Next-Gen Health: from Multimodal AI to Foundation Models

Rosa Sicilia[1,*], Fatih Aksu[1,2], Alessandro Bria[3], Alice Natalina Caragliano[1], Camillo Maria Caruso[1], Ermanno Cordelli[1,4], Arianna Francesconi[1], Valerio Guarrasi[1], Giulio Iannello[1], Guido Manni[1,5], Massimiliano Mantegna[1], Giustino Marino[1], Daniele Molino[1], Elena Mulero Ayllón[1], Filippo Ruffini[1], Linlin Shen[6], Matteo Tortora[7] and Paolo Soda[1,8]

[1] *Unit of Artificial Intelligence and Computer Systems, Department of Engineering, University Campus Bio-Medico of Rome, Italy*
[2] *Department of Biomedical Sciences, Humanitas University, Milan, Italy*
[3] *Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Cassino, Italy*
[4] *Department of Experimental Medicine, Università del Salento, Italy*
[5] *Unit of Advanced Robotics and Human-Centered Technologies, Department of Engineering, University Campus Bio-Medico of Rome, Italy*
[6] *College of Computer Science and Software Engineering, Shenzhen University, China*
[7] *Department of Naval, Electrical, Electronics and Telecommunications Engineering, University of Genoa, Genoa, Italy*
[8] *Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umeå University, Sweden*

Abstract

Artificial intelligence is reshaping every aspect of health and well-being—from early prevention to day-to-day self-care. Our research advances this shift on three cutting-edge fronts: multimodal AI, resilient AI, and foundation models. By blending diverse data streams with resilient architectures, we bring forward the research in AI for health and well-being, trying to bridge the gap between cutting-edge computation and real-world health services, paving the way for next-generation AI that supports individuals, clinicians, and public-health systems alike.

Keywords

Artificial Intelligence, Medical Foundation Models, Multimodal Learning, Resilient AI, Stress Detection

1. Introduction

AI is emerging as a transformative force in medicine and human well-being [1, 2], offering the potential to make healthcare safer, faster, and more accessible. However, this shift brings challenges, particularly the reliance of AI on large, annotated datasets—resources that are often costly and time-intensive to acquire in healthcare. To address this, self-supervised pre-training on unlabeled data has gained traction, leading to the development of foundation models that can be adapted for diverse clinical tasks.

Our research advances this field through three interconnected directions: (i) Multimodal AI for health and well-being (section2); (ii) Resilient AI for health (section3); (iii) Foundation Models in Medicine(section4). Figure 1 illustrates the solutions explored in these areas, highlighting both current achievements and the open questions guiding our future research.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

r.sicilia@unicampus.it (R. Sicilia); Fatih.Aksu@unicampus.it (F. Aksu); a.bria@unicas.it (A. Bria); a.caragliano@unicampus.it (A. N. Caragliano); CamilloMaria.Caruso@unicampus.it (C. M. Caruso); e.cordelli@unicampus.it (E. Cordelli); Arianna.Francesconi@unicampus.it (A. Francesconi); valerio.guarrasi@unicampus.it (V. Guarrasi); G.Iannello@unicampus.it (G. Iannello); Guido.Manni@unicampus.it (G. Manni); M.Mantegna@unicampus.it (M. Mantegna); Giustino.Marino@unicampus.it (G. Marino); daniele.molino@unicampus.it (D. Molino); e.muleroayllon@unicampus.it (E. Mulero Ayllón); Filippo.Ruffini@unicampus.it (F. Ruffini); llshen@szu.edu.cn (L. Shen); matteo.tortora@unige.it (M. Tortora); p.soda@unicampus.it (P. Soda)

