Enabling Trustworthy Predictions using AI, ML and Multi-Modal Data

Andrea Bianchi, Giordano d'Aloisio, Antinisca Di Marco, Giovanni Stilo
Department of Engineering and Information Sciences and Mathematics
University of L'Aquila
{andrea.bianchi,giordano.daloisio}@graduate.univaq.it
{antinisca.dimarco,giovanni.stilo}@univaq.it

***Abstract*—Artificial Intelligence (AI) and Machine Learning (ML) systems can help stakeholders addressing challenges in several domains. To this aim, a huge amount of heterogeneous data is available and can be fed into such systems to solve different tasks (like image recognition, data predictions and so on). However, managing the complexity and multi-modality of data is always a complex task. Moreover, nowadays developing effective (i.e., accurate) AI and ML systems is no longer sufficient to have a system which is of high quality. Indeed, AI and ML systems must be *fair*, *explainable*, and *private* with respect to sensitive information of the user (in other words, *trustworthy*).**

In this extended abstract, we present an approach to allow the trustworthy predictions based on AI, ML and multi-modal data. We describe the high-level architecture of the system and discuss its possible applications in the medical domain.

***Index Terms*—machine learning, multi-modal data, trustworthiness, quality, fairness, explainability, privacy**

I. Introduction

Nowadays, a huge amount of multi-modal data is available and can be used to help stakeholders in different domains through Artificial Intelligence (AI) and Machine Learning (ML) systems. Thinking about the medical domain, a considerable amount of sensible data produced by a large set of diagnostic and instrumental tests integrated with the data obtained by high-throughput technologies might be used to strengthen predictions to improve the prevention, to reduce the time-to-diagnosis and hence the costs of the health system, while bringing out hidden knowledge.

In general, the complexity and the *multi-modality* of data, i.e., data that spans different types and contexts (e.g. images, audios, text and others), lead to challenges in developing methods that can extract peaces of knowledge (*sub-domain*) from each type of data (e.g., identify a person in a video, identify a cancer in a diagnostic image) and can integrate them in a wider knowledge useful to improve predictions in a specific domain (*domain-knowledge*). On the other hand, the *trustworthiness* of AI and ML systems, i.e., systems that are *explainable* [1], *fair* [2] and *private* with respect to sensitive information of the people [3], is nowadays a paramount to develop systems that are general and can be used by different groups and populations. Considering again the medical domain, the results obtained by AI and ML systems must be *explainable* (i.e., they must describe why the prediction has been defined), *fair* towards individuals belonging to *sensitive* groups (e.g., women, non-white people, young people, and so on) and must *protect* sensitive information of the users to be actually meaningful. In addition, given the wide adoption of AI and ML in many domains, the development of trustworthy systems must be made available also to non-expert users (namely, *democratize*).

In this work, we aim to solve the issues described above by proposing an approach for the trustworthy prediction of multi-modal data. In particular, we aim to answer these two challenges: *i)* how can we analyze multi-modal data in order to learn specific (i.e., sub-domain) and integrated (i.e., domain) knowledge? *ii)* how can we assure the trustworthiness of the predictions obtained using multi-modal data, both for specific and integrated knowledge?

In the following, we present a system we are developing to answer the challenges mentioned above. First, we discuss how we can learn domain knowledge by using multi-modal data. Next, we describe how we can assure the trustworthiness of these predictions, democratizing the trustworthy assurance process. To help the reader, we use the medical domain as motivating scenario.

II. Combining Predictions in multi-modal heterogeneous Health data

In order to achieve goals for the trustworthiness, let's take into account a very common scenario in medicine, where various data sources are available, but data are stored in different physical places and, for legal reasons, cannot be freely distributed even between several medical departments. In such scenario, the implementation of a single monolithic learning model, trained using multi-modal data, cannot be applied. Rather, we envision a hierarchical AI system that masters multi-modal data by learning from the single sources using trustworthy models to maximize the trust of stakeholders (such as, domain experts and patients). We also believe that a hierarchical AI system eases the trustworthiness of the predictions. We observe that the envisaged AI system is an extension of the federated learning approach [4] by

This work was supported in part by the Territori Aperti (a project funded by Fondo Territori, Lavoro e Conoscenza CGIL CISL UIL), and in part by the SoBigData RI (a project funded by SoBigData-plusplus H2020-INFRAIA-2019-1 EU Project - Contract n. 871042, SoBigData RI PPP HORIZON-INFRA-2021-DEV-02-01 EU project - Contract n. 101079043, and SoBigData.it PNRR project - CUP n. B53C22001760006).

also adding trustworthiness of the system. Subsequently, put together all the obtained predictions to train a holistic model able to semantically link and backtrack the motivations of the predictions by using explainable AI techniques. We call this holistic model, Hyper-Model (HM).

