

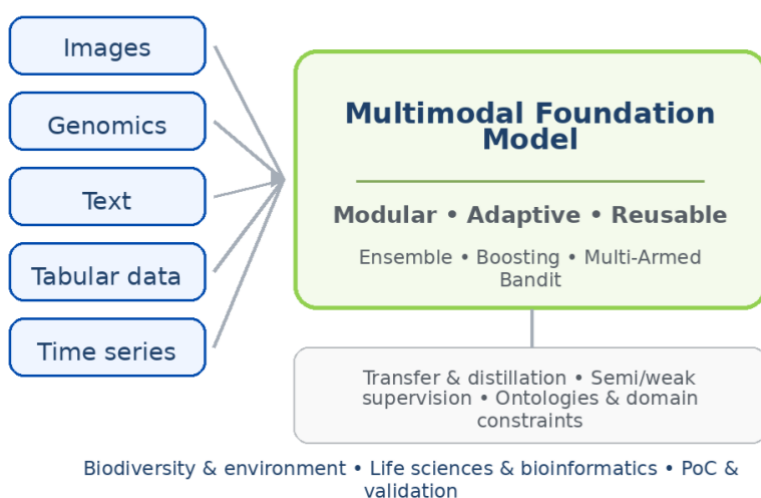


RISE

Infrastruttura Strategica della Ricerca
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Servizio integrato per modelli fondazionali multimodali per dati scientifici, biologici e ambientali

MULTIMODAL SCIENTIFIC DATA



Architettura concettuale del servizio AI4NATURE

Per informazioni rivolgersi al responsabile scientifico:

Prof. Michelangelo Ceci
michelangelo.ceci@uniba.it

Gruppo di ricerca KDDE

Polo Scientifico Tecnologico
Dipartimento di Informatica
Via Orabona 4, 70125 (BA)

ENG: Within the PN-RIC 2021-2027 project AI4NATURE (Artificial Intelligence for Environmental Risk and Biodiversity Restoration), the University of Bari proposes an integrated service for the co-design, development, adaptation, and validation of multimodal foundation models for complex scientific data. The service builds on the activities of the UNIBA work package “The AI4Nature Foundation Model” and on the expertise of the KDDE research group of the Department of Computer Science, coordinated by Prof. Michelangelo Ceci.

The service is designed to integrate heterogeneous information sources, including images, genomic signals, textual and tabular data, and time series, within modular, adaptive, and reusable AI architectures. Specialist models for different modalities can be combined through ensemble learning and selected or weighted adaptively using boosting and Multi-Armed Bandit techniques [Simeon et al., 2024]. This architecture supports the addition, update, or replacement of specialised components while reducing the need for complete retraining.

To operate effectively in scientific domains where labelled data are scarce, the service can employ semi-supervised and weakly supervised learning pipelines [Serafino et al., 2018; Pio et al., 2018], together with transfer learning [Mignone et al., 2024], knowledge distillation, and selective reuse of pre-trained components. Physical constraints, ontologies, taxonomies, scientific knowledge, and

ecological rules can be incorporated into the learning process to improve robustness, interpretability, and consistency with domain knowledge.

The service supports feasibility analysis and co-design of multimodal AI applications [Buttaro et al., 2025]; design and adaptation of modular AI architectures; integration of specialist models and multiple data modalities; development of learning solutions for sparsely labelled datasets; integration of domain knowledge and scientific constraints; and implementation and validation of proof of concept on selected use cases. Potential beneficiaries include universities, research organisations and laboratories in life sciences, genomics and bioinformatics; companies developing AI and advanced data-analysis solutions; and public administrations, agencies, and organisations involved in biodiversity and environmental monitoring.

The objective is to help public and private stakeholders exploit heterogeneous scientific data, reduce annotation and training costs, and develop models that can be adapted and reused across different application contexts. The service is consistent with the RISE Scientific and Technological, Bio-Health, and Agr-Vet thematic poles, with particular relevance to biodiversity and environmental monitoring, genomics, life sciences, and biological data.

