Building an AI-Enabled Metaverse for Intelligent Healthcare: Opportunities and Challenges

Valerio Guarrasi[1,*], Lorenzo Tronchin[1], Camillo Maria Caruso[1], Aurora Rofena[1], Guido Manni[1], Fatih Aksu[2], Domenico Paolo[1], Giulio Iannello[1], Rosa Sicilia[1], Ermanno Cordelli[1] and Paolo Soda[1,3]

[1]Unit of Computer Systems and Bioinformatics, Department of Engineering, University Campus Bio-Medico of Rome, Italy

[2]Department of Biomedical Sciences, Humanitas University, Milan, Italy

[3]Department of Radiation Sciences, Radiation Physics, Biomedical Engineering, Umeå University, Sweden

Abstract

This abstract discusses the development of a metaverse for intelligent healthcare, which involves creating a virtual environment where healthcare professionals, patients, and researchers can interact and collaborate using digital technologies. The metaverse can improve the efficiency and effectiveness of healthcare services and provide new opportunities for research and innovation. AI models are necessary for analyzing patient data and providing personalized healthcare recommendations, but the data in a metaverse setting is inherently multimodal, unstructured, noisy, incomplete, limited, or partially inconsistent, which poses a challenge for AI models. However, it becomes necessary the integration of AI models for the development of virtual scanners to simulate image modalities, and robotics to simulate surgical procedures within a virtual environment. The ultimate goal is to leverage the power of AI to enhance the quality of healthcare in a metaverse for intelligent healthcare, which has the potential to transform the way healthcare services are delivered and improve health outcomes for patients worldwide.

Keywords

Artificial Intelligence, Multimodal Learning, Metaverse, Healthcare

1. Introduction

The development of a metaverse for intelligent healthcare involves the creation of a virtual environment where healthcare professionals, patients, and researchers can interact and collaborate using digital technologies [1]. The metaverse can be thought of as a virtual world where users can engage with each other in real-time, create and manipulate digital objects, and perform complex tasks in a simulated environment. In the context of healthcare, the metaverse can be used to improve the efficiency and effectiveness of healthcare services, as well as provide new opportunities for research and innovation. For example, healthcare professionals can use the metaverse to perform virtual consultations with patients, monitor their health remotely, and collaborate with other healthcare providers in real-time.

The development of a metaverse for intelligent healthcare requires a range of technologies and skills, including artificial intelligence (AI), virtual and augmented reality, natural language processing, and data analytics. These technologies can be used to create realistic simulations of healthcare scenarios, analyze patient data, and provide personalized recommendations based on individual health needs.

The development of AI models for a healthcare metaverse holds great potential to improve the quality and accessibility of healthcare. By analyzing vast amounts of data on an individual's health, lifestyle, and medical history, AI models can provide personalized healthcare recommendations tailored to each patient's needs. However, the data in a metaverse setting is inherently multimodal, which requires the development of AI models capable of exploiting the relevant features of each modality. Moreover, the data may be unstructured, noisy, incomplete, limited, or partially inconsistent, which poses a challenge for AI models. Therefore, there is a need for resilient AI models that can work under these conditions. There are many potential applications of the healthcare metaverse; one of the most promising is the creation of virtual scanners to generate virtual medical images for patients. This requires further research on AI systems capable of translating one image modality to another. In addition, the integration of AI models with robotics is necessary to simulate surgical procedures within a virtual environment.

2. Multimodal Learning

The healthcare metaverse can use multimodal data in various ways to improve healthcare services and outcomes. Multimodal data refers to data that is generated from

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.
valerio.guarrasi@unicampus.it (V. Guarrasi)
0000-0002-1860-7447 (V. Guarrasi)

multiple sources [2].

By combining data from different sources, such as medical images, genetic information, medical history available within the electronic health records, and lifestyle data, the healthcare metaverse can exploit multimodal data to create personalized patient health profiles that, in turn, offer the chance to tailor treatments and interventions. This would help improve the accuracy and efficiency of medical diagnoses and prognoses.

Next sections overview research ongoing in our lab on this topic.