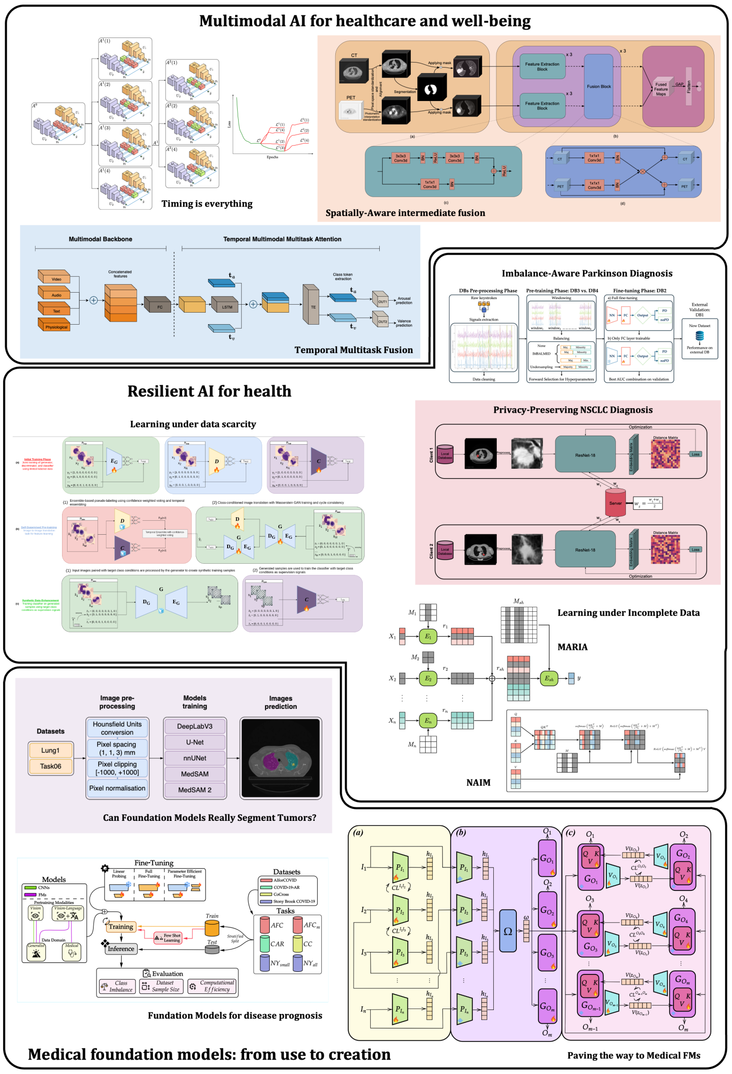
Multimodal AI for healthcare and well-being
Timing is everything
Spatially-Aware intermediate fusion
Multimodal Backbone
Temporal Multimodal Multitask Attention
Temporal Multitask Fusion
Imbalance-Aware Parkinson Diagnosis
Resilient AI for health
Learning under data scarcity
Privacy-Preserving NSCLC Diagnosis
Learning under Incomplete Data
MARIA
NAIM
Datasets
Image pre-processing
Models training
Images prediction
Lung1
Task06
Hounsfield Units conversion
Pixel spacing (1, 1, 3) mm
Pixel clipping [-1000, +1000]
Pixel normalisation
DeepLabV3
U-Net
nnUNet
MedSAM
MedSAM 2
Can Foundation Models Really Segment Tumors?
Fine-Tuning
Models
Datasets
Tasks
Training
Inference
Evaluation
Fundation Models for disease prognosis
Medical foundation models: from use to creation
Paving the way to Medical FMs

Figure 1: Outline of the presented research activities.

2. Multimodal AI for healthcare and well-being

Our inherently multimodal world necessitates integrating diverse data streams to fully capture its complexity. Multimodal Deep Learning (MDL) excels at this by identifying patterns across heterogeneous data sources, offering richer representations that drive improved insights and decision-making [3]. In healthcare, such integration is vital for understanding patient health through imaging, clinical, genetic, and other data types.

Our systematic review [4] compiles intermediate-fusion strategies in biomedical MDL, introducing a unified notation and sharing the dataset via an open repository (https://github.com/cosbidev/Intermediate-Multimodal-Fusion-Bio). This review identifies key gaps in current approaches, such as the need for broader modality combinations, adaptive fusion strategies, improved explainability [5, 6, 7, 8], and robust handling of missing data and domain adaptation [9, 10, 11, 12].