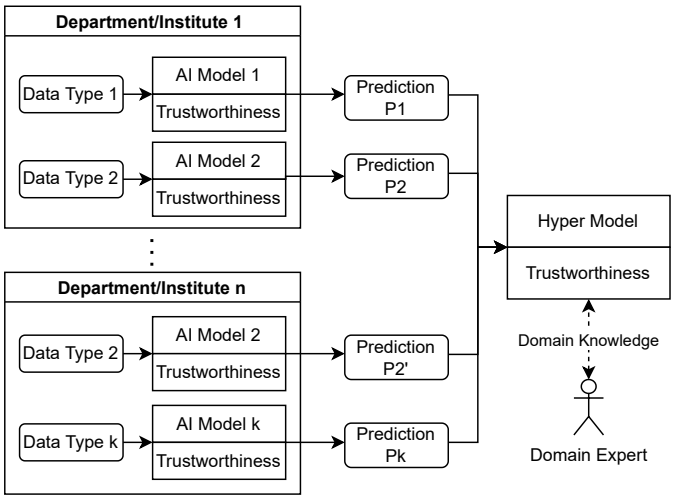

Fig. 1. Trustworthy Prediction System Architecture

Figure 1 depicts a high-level architecture of our system that is composed by two layers. The first layer is fed by available data coming from different sources (for instance, recalling the medical domain different institutes or institutes' departments) and it will implement a specific learning pipeline for each data type. This holds for all the different sources managed in the system. The second layer corresponds to the HM itself, fed by the predicted results of the underlying individual learning pipelines eventually supported by the *domain experts' knowledge* (e.g., clinical knowledge) to semantically link the prediction obtained by different models. The domain knowledge can also be exploited to improve pipeline's phases (e.g., Data Collection, Data Cleaning, Feature Engineering). Of course, the prediction obtained by the whole system also contribute to increase the domain knowledge. In critical domains, such as the healthcare, the knowledge integration and improvement play a fundamental role since it allows for calibrating and verifying the inputs and the output obtained in the learning process, it also permits to increase trustworthiness both in the first layer (composed by individual models) and in the second layer (i.e., the hyper-model), and to develop new knowledge.

III. Assuring the Trustworthiness of the System

In order to assure that the prediction obtained by our system is trustworthy, we must select a set of methods that can better satisfy the fairness, explainability and privacy of our system. However, these properties are strictly related to the involved dataset and must be assessed each time a new training phase of the single pipelines or of the whole HM is performed. In addition, given the wide adoption of AI and ML systems, this trustworthiness evaluation process must be made available also to non-technical users (e.g., doctors in the medical domain). To solve the challenges mentioned above, we propose a solution to democratize the trustworthy assurance process of ML systems through a low-code platform [5], [6]. This framework aims to provide an environment for the configuration of experiments that automatically selects the algorithm better satisfying trustworthiness. This will simplify the work of the domain experts and will democratize the trustworthy assurance process. We rely on the concept of Extended Feature Models (ExtFM) to model such trustworthy evaluation experiment as Software Product Lines (SPL) [7], [8]. Each experiment evaluates a set of methods to enhance a given trustworthy attribute (i.e., explainability, fairness, and privacy) and computes a set of metrics related that attribute. Finally, for each trustworthy attribute, a report of such metrics is returned to the user. Using this report, the domain expert can evaluate which algorithm performs better for each trustworthy attribute and can embed it in the final system.

As a further challenge, critical AI and ML systems, such as healthcare, requires a continuous learning process to be always updated with new knowledge. For this reason, a continuous assessment of trustworthiness is needed.

References

[1] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, Jul. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3457607

[3] R. Xu, N. Baracaldo, and J. Joshi, "Privacy-Preserving Machine Learning: Methods, Challenges and Directions," *arXiv:2108.04417 [cs]*, Sep. 2021, arXiv: 2108.04417. [Online]. Available: http://arxiv.org/abs/2108.04417

[4] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[5] G. d'Aloisio, A. Di Marco, and G. Stilo, "Modeling Quality and Machine Learning Pipelines through Extended Feature Models," Jul. 2022.

[6] G. d'Aloisio, "Quality-driven machine learning-based data science pipeline realization: a software engineering approach," in *2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE, 2022, pp. 291–293.

[7] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, and A. S. Peterson, "Feature-oriented domain analysis (foda) feasibility study," Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, Tech. Rep., 1990.

[8] J. A. Galindo, D. Benavides, P. Trinidad, A.-M. Gutiérrez-Fernández, and A. Ruiz-Cortés, "Automated analysis of feature models: Quo vadis?" *Computing*, vol. 101, no. 5, pp. 387–433, May 2019. [Online]. Available: http://link.springer.com/10.1007/s00607-018-0646-1