2.1. When, which and how?

Deep multimodal learning has shown promising performance also encompassing those attained by traditional machine learning approaches. This happens because deep neural networks permit us to fuse the learners exploiting the loss backpropagation at different depths. However, understanding "when", "which", and "how" to fuse the modalities is the main open methodological research question now opened.

The "when" question involves determining the optimal depth in the network architecture to combine different types of data, such as imaging and clinical data, to improve accuracy and reliability. To determine the ideal point of fusion among the different modalities in a joint fusion scenario, we have developed an iterative algorithm that increases the number of fusion connections among convolutional networks, enabling us to obtain optimal results. Our findings indicate that the gradual fusion mechanism among modalities is highly effective, resulting in superior performance compared to traditional fusion techniques.

The "which" question involves determining which modalities and which models to fuse. Indeed, given a task characterized by a specific source of data available, practitioners and researchers often need to find out the most relevant and useful models. In this respect in [3] we present a multi-objective optimized ensemble search that selects the best ensemble of networks to satisfy a classification task, outperforming individual models. Our method optimizes the search of this optimal ensemble by exploiting both evaluation and diversity metrics and it was applied to various contexts like COVID-19 diagnosis [4] and lung cancer overall survival [5] (deepened in the next subsection).

The "how" question deals with determining how to fuse the modalities. This involves determining the best method for integrating different types of data, such as using convolutional neural networks to analyze images and recurrent neural networks to analyze time-series data.

To cope with these three questions in [6] we present a novel approach optimizing the setup of a multimodal end-to-end model. It exploits Pareto multi-objective optimization working with a performance metric and the diversity score of multiple candidate unimodal neural networks to be fused. We attain state-of-the-art results, not only outperforming the baseline performance but also being robust to external validation. Via this method, we automatically understand "which" are the most suited modalities and models for the task, "when" the fusion of the modalities should occur, and "how" to optimally fuse them.

2.2. Multimodal Ensemble for Overall Survival

In the context of lung cancer, multimodal learning is becoming an increasingly important research area as it can provide insights into the optimal treatment for aggressive tumors. In this particular work [5], we utilized the CLARO dataset [7], which includes CT images and clinical data from non-small-cell lung cancer patients, to investigate the use of multimodal learning for predicting overall survival.

We employed a late fusion approach, which involves training unimodal models on each modality separately and then combining the results using an ensemble method. We selected the optimal set of classifiers for the ensemble by solving a multi-objective optimization problem that maximizes both performance and diversity of the unimodal models. The results of this study demonstrate the potential of multimodal learning in improving the prediction of overall survival in lung cancer patients. The proposed ensemble outperformed models trained on a single modality and achieves state-of-the-art results on the task at hand.

2.3. Multimodal XAI

Explainable Artificial Intelligence (XAI) methods are becoming increasingly important in the development of intelligent healthcare systems, particularly in the context of multimodal learning [8].

The use of multimodal learning can also make it more challenging to understand how the model arrived at its decision, particularly when the model is making decisions based on multiple inputs. This is where XAI methods become essential. XAI methods can help healthcare professionals and patients understand how the model arrived at its decision by providing interpretable and transparent explanations. This can help increase trust in the system and ultimately improve patient outcomes.

In the context of a metaverse for intelligent healthcare, XAI methods could be used to explain how the system is diagnosing a patient based on their symptoms and medical history. This explanation could be provided in a visual or interactive format, making it easy for the patient to understand and engage with.

Practical applications have impacted supervised multimodal fusion to explain decisions taken when identifying patients affected by SARS-CoV-2 at risk of severe outcomes, such as intensive care or death. Using chest X-ray scans and clinical data, our approach creates a multimodal embedded representation of the data using a convolutional Autoencoder for the imaging modality and an Autoencoder for the tabular modality [9]. This embedded representation is then joined with a multilayer perceptron that performs the classification task. To help us understand how the network performs classifications, our novel XAI algorithm works on the variations of the embedding, computed by applying a latent shift that simulates a counterfactual prediction. This reveals the features of each modality that contribute the most to the decision and computes a quantitative score indicating the modality's importance. By reducing the model's opacity, this approach improves trust and transparency for doctors and regulators who may have difficulty trusting the MDL models: indeed, the results on the AIforCOVID dataset [10], which contains multimodal data of clinical records and chest X-ray images of patients from six Italian hospitals, show that the proposed method provides meaningful explanations without degrading the classification performance for the early identification of COVID-19-positive patients at risk of severe outcome.