Building on these insights, our research addresses: (i) When to fuse? Determining optimal fusion timing (section 2.1); (ii) Spatial information preservation: Retaining spatial cues in multimodal medical images (section 2.2); (iii) Stress detection: Using multimodal data to understand and detect stress (section 2.3).

2.1. Timing is everything

In [13], we explore the crucial but often neglected question of *when* to fuse modalities in multimodal deep learning for medical imaging. Most existing strategies (early, late, or intermediate fusion) or neural architecture searches are either rigid or computationally prohibitive. We propose a Sequential Forward Search Algorithm (SFSA) that incrementally tests fusion modules at different network layers, selecting the best configurations based on validation performance and stopping when gains stabilize. SFSA leverages pretrained weights for efficient design space exploration and integrates modality-specific subnetworks with Multimodal Transfer Module blocks, enhancing accuracy and efficiency. Applied to MRI-based tasks (focal cortical dysplasia detection and Alzheimer's diagnosis), SFSA surpasses unimodal, late-fusion, and brute-force baselines in accuracy, F-score, specificity, and AUC.

2.2. Spatially-Aware intermediate fusion

Conventional intermediate fusion techniques typically merge features from separate backbones into compressed vectors, losing spatial correlations [4]. However, imaging modalities like CT and PET are spatially aligned during acquisition. To leverage this, we introduce MINT [14], a Multi-stage INTermediate fusion method that iteratively fuses voxel-wise features early in the extraction process. This preserves spatial structure and allows fused features to inform unimodal branches. Evaluated on PET/CT data for NSCLC histological subtype classification, MINT outperforms unimodal, early, late, and the only existing intermediate-fusion baseline. It effectively handles class imbalance and demonstrates the benefits of spatially-aware multimodal fusion in tumor characterization.

2.3. Temporal Multitask Fusion

Stress is a critical factor in well-being, and unmanaged stress can lead to serious health issues. Multimodal AI enables real-time, proactive stress monitoring [15]. We introduce Temporal Multimodal Multitask Attention (TMMA), an architecture that estimates arousal and valence by combining video, audio, text, and physiological features. TMMA uses LSTM for temporal modeling and a Transformer Encoder to capture long-range dependencies, achieving superior performance (valence: 60.3%, arousal: 61.1%) on the MuSe-2022 ULM-TSST dataset while remaining lightweight (≈1.8M parameters) and fast (≈20ms/sample on CPU). To support future work, we collected a new multimodal dataset (Dec 2024) from 28 air-traffic-controller trainees, capturing video, audio, physiological signals, self-reported stress scores, and workload labels. This dataset enables further exploration of real-time stress monitoring using TMMA's efficient, multimodal multitask approach with potential for foundation model integration.

3. Resilient AI for health

AI systems in healthcare must remain robust despite data corruption, incompleteness, and changing environments—essential for reliable clinical deployment. Our research prioritizes: (i)coping with data scarcity(section 3.1), (ii)handling imbalance in typing-based Parkinson detection(section 3.2), (iii)privacy-preserving learning(section 3.3), and (iv)robustness to missing data and modalities(section 3.4).

3.1. Learning under Data Scarcity

Deep learning in medical imaging is hampered by the lack of labeled data due to privacy and annotation challenges [16, 17, 18]. We propose a GAN-based semi-supervised framework that performs well with as few as 5–10 labeled samples/class. It combines a generator ($\mathscr{G}$), a discriminator ($\mathscr{D}$), and a classifier ($\mathscr{C}$) in three phases: supervised joint training, self-supervised pre-training on unlabeled data with pseudo-labeling and translation tasks, and synthetic data enhancement. Evaluated on MedMNIST, our ensemble outperforms six state-of-the-art methods across 5–50 shot settings, achieving up to 80.64% accuracy, marking progress toward resilient, data-efficient AI in healthcare.

