises the model's capacity for extending to broader multimodal prognostic systems. We are now working not only on testing the model performance on several benchmark datasets but also on extending its definition and its learning strategy to cope with its application to a broader range of situations.

We also worked on multi-task training, using as a case study the resilience to COVID-19 pandemic. In this context, our strategy involved the development of a multi-dataset multi-task training framework to enhance disease outcome predictions from chest X-ray images. Utilizing a deep convolutional network architecture that accommodates multiple datasets and branches into task-specific outputs, this framework amplifies the robustness and predictive accuracy. The multi-task model was trained with a weighted loss function that dynamically adjusted the importance of each task based on the training progress, further improving performance. The versatility and effectiveness of this framework were validated across diverse neural network architectures and evaluation metrics, illustrating its broad applicability in clinical settings.

Furthering our knowledge in multimodal deep learning, we have continued our systematic review of intermediate fusion methods, particularly focusing on biomedical applications. Our analysis now encompasses a complete review of 54 papers, enriching our understanding of how multimodal data integration can enhance AI learning outcomes and application efficacy.

In the field of multimodal deep learning, our research focused on efficient integration methods, especially in scenarios with missing modalities, leading to incomplete data acquisitions. Specifically, in the task of assessing dementia severity, we added the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which includes clinical information and images from approximately 2000 patients. We evaluated a proposed Multi Input-Multi Output convolutional neural network designed to analyze both Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) data. This network employs an efficient training strategy capable of handling incomplete acquisitions by adapting training iterations based on the availability of specific image modalities. Our experiments on the ADNI dataset demonstrated that our network outperformed approaches using single image modalities and alternative multimodal deep learning fusion techniques. To further improve the integration of heterogeneous modalities, we introduced a Transfer Module (TM) within the network architecture. This module, placed between layers dedicated to different modality-specific paths, performs cross-modality calibration of extracted features, effectively mitigating the impact of less discriminative features while enhancing integration across multiple sources. We conducted additional experiments on both the OASIS-3 and ADNI datasets to evaluate the generalization ability of the 3D Multi Input-Multi Output convolutional neural network with the TM. The results confirmed the efficacy of our approach, demonstrating robust performance across different datasets and further validating the benefits of our intermediate fusion strategy.

In the context of multimodal information extraction, we continued the experimentation of a transformer-based Visual Question Answering (VQA) pipeline across different configurations, according to a dual-stream architecture, evaluating the impact of various parameters on performance, training duration and memory footprint. Moreover, we also implemented and tested another VQA dual-stream architecture using a pre-training phase based on CLIP (Contrastive

Language-Image Pre-Training), in comparison with some VQA single-stream architecture based on multimodal vision-language models such as ViLT and BLIP, adopting various general purpose datasets, such as DAQUAR dataset, and some datasets specific for the medical domain (such as the SLAKE dataset, and the VQA-RAD dataset).
Furthermore, we continued with the experimentation of two different approaches of Quantum Natural Language Processing on actual quantum hardware (NISQ machines). The first approach, based on Quantum Transfer Learning, was evaluated for one syntactic and one semantic classification task, i.e. acceptability judgement and sentiment analysis, in both English and Italian, where the representation power of classical language models was combined with the classification power of a quantum circuit. About the second approach, which was quantum native and based on the Lambek categorial grammar and Distributional-Compositional Framework, we are working on the realisation of a full learning pipeline for the Italian Language (tailored for acceptability judgement) and in particular, as a first component, on the setup of an Italian parser trained on real corpora.
Moreover, our team has continued the analysis of energy-based models to prove their effectiveness on image and natural language processing tasks, with a focus on dense associative memories. The effort was centred on theoretically understanding the difference between classical and dreaming dynamics, using Monte Carlo simulations. At the same time, we are exploring the potential of Kolmogorov-Arnold networks (KAN) to further modify the energy transformer and obtain a lighter model in terms of parameters when applied to multimodal tasks.
In the context of multimodal systems, we started a research activity focused on developing and analysing algorithms for activity monitoring and recognition using computer vision and pattern recognition techniques. In particular, we experimented an application of these approaches in the medical domain, annotating a dataset comprising various therapeutic gesture activities and training on it algorithms for automatic evaluation of activity correctness. Additionally, biases related to human physical and health characteristics were analysed during the training of various state-of-the-art deep neural network architectures. Moreover, we also investigated multi-channel time series models through a multiscale feature learning approach. We experimented with traditional multi-channel time series algorithms with their multiscale counterparts' architectures, designed to independently extract features at different scale resolutions, operating on an open dataset including both healthy individuals and patients with various disabilities.
Finally, research activities regarding the setup of a multimodal acquisition system for autonomous vehicle driving continued. In particular, they regarded the design of an acquisition campaign using the multimodal sensor set equipped in an off-road autonomous rover. The rover was configured to be remote-controlled during data acquisition. Different off-road locations were identified, including environments like rural areas and gardens. Moreover, the hardware setup to support continuous acquisition on SSD drive was completed. After that, the first data were acquired using this setup.