2.4. Federated Learning

It is common practice in AI research to collect data from multiple sources and send it to a central server for computation. However, this approach poses several challenges in healthcare where sensitive patient data must be protected to avoid privacy violations.

To address this issue, federated learning has emerged as a potential solution [11]. This is a critical requirement for the development of an efficient metaverse of intelligent healthcare, where patient data is protected and used to enhance patient care. Federated learning allows for training of a shared global model with a central server while keeping the data private in local institutions, thereby promoting a greener and more secure practice.

In this area, we propose a new paradigm as a variant of the most widely used approach, in which, instead of training individual client models independently for a number of epochs each turn, and then sending their weights back to the server, which takes care of aggregating them into a single set of weights and redistributing them back to the models, the concept of a token is introduced, passed to each epoch sequentially or randomly among the clients, which is intended to allow the weights to be sent to the server only by its owner, who redistributes them directly to all models. The absence of local training epochs and the immediate broadcast allows the entire paradigm to be restructured by building a single model passed between clients, even eliminating the role of the server and thus halving the number of parameters sent in each round.

3. Resilient AI

The problem of unstructured, noisy, incomplete, limited in number, or partially inconsistent data is a significant challenge for many areas of data-driven research, especially in healthcare. In AI, such situations could impact models' accuracy and reliability, leading to incorrect or biased outcomes. Hence, developing resilient AI systems able to handle such types of data is crucial to ensure the ethical and responsible use of AI in various applications.

3.1. Missing Features

Missing data is a common problem in healthcare datasets, occurring when some information is not available for some patients or variables in a dataset. Missing data not only could bias the results, but it often contrasts with the needs of AI models, which often require complete data to function properly.

There are several methods for dealing with missing data in healthcare datasets, including imputation or deletion techniques. However, each approach has its own strengths and limitations, and the appropriate method depends on the specifics of the dataset and the research question being addressed.

To address such issues, we developed a Transformer model that masks the missing features during training [12], so that the model ignores any missing data during training. This approach eliminates the need for imputation or deletion techniques, as the model simply does not consider the missing data. In practice, the proposed method leverages the idea of the mask inside the self-attention module to learn from incomplete input data, which signals to the model only the positional encoding eliminating the input embeddings. Furthermore, to shift to the use of tabular data with a Transformer model, we introduce a novel type of positional encoding that identifies the feature and not the position.

We experimentally validated the method on a classification task that aims to predict the overall survival in patients affected by non-small-cell lung cancer using clinical data from the CLARO [7] project. The results show that this approach overcomes the limitations of traditional imputation methods, reduces bias, and improves the accuracy of the final analysis.

3.2. Siamese Networks

It is well known that the AI's power of analyzing vast amounts of data is an element lying behind models' performance. However, data availability is a major barrier in many domains, healthcare and metaverse included. To overcome this limitation, several works in the literature have studied how to learn in case of limited training data, and, among them, Siamese networks are a viable alternative. They consist of two or more identical networks working in parallel and triplet networks is an established subtype of this approach, where three identical networks are used. During training, two of the inputs of these three networks belong to the same class, while the third belongs to a different class. The main goal is to learn a feature space where each class forms a cluster, used in the inference step to classify new instances. Since triplet networks utilize inter-class diversities in addition to intra-class similarities, and the number of possible triplets is much higher than the number of instances, they can be a convenient option for applications where the data is limited.

On a private dataset with 86 patients, representing a real case with limited data, we demonstrated that triplet networks outperform deep networks using the softmax classifier to predict histological NSCLC subtypes from CT scans.

3.3. Name-entity Recognition

The use of electronic health records (EHRs) can provide valuable data for medical research since physicians register information that describes symptoms, the diagnosis, treatments, and the evolution of the patient over time, which represent an invaluable source of data to build patient metaverse. However, the analysis of EHRs can be difficult due to the presence of a large amount of unstructured data and complex clinical language used by physicians. Clinical name-entity recognition (NER) is the natural language processing task of identifying and categorizing medical information (entities) in clinical text.