3.2. Imbalance-Aware Parkinson Diagnosis

Parkinson's disease (PD) remains hard to detect early. We introduce a digital biomarker based on keyboard dynamics (KD), analyzing keystroke timing during normal typing [19]. Our three-phase pipeline includes data preprocessing (IMBALMED [20]), pre-training of eight deep models with optimized hyperparameters, and fine-tuning with external validation. Hybrid architectures like GRU-FCN and LSTM-FCN achieve AUC-ROC $>90\%$ and F1-Scores $>80\%$. TCN peaks at 92.22% AUC-ROC, outperforming internal-only validation baselines. These results position KD as a scalable, non-invasive PD screening tool. Future directions include integrating multimodal sensors and enhancing model interpretability.

3.3. Privacy-Preserving NSCLC Diagnosis

NSCLC subtype classification faces data fragmentation and privacy challenges. We adopt a triplet-loss-based horizontal federated learning (HFL) framework [21] using two datasets: a private PET/CT cohort and the public NSCLC Radiomics dataset. Clients train modified ResNet-18 models (no classification layer, triplet loss) and aggregate weights centrally. Predictions are aggregated per patient via k-NN in feature space and majority voting. Federated learning achieves AUCs of 0.664 and 0.654, demonstrating the feasibility of collaborative, privacy-preserving subtype diagnosis.

3.4. Learning under Incomplete Data

Missing data is pervasive in clinical datasets, and imputation can mask important signals. We propose two transformer-based models treating missingness as informative:

NAIM [22, 23] embeds tabular data as tokens, removing missing entries from self-attention and gradients, and uses epoch-wise regularization. Benchmarked on five datasets, it outperforms 35 pipelines (Wilcoxon tests: 98.4% of cases) and retains strong AUROC even with 75% missing inputs.

MARIA [24] generalizes NAIM to multimodal data (demographics, labs, imaging biomarkers). Each modality is encoded via a NAIM module, then fused with masked-attention to aggregate available signals. Evaluated on AIforCovid and ADNI for eight diagnostic tasks, MARIA outperforms 32 baselines [25, 26] in 60% of tests under missing-modalities and 12% in all-missing regimes. These models show how treating missingness explicitly—rather than via imputation—enables robust, fair predictions in real-world clinical environments.

4. Medical foundation models: from use to creation

Foundation Models (FMs) hold transformative potential for healthcare. We present our work spanning two key areas: **(i)**applying existing foundation models to clinical tasks(section 4.1), and **(ii)**developing a new medical foundation model from scratch(section 4.2).

4.1. Foundation Models for Medical Images

Our investigations into FMs for medical images span segmentation, histological classification, and prognosis prediction:

Tumor Segmentation: We benchmarked MedSAM and MedSAM2—promptable foundation models—against U-Net, DeepLabV3, and nnUNet for lung tumor segmentation [27]. On NSCLC-Radiomics and Task06 datasets, traditional CNNs performed well on lung segmentation but failed on tumors (IoU $\approx$ 0.04, Dice $\approx$ 0.05). nnUNet achieved high tumor accuracy (IoU 0.84, Dice 0.90), while MedSAM2 surpassed all (IoU 0.86, Dice 0.91), offering both accuracy and computational efficiency. These results highlight MedSAM2's potential for clinical deployment, with future work focused on automating prompt generation.

Histological Image Analysis: We evaluated three general-purpose medical FMs—Merlin, CT-CLIP, and CDUM—for NSCLC subtype classification, comparing them to task-specific models MINT [14], DETECT-LC, and LUCY. Foundation models consistently outperformed task-specific baselines in balancing sensitivity and specificity, showing robustness and generalizability for real-world histological tasks.

Prognosis Prediction: We benchmarked CNNs and FMs on COVID-19 chest X-ray datasets for severity escalation, ICU admission, and mortality prediction, evaluating fine-tuning strategies (FFT, LP, PEFT including LoRA, BitFit, VeRA, IA^3). CNNs excelled in extreme low-data settings, while PEFT methods (notably LoRA) enabled stable FM adaptation with minimal updates, outperforming LP in few-shot regimes. These findings offer practical guidance for choosing models and strategies based on data and task constraints. Our open repository is at github.com/fruffini/PEFT_Prognosis.