ITA: Nell'ambito del progetto PN-RIC 2021-2027 AI4NATURE (Intelligenza Artificiale per il Rischio Ambientale e il Restauro della Biodiversità), l'Università di Bari propone un servizio integrato per il co-design, lo sviluppo, l'adattamento e la validazione di modelli fondazionali multimodali per l'analisi di dati scientifici complessi. Il servizio valorizza le attività del work package di competenza UNIBA "The AI4Nature Foundation Model" e le competenze del gruppo di ricerca KDDE del Dipartimento di Informatica, coordinato dal Prof. Michelangelo Ceci.

Il servizio è concepito per integrare fonti informative eterogenee, tra cui immagini, segnali genomici, dati testuali e tabellari e serie temporali, all'interno di architetture di IA modulari, adattive e riusabili. Modelli specialistici relativi a differenti modalità possono essere combinati mediante tecniche di ensemble learning e selezionati o pesati in modo adattivo attraverso boosting e Multi-Armed Bandit [Simeon et al., 2024]. Tale impostazione consente di aggiungere, aggiornare o sostituire componenti specialistiche riducendo la necessità di un riaddestramento completo.

Per operare in domini scientifici caratterizzati da una disponibilità limitata di dati etichettati, il servizio può adottare pipeline di apprendimento semi-supervisionato e debolmente supervisionato [Serafino et al., 2018; Pio et al., 2018], insieme a tecniche di transfer learning [Mignone et al., 2024], knowledge distillation e riuso selettivo di componenti già addestrati. Vincoli fisici, ontologie, tassonomie, conoscenze scientifiche e regole ecologiche possono essere integrate nei processi di apprendimento per favorire robustezza, interpretabilità e coerenza con le conoscenze di dominio.

Il servizio supporta l'analisi di fattibilità e il co-design di applicazioni basate su IA multimodale [Buttaro et al., 2025]; la progettazione e l'adattamento di architetture AI modulari; l'integrazione di modelli specialistici e di differenti modalità informative; lo sviluppo di soluzioni di apprendimento in presenza di dati scarsamente etichettati; l'integrazione di conoscenze di dominio e vincoli scientifici; la realizzazione e la validazione di proof of concept su casi d'uso selezionati. I potenziali beneficiari

comprendono università, enti di ricerca e laboratori attivi nelle scienze della vita, nella genomica e nella bioinformatica; imprese che sviluppano soluzioni di IA e servizi avanzati di analisi dei dati; Pubbliche Amministrazioni, agenzie ed enti impegnati nel monitoraggio della biodiversità e dell'ambiente.

L'obiettivo è supportare stakeholder pubblici e privati nella valorizzazione di dati scientifici eterogenei, nella riduzione dei costi di annotazione e addestramento e nello sviluppo di modelli adattabili e riusabili in differenti contesti applicativi. Il servizio è coerente con i Poli Tematici RISE Scientifico e Tecnologico, Bio-Health e Agr-Vet, con particolare riferimento al monitoraggio ambientale e della biodiversità, alla genomica, alle scienze della vita e ai dati biologici.

References

- 1) Simeon, M. Radovanović, T. Lončar-Turukalo, M. Ceci, S. Brdar, G. Pio, Multi-class boosting for the analysis of multiple incomplete views on microbiome data, BMC Bioinformatics, 25, 18, 2024
- 2) F. Serafino, G. Pio, M. Ceci, Ensemble Learning for Multi-type Classification in Heterogeneous Networks, IEEE Transactions on Knowledge and Data Engineering, 30(12): 2326-2339, 2018
- 3) G. Pio, F. Serafino, D. Malerba, M. Ceci, Multi-type clustering and classification from heterogeneous networks, Information Sciences, 425:107-126, 2018
- 4) P. Mignone, G. Pio, M. Ceci, Distributed Heterogeneous Transfer Learning, Big Data Research, 37, 100456, 2024
- 5) V. Buttarò, G. Lamanna, D. Massaro, C. B. Caporusso, G. Pio, M. Ceci, Novel AI Approach for the Diagnosis of Alzheimer's Disease from Multi-modal Incomplete Data, Discovery Science 2025