We investigate the use of clinical NER to extract relevant clinical information from Italian-language EHRs of NSCLC patients included in the CLARO project, investigating a transformed-based approach. In particular, we aim to fine-tune bioBIT, an Italian biomedical pretrained model, that derives from BERT, for the task of clinical NER. To this end, we also first identified a set of relevant entities, which are then used to annotate the data cohort.

Class imbalance learning is a common issue for clinical NER since some entities may appear much less frequently than others. To overcome it, we propose to use the Focal Loss, making the model more focused on rarer entities. On a cohort of patients affected by Non-Small Cell Lung Cancer, we obtained promising f1 score results in the recognition of clinical entities.

4. Virtual Scanner

In a metaverse for intelligent healthcare, a virtual scanner refers to a computer-generated imaging device that uses virtual reality technology to create medical images of a patient's body. Virtual scanners can be used to create detailed 3D images of internal organs, bones, tissues, and other structures without the need for invasive procedures, also minimizing patient discomfort as well as allowing medical professionals to view and manipulate images in ways that would not be possible with traditional imaging techniques. AI image translation models can be particularly useful in this context. These models use deep learning (DL) algorithms to translate medical images from one modality to another. In this context, we are working in three directions.

In the first [13], we aim to develop and validate DL models to perform Virtual Contrast Enhancement (VCE) in Contrast Enhanced Spectral Mammography (CESM). It is a diagnostic imaging technique in breast cancer that, unlike standard mammography, involves the injection of an iodinated contrast medium, which diffuses into the tumor tissue and enhances lesion visibility. Although this results in improved diagnostic accuracy, especially in patients with dense parenchymal tissue, and although it is a more affordable and accessible alternative to contrast-enhanced MRI, CESM has two main weaknesses. One is a biological risk due to ionizing radiation, whilst the other refers to possible reactions to iodinated contrast agents, such as nephropathy. CESM acquires a low-energy (LE) and a high-energy (HE) image in quick succession for each breast: the LE image is equivalent to standard mammography, whereas the HE image is post-processed with the corresponding LE image to compute the recombined image, which suppresses parenchymal tissue so that only areas of contrast enhancement are visible. Our VCE task asks AI models to output a virtually recombined image having only the LE image as input. To this end, we compared three DL models: an autoencoder, a CycleGAN, and the Pix2Pix. Results on a public [14] and on a private dataset with 1003 and 105 patients scans, respectively, show that the CycleGAN is able to effectively perform the image-to-image translation task in the lesion areas, suggesting VCE is a viable alternative that deserves more research efforts. Therefore, VCE can be useful in a metaverse for intelligent healthcare since it can reduce the need for the injection of an iodinated contrast medium and it can eliminate the need for undergoing double radiation. This would make imaging procedures safer and it would reduce the time and cost associated with obtaining diagnostic images.