4.2. Paving the way to Medical FMs

Traditional generative modeling in healthcare has focused on unimodal data [28, 29, 30]. MedCoDi-M [31] addresses this by introducing a multimodal foundation model for any-to-any translation across medical images, reports, biosignals, and structured data. MedCoDi-M leverages a shared latent representation learned through contrastive objectives, with modality-specific encoders feeding into this common space. It supports flexible generation across modalities—e.g., images from text or structured data from imaging—capturing the rich multimodality of clinical workflows. We validated MedCoDi-M on MIMIC-CXR, evaluating not only standard benchmarks for image/text generation but also the real-world clinical utility of synthetic data for: (i) Anonymization: Training solely on synthetic data and testing on real samples; (ii) Class imbalance: Oversampling underrepresented disease classes with synthetic images; (iii) Data scarcity: Augmenting small real datasets with synthetic data to boost performance. These experiments underscore the promise of foundation models to promote fairness, accessibility, and robustness in clinical AI pipelines.

Acknowledgments

Fatih Aksu, Alice Natalina Caragliano, Camillo Maria Caruso, Arianna Francesconi, Guido Manni, Massimiliano Mantegna, Giustino Marino, Daniele Molino and Filippo Ruffini are Ph.D. students enrolled in the National Ph.D. in Artificial Intelligence, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. We acknowledge financial support from: i) PNRR MUR project PE0000013-FAIR; ii) PRIN 2022 MUR 20228MZFAA-AIDA (CUP C53D23003620008); iii) Unione europea- Next Generation EU, Missione 4 Componente C2 - Investimento 1.1 “Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)” CUP di progetto (CUP C53D23009280001); iv) FCS MISE (CUP B89J23000580005); v) MAECI (grant n. CN23GR09, CUP C83C22001650001); vi) ECS 0000024 Rome Technopole - CUP C83C22000510001; vii) University Campus Bio-Medico di Roma under the program “University Strategic Projects” within the project “AI-powered Digital Twin for next-generation lung cancEr cAre (IDEA)”; viii) Kempe Foundation project JCSMK24-0094. This work was also partially supported by the following companies: Eustema S.p.A. and ENAV S.p.A..

Declaration on Generative AI

During the preparation of this work, the authors did not use any AI tool.

References

[1] V. Guarrasi, et al., Building an AI-enabled metaverse for intelligent healthcare: opportunities and challenges, in: CEUR WORKSHOP PROCEEDINGS, volume 3486, 2023, pp. 134–139.

[2] F. Aksu, et al., Towards AI-driven next generation personalized healthcare and well-being, in: 2024 Ital-IA Intelligenza Artificiale-Thematic Workshops, Ital-IA 2024, Naples, Italy, May 29-30, 2024, CEUR-WS, 2024, pp. 360–365.

[3] M. Tortora, et al., RadioPathomics: multimodal learning in non-small cell lung cancer for adaptive radiotherapy, IEEE Access 11 (2023) 47563–47578.

[4] V. Guarrasi, et al., A systematic review of intermediate fusion in multimodal deep learning for biomedical applications, Image and Vision Computing (2025) 105509.

[5] O. Coser, et al., Deep Learning for Human Locomotion Analysis in Lower-Limb Exoskeletons: A Comparative Study, Frontiers in Computer Science Volume 7 - 2025 (2025).

[6] V. Guarrasi, et al., Multimodal explainability via latent shift applied to COVID-19 stratification, Pattern Recognition 156 (2024) 110825.

[7] A. N. Caragliano, et al., Doctor-in-the-Loop: An explainable, multi-view deep learning framework for predicting pathological response in non-small cell lung cancer, Image and Vision Computing (2025) 105630.

[8] A. N. Caragliano, et al., Multimodal Doctor-in-the-Loop: A Clinically-Guided Explainable Framework for Predicting Pathological Response in Non-Small Cell Lung Cancer, arXiv preprint arXiv:2505.01390 (2025).