WP 3.6: Automated Support for Resilient, Dependable, and Interpretable AI

Objectives: The goal of this Work Package is to conceive software engineering techniques aimed at supporting the development of AI-intensive software systems, providing quality attributes such as resilience, dependability, and interpretability. To this aim, the WP will foresee techniques aimed at supporting developers by automatically providing code templates for ML tools; performing quality assessment of ML-intensive systems, also in a highly-distributed context, and performing verification and validation of ML-intensive systems. Last, but not least, the WP will provide approaches for explainability and interpretability support.

Current Results: An ante-hoc interpretability method for deep learning, i.e., the Latent Diffusion Model (LDM), has been investigated for the super-resolution of medical images, namely MRIs. Results are good; a problem has emerged with the evaluation metrics commonly used in literature, i.e., the Peak-Signal-to-Noise Ratio (PSNR) and the Structural SIMilarity index (SSIM) metrics. Therefore, a study to overcome these problems has started.
Furthermore, we worked on a new approach to Federated Learning that leverages Neuroevolution running on the clients to optimize the structure of Artificial Neural Networks through Grammatical Evolution. The structures of the obtained networks are very small, so they provide easy-to-understand explainability in terms of a clear relationship between the inputs and the outputs.

Regarding the resiliency of the XAI technique Shadow Clustering, activities have been carried out to improve the approach's robustness to small random perturbations of sample values. This has led to alternative measures for evaluating rule quality that can better face noise.
Trustworthy AI has been investigated in terms of explainability and safety. This raises the need for performance guarantees and interpretability. Conformal prediction (CP) allows the generation of prediction sets in the label space with predefined probabilistic guarantees for any machine learning model. The definition of a score function with higher values for misclassifications and lower for good performance is crucial for CP.
A specific score function suitable for applying CP to rule-based classifiers of if-then type, considered the most interpretable XAI method, has been proposed and implemented.
From the other side, in the same period we continued the research activities on verification techniques for autonomous vehicles. We focused from one hand on machine learning model robustness for ADAS applications, particularly in image analysis. The methodology adopts quality assessment through error analysis and performance auditing. To enhance overall robustness, we are developing a novel framework with an adversarial detector and an input distillation model for improved image quality and classification accuracy, and some preliminary experiments have been carried out. From the other one, contingency plans involving extra training with new data or architectural modifications have been performed, aiming for a resilient framework adaptable to various scenarios in the dynamic ADAS environment.
We are also working on simulation-based testing approaches for autonomous vehicles. In this research line we continued experimenting with the selected simulation platforms (i.e., CARLA and BeamNG) to develop a customizable pipeline prototype capable of executing simulation-based test scenarios in both platforms and collecting data from a large array of virtual sensors, which could prove useful also in the definition of new datasets for autonomous vehicles. Research activities also included the definition of novel test generation strategies, including both AI-based approaches such as Reinforcement Learning and search-based approaches, such as genetic algorithms, to generate simulation-based test scenarios to run using the aforementioned pipeline.

WP 3.7: Resilient Strategic Reasoning in AI

Objectives: In this WP we will investigate resilience in the context of AI reactive systems that autonomously take decisions and program themselves to act strategically in a nondeterministic partially known environment, while being resilient to unexpected changes in it. We will make use of formal methods for the strategic reasoning and synthesis of best effort strategies.

Current Results: IW have continued the study aiming at extending formal aspects of strategic reasoning, especially in the context of multi-agent systems (MAS) to incorporate strong forms of resilience, both for finite and infinite-state models.
On the finite-state case, we have investigated the application of logics for the strategic reasoning, such as ATL and Strategy Logic, in the setting of Stochastic Environments, with full and partial information among the players, both in the qualitative and qualitative setting. Also, we have investigated along these settings the application of the concept of natural strategies. We have also investigated the connection of Strategy logic with GDL-II as well as its use in the incentive design for rational agents.
On the topic of infinite-state real-time systems, we are investigating a wider range of models where it may be possible to perform automatic model-checking of linear temporal properties. We also initiated contact with Collins Aerospace to discuss potential collaborations.
Regarding stochastic models of sequential decision making, we started an investigation on advanced solution concepts for Markov decision processes, with the aim of identifying best-effort policies.