Turning our attention to the second task, we aim to develop and validate DL models to perform MR-to-CT image translation—to generate synthetic CT (sCT) images from real MR images. While CT is required for treatment planning, due to its excellent anatomic localisation, it is often complemented with MR imaging, because of its superior soft-tissue contrast. However, taking multiple images can be cost-prohibitive, burdensome to the patient, and problematic in light of CT ionization risk. For these reasons, MR-only treatment planning has become an attractive alternative, and one of the main tasks in the MR-only framework is MR-to-CT image translation—to generate sCT images from real MR images. Many DL models have been developed for the MR-to-CT translation task, using convolutional neural networks to generate sCT images, by minimizing pixel-wise differences to reference CT volumes. More recently, GAN models have been used to solve this task: the pix2pix model has been used, obtaining very promising results when there are co-registered MR and CT images, and the CycleGAN model has been used to handle the case with unpaired MR and CT images. One of the main issues, however, has consistently been the limited access to large datasets. Most ML tasks in medical imaging suffer from a lack of data and/or a lack of manually annotated data. To manage this issue, a technique called data augmentation has been widely utilized to improve generalization and robustness when training deep neural networks. We developed a novel strategy for latent data augmentation, exploiting the learned internal latent representation of a StyleGAN2 model, to generate medical images of sufficiently high resolution and quality that can be used to augment the real patient data available to develop a model that can generate sCT images from MR images. The StyleGAN2 model allows to generate multiple imaging modalities (i.e., MR and CT) of the same underlying synthetic patient. A number of augmented/generated images are used together with a number of real images (the fraction decided by the analyst) to train the MRI-to-CT translation model. The StyleGAN2 model is able to generate highly realistic synthetic images, but it does not provide any control over the generated images. We developed a novel generative procedure that has two main components: 1) an improved inverse generator, that embeds real MR and CT images in the StyleGAN2 model's learnt latent space by enforcing the resulting regenerated images to retain the semantic properties of the original real images at multiple levels; 2) a regularisation term controlling the distance in the latent space of a new purely generated image to the real images, which allows the analyst to tune the variability in the generated images. In particular, the proposed approach allows to maximize the diversity of the data sampled from the GAN manifold while retaining their realness, i.e., the synthesized images will be plausible datapoints from the learned real manifold (of real CT-MR images). A comparative analysis between the performance of the pix2pix model trained with and without data augmentation procedure is performed, showing a reduction of both scores using latent augmentation procedure. Thus, latent DA can be used to develop better MR-to-CT allowing the doctor to inspect multiple image modalities without the need for invasive procedures.

In the context of sCT generation, another pillar of our research focuses on developing denoise techniques with the aim of reducing the dose while maintaining the quality of the virtual scanner provided to the clinician. Indeed, under the ALARA (As Low As Reasonably Achievable) principle, the use of low-Dose (LD) acquisition protocol has become a clinical practice to be preferred over high-Dose (HD) protocols. However, if on the one side a reduction of radiations can be achieved, on the other, the overall quality of the reconstructed CT decreases. Thereby, a trade-off between dose and noise level must be tackled to ensure sufficient diagnostic image quality. This is the reason why a lot of effort has been put into investigating denoising strategies, trying to obtain high-quality CT images at the lowest cost in terms of radiation. To overcome this task, we develop a texture-based loss function to be included in the objective of the CycleGAN during training from LDCT to HDCT images. We move from the hypothesis that the noise due to LD protocols has a textural nature. Thus a texture-based loss will be beneficial during training allowing a better denoising quality and faster training.

Overall, virtual scanners and AI image translation models have the potential to revolutionize healthcare by making it more efficient, less invasive, and more personalized.

5. Virtual Surgery

Robotic surgery has revolutionized complex medical applications such as minimally invasive, orthopedic, brain, and radiotherapy surgeries. With the adoption of robotic surgery, automation in the surgical domain has become a crucial topic, presenting potential benefits like improved consistency, increased dexterity, and access to standard techniques [15].

Path planning of surgical robots' end effectors is a promising area for automation, which can automate surgical tasks or provide suggestions to the surgeon. However, classical path planning algorithms have limitations in dealing with dynamic environments like the surgical environment. In this respect, early attempts in this field have proven the feasibility and the potential impacts of the automation, but also the limitations of classical path planning algorithms. Reinforcement learning, even exploiting deep architectures, could be a viable alternative for path planning helping overcome such limitations.

For this reason, we are developing a deep reinforcement learning (DRL) framework that can plan an optimal path from the entry point to the target point within the patient, and update it during surgery using the endoscopic camera feedback. Our study is divided into two subproblems: global and local. The global subproblem involves planning the optimal trajectory, given information about the entry point, target point, and a digital twin of the patient obtained from CT/MRI scans. Conversely, the local subproblem focuses on updating the path based on the endoscopic video feed.

To tackle both subproblems we developed a simulation environment that generates a digital twin of the patient from CT/MRI scans and constructs a 3D environment of the surgical environment from monocular endoscopic images. The DRL agent is then trained to plan a global optimal trajectory and learn to update the local trajectory in real-time from the video feedback.