[9] A. Rofena, et al., A deep learning approach for virtual contrast enhancement in Contrast Enhanced Spectral Mammography, Computerized Medical Imaging and Graphics 116 (2024) 102398.

[10] M. Salmè, et al., Evaluating Vision Language Model Adaptations for Radiology Report Generation in Low-Resource Languages, arXiv preprint arXiv:2505.01096 (2025).

[11] A. Rofena, et al., Augmented Intelligence for Multimodal Virtual Biopsy in Breast Cancer Using Generative Artificial Intelligence, arXiv preprint arXiv:2501.19176 (2025).

[12] A. Rofena, et al., Lesion-Aware Generative Artificial Intelligence for Virtual Contrast-Enhanced Mammography in Breast Cancer, in: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems, 2025, pp. 141–146.

[13] V. Guarrasi, et al., Timing Is Everything: Finding the Optimal Fusion Points in Multimodal Medical Imaging, arXiv preprint arXiv:2505.02467 (2025).

[14] F. Aksu, et al., Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from CT and PET, Pattern Recognition Letters 193 (2025) 86–93.

[15] L. Furia, et al., Exploring early stress detection from multimodal time series with deep reinforcement learning, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 1917–1920.

[16] F. Ruffini, et al., Multi-Dataset Multi-Task Learning for COVID-19 Prognosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 251–261.

[17] L. Nibid, et al., Deep pathomics: A new image-based tool for predicting response to treatment in stage III non-small cell lung cancer, Plos one 18 (2023) e0294259.

[18] C. Z. Liu, et al., Exploring deep pathomics in lung cancer, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 407–412.

[19] A. Francesconi, D. Cappetta, F. Rebecchi, P. Soda, V. Guarrasi, R. Sicilia, Cross-dataset multivariate time-series model for parkinson’s diagnosis via keyboard dynamics, arXiv preprint arXiv:2510.15950 (2025).

[20] A. Francesconi, et al., Class balancing diversity multimodal ensemble for Alzheimer’s disease diagnosis and early detection, Computerized Medical Imaging and Graphics 123 (2025) 102529.

[21] F. Aksu, et al., Enhancing NSCLC histological subtype classification: A federated learning approach using triplet loss, in: International Conference on Pattern Recognition, Springer, 2024, pp. 154–168.

[22] C. M. Caruso, et al., Not Another Imputation Method: A Transformer-based Model for Missing Values in Tabular Datasets, arXiv preprint arXiv:2407.11540 (2024).

[23] C. M. Caruso, et al., A deep learning approach for overall survival prediction in lung cancer with missing values, Computer Methods and Programs in Biomedicine 254 (2024) 108308.

[24] C. M. Caruso, et al., MARIA: A multimodal transformer model for incomplete healthcare data, Computers in Biology and Medicine 196 (2025). doi:10.1016/j.compbiomed.2025.110843.

[25] V. Guarrasi, et al., Multi-objective optimization determines when, which and how to fuse deep networks: An application to predict COVID-19 outcomes, Computers in Biology and Medicine

154 (2023) 106625.

[26] K. Mogensen, et al., An optimized ensemble search approach for classification of higher-level gait disorder using brain magnetic resonance images, Computers in Biology and Medicine 184 (2025).

[27] E. M. Ayllón, et al., Can Foundation Models Really Segment Tumors? A Benchmarking Odyssey in Lung CT Imaging, in: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2025, pp. 375–380.

[28] V. Guarrasi, et al., Whole-body image-to-image translation for a virtual scanner in a healthcare digital twin, in: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems, 2025, pp. 528–534.

[29] M. Salmè, et al., Beyond the Generative Learning Trilemma: Generative Model Assessment in Data Scarcity Domains, arXiv preprint arXiv:2504.10555 (2025).

[30] M. Mantegna, et al., Benchmarking GAN-Based vs Classical Data Augmentation on Biomedical Images, in: International Conference on Pattern Recognition, Springer, 2024, pp. 92–104.

[31] D. Molino, et al., MedCoDi-M: A multi-prompt foundation model for multimodal medical data generation, arXiv preprint arXiv:2501.04614 (2025).