WP 3.8: Ethical, Legal and Societal issues in resilient AI systems

Objectives: The objectives of this WP are the coexistence of different regulatory systems; the division of tasks between European existing Authorities and the future one in a multi-level order; the resilience of rules and Authorities. The above objectives are intertwined in such a way as to overturn the traditional systems of competencies' distribution among several Authorities. The objectives necessarily become a macro-objective, consisting of hybridization of principles, confusion of rules, and intersection in the fields of fundamental rights and Authorities. The latter will have to dialogue in order to achieve an ambitious result: to deliver to Europe a system of AI regulation able to resist future crises. Finally, the WP will also support the process of verifying whether a given information/knowledge system complies with the applicable regulatory framework – such as the GDPR, for example.

Current Results: The three research axes (constitutional, civil and labour law) have been continuing their activities in the considered reporting period.
The constitutional axis continued by delving into five specific aspects.

1) Governance of Artificial Intelligence Systems: This aspect refers to both national and supranational frameworks outlined in Articles 70 ff. and 64 ff. of the Proposal for a Regulation on Artificial Intelligence, COM/2021/206 final (AI Act). The governance structure could have taken two forms: full decentralization or fully centralized supervision. The Commission opted for an intermediate approach. The European rule-maker designates Agencies with functional independence (though not genetic one) rather than independent authorities to oversee AI systems. This approach reflects a government-centric strategy justified by AI's potential role as a driving force behind public policies, crucially remaining under current political majorities' control.
2) Regulation of Generative AI Systems: This involves interactions within the digital ecosystem and between three regulations: the Regulation on Artificial Intelligence, Regulation 2022/2065 Digital Services Act (DSA), and Regulation 2022/1925 Digital Markets Act (DMA). Addressing the goal of establishing a unified European digital ecosystem, the research explores how the DSA and DMA can apply to generative intelligences by analogical reasoning.
3) Insight into Obligations Introduced by the AI Act (Articles 51 et seq.): Analysis of generative AI under the AI Act reveals differing treatment compared to traditional AI. While both must comply with harmonized rules for AI model deployment, traditional AI faces stricter regulations. This strategic distinction aims to foster generative AI development, ensuring European competitiveness against U.S. and Chinese markets.
4) Prohibitions Introduced in the AI Act (Art. 5): Detailed investigation on prohibitions outlined in Article 5 of the AI Act, specifically addressing subliminal techniques, exploitation of vulnerabilities, and emotion capture. These prohibitions are analyzed in relation to new rights and their effectiveness in safeguarding citizens' digital *habeas mentem*.
5) Criterion of Digital Ecosystems: assess how the criterion of digital ecosystems has been replacing that of substitutability in current online markets, consequently analysing data governance in light of competition and human rights protection.

Additionally, the constitutional group collaborated with WP3.3: Resilient Multi-task Learning on the Edge from Incomplete and/or Noisy Data. This collaboration aimed to develop a generative AI prototype in response to the Italian Parliament's call for interest. The project aims to integrate recent generative AI methodologies into parliamentary processes, specifically facilitating the preparation of bills and enabling effective oversight by Members of Parliament on gender issues and empowerment. The prototype serves as a foundation for developing AI solutions within Parliament's activities: legislative initiative, policy-making and control of government activities.
Activities carried out.
The research in civil law investigates antinomies and resolution criteria in the circulation of personal data, especially health data, also considering the impact of AI. Emerged lines of research concern the identification of the legal basis for the processing of electronic health data - as a prerequisite for the implementation of AI in this field - articulating the question regarding both primary and secondary use. Having outlined the existing legal framework for the secondary use

of health data and the existing antinomies, research questions are asked about legitimate use in proactive medicine by verifying the necessity and nature of explicit consent of the data subject.
The labor law axis has continued to study the impact of AI in the workplace, with particular attention to the European AI regulation and the bill presented to the Senate (A.S. 1146 - Disposizioni e deleghe al Governo in materia di intelligenza artificiale). These documents have also been examined regarding the impact on the University's missions. During this semester, the research also focused on the study of generative AI, taking part in the Privacy Tour 2024 organized by the Italian Data Protection Authority. The issue of metadata preservation was also addressed, examining the topic from the perspective of the AI Act, the GDPR, and the Digital Market Act. Furthermore, were also examined the topics of the relationship between employer liability and AI systems and the relationship between fundamental rights and the risk-based approach of the AI Act.