Potential future applications of this DRL framework into the metaverse are straightforward. For example, in augmented reality, the DRL agent could be utilized to provide suggestions to the surgeon on how to plan the next move in the surgical environment, by overlaying the DRL-generated path onto the surgeon's field of vision. Furthermore, the simulation environment and DRL agent can also be used to train surgeons on different cases in a safe and controlled environment.

Acknowledgments

We acknowledge: FONDO PER LA CRESCITA SOSTENIBILE (F.C.S.) Bando Accordo Innovazione DM 24/5/2017 (Ministero delle Imprese e del Made in Italy), under the project entitled "Piattaforma per la Medicina di Precisione. Intelligenza Artificiale e Diagnostica Clinica Integrata" (CUP B89J23000580005); PNRR MUR project PE0000013-FAIR; the project n. F/130096/01-05/X38 - Fondo per la Crescita Sostenibile - ACCORDI PER L'INNOVAZIONE DI CUI AL D.M. 24 MAGGIO 2017 - Ministero dello Sviluppo Economico (Italy), iii) Programma Operativo Nazionale (PON) "Ricerca e Innovazione" 2014-2020 CCI2014IT16M2OP005 Azione IV.4.; Regione Lazio PO FSE 2014-2020 Avviso Pubblico "Contributi per la permanenza nel mondo accademico delle eccellenze" Asse III – Istruzione e formazione - Priorità di investimento 10 ii) - Obiettivo specifico 10.5 Azione Cardine 21.

References

[1] G. Wang, A. Badal, X. Jia, J. S. Maltz, K. Mueller, K. J. Myers, C. Niu, M. Vannier, P. Yan, Z. Yu, et al., Development of metaverse for intelligent healthcare, Nature Machine Intelligence (2022) 1–8.

[2] T. Baltrušaitis, et al., Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[3] V. Guarrasi, N. C. D'Amico, R. Sicilia, E. Cordelli, P. Soda, Pareto optimization of deep networks for covid-19 diagnosis from chest x-rays, Pattern Recognition 121 (2022) 108242.

[4] V. Guarrasi, N. C. D'Amico, R. Sicilia, E. Cordelli, P. Soda, A multi-expert system to detect covid-19 cases in x-ray images, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 395–400.

[5] C. M. Caruso, V. Guarrasi, E. Cordelli, R. Sicilia, S. Gentile, L. Messina, M. Fiore, C. Piccolo, B. Beomonte Zobel, G. Iannello, et al., A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer, Journal of Imaging 8 (2022) 298.

[6] V. Guarrasi, P. Soda, 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.

[7] N. C. D'Amico, R. Sicilia, E. Cordelli, L. Tronchin, C. Greco, M. Fiore, A. Carnevale, G. Iannello, S. Ramella, P. Soda, Radiomics-based prediction of overall survival in lung cancer using different volumes-of-interest, Applied Sciences 10 (2020) 6425.

[8] G. Joshi, et al., A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821.

[9] V. Guarrasi, L. Tronchin, D. Albano, E. Faiella, D. Fazzini, D. Santucci, P. Soda, Multimodal explainability via latent shift applied to covid-19 stratification, arXiv preprint arXiv:2212.14084 (2022).

[10] P. Soda, N. C. D'Amico, J. Tessadori, G. Valbusa, V. Guarrasi, C. Bortolotto, M. U. Akbar, R. Sicilia, E. Cordelli, D. Fazzini, et al., Aiforcovid: Predicting the clinical outcomes in patients with covid-19 applying ai to chest-x-rays. an italian multicentre study, Medical image analysis 74 (2021) 102216.

[11] N. Truong, et al., Privacy preservation in federated learning: An insightful survey from the gdpr perspective, Computers & Security 110 (2021) 102402.

[12] C. M. Caruso, P. Soda, V. Guarrasi, A deep learning approach for overall survival analysis with missing data, Submitted to: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) (2023).

[13] A. Rofena, P. Soda, V. Guarrasi, A deep learning approach for virtual contrast enhancement in contrast enhanced spectral mammography (cesm), Submitted to: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) (2023).

[14] R. Khaled, et al., Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research, Scientific Data 9 (2022) 122.

[15] P. Fiorini, Automation and autonomy in robotic surgery, 2021.