Trustworthy Machine Learning Predictions to Support Clinical Research and Decisions

Andrea Bianchi*, Antinisca Di Marco*, Francesca Marzi*, Giovanni Stilo*, Cristina Pellegrini*,
Stefano Masi †, Alessandro Mengozzi †, Agostino Virdis †, Marco Salvatore Nobile ‡, and Marta Simeoni ‡
*University of L'Aquila, Italy, †University of Pisa, Italy, ‡Ca' Foscari University of Venice, Italy

2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CBMS58004.2023.00222

***Abstract*—Nowadays, physicians have at their hands a huge amount of data produced by a large set of diagnostic and instrumental tests integrated with data obtained by high-throughput technologies. If such data were opportunely linked and analysed, they might be used to strengthen predictions, so that to improve the prevention and the time-to-diagnosis, reduce the costs of the health system, and bring out hidden knowledge. Machine learning is the principal technique used nowadays to leverage data and gain useful information. However, it has led to various challenges, such as improving the interpretability and explainability of the employed predictive models and integrating expert knowledge into the final system. Solving those challenges is of paramount importance to enhance the trust of both clinicians and patients in the system predictions. To solve the aforementioned issues, in this paper we propose a software workflow able to cope with the trustworthiness aspects of machine learning models and considering a multitude of heterogeneous data and models.**

***Index Terms*—heterogeneous data, machine learning, explainability, interpretability, risk prediction, clinical decision support system**

I. Introduction

21st-century healthcare systems face a hard challenge due to the ageing population. An example comes from cardiovascular disease (CVD), which counts 17.9 million deaths per year. According to recent predictions, the situation is worsening: the prevalence of the metabolic disease will rise, and in 2030 1 out of 2 U.S. adults will suffer from obesity [25]. With the ageing population [8], high-risk cardiovascular phenotypes will prevail, negatively impacting cardiovascular mortality and morbidity. These considerations can be extended to several chronic diseases (e.g. cancer, neurological and autoimmune disease) that, if not prevented, are deemed to encumber the healthcare systems. To face this challenge means adequately utilising medical resources and data to provide accurate, feasible, and easily implementable disease risk prediction models. Machine Learning (ML) can be used to extrapolate hidden information and improve disease prediction, such as CVD [12], [19], [22].

However, there is often a strong lack of trust when the results are obtained by applying black-box methods that cannot explain, or motivate, their response. Moreover, many of the proposed approaches work on a single data source. Instead, in [17] the authors highlight the importance of efficiently and effectively integrate all types of health data, from images to clinical to omics, to achieve accurate diagnosis or to suggest an efficient treatment. Nowadays, physicians are provided with many clinical risk scores. However, these suffer from three major flaws: i) they still fail to identify high-risk phenotypes; ii) they trade accuracy for feasibility: the advent of genetic and polygenic risk scores provided good estimators of lifetime risk that are poorly helpful if the patient already suffers from the disease [3]; iii) they are often perceived as obscure, limiting the trust of both physicians and the patients. On top of this scenario, patients' data are heterogeneous and often unavailable in the whole to both the physician and the researcher.

This paper aims to describe a methodology for defining ML systems in the health domain that exploit heterogeneous data sources and generate accurate, interpretable and explainable prediction results. The methodology makes use of domain specific knowledge – medical knowledge in this case – to ensure robust predictions. We expect our methodology, fuelled by population data and guidelines' recommendations, to be able to suggest the clinician the most effective, reliable, direct and cost-effective pathways that lead to an accurate, patient-tailored final diagnosis and treatment. While presenting specific results or sharing data is beyond the scope of this paper, we are working to validate our approach using real-world data in the near future.

II. Software Process for trustworthy health prediction Systems

In this section, we describe the workflow for trustworthy ML predictions shown in Figure 1, which guarantees interpretability and explainability of the predictions leveraging on learning pipelines. We summarise the requirements and constraints to be considered in the workflow design (Section II-A) and describe the software process we propose to achieve the challenge of a trustworthy ML prediction in case of multiple data sources (Section II-B).

This work is supported by European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021, by "ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing", funded by European Union – NextGenerationEU, and by LIFEMAP-Dalla patologia pediatrica alle malattie cardiovascolari e neoplastiche nell'adulto: mappatura genomica per la medicina e prevenzione personalizzata Traiettoria 3 "Medicina rigenerativa, predittiva e personalizzata" - Linea di azione 3.1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale" of the Ministry of Health.

2372-9198/23/$31.00 ©2023 IEEE
DOI 10.1109/CBMS58004.2023.00050

A. Challenges

Goal #1: Recent advances in multimodal learning have shown that the integration of different types of data (for instance clinical data, molecular diagnostics, radiological and histological imaging) leads to promising results. In [5] and [13] the authors show that prediction models that combine multiple types of data perform better than models that consider only a single data type. However, in real-world scenarios, accessibility to different data sources is often limited, making it challenging to perform a meaningful data integration and analysis. When heterogeneous data coming from different sources need to be analysed through ML techniques, a single pipeline is not adequate to generate the outcomes, especially if explainability and interpretability are to be imposed. Indeed, systems that apply those techniques to real contexts must use complex models that can combine multiple simpler sub-models.

Goal #2: Imagine a scenario where various data sources are available, but data are stored in different physical places and, for some reason, cannot be freely distributed. This is a typical situation in the health domain, where patients undergo various diagnostic exams in different hospitals or clinics that do not share the same repository for the outcomes. In such scenario, it is necessary to learn from each single source separately and, subsequently, put together all the obtained predictions to train a holistic model that we call *hyper model*. More details on the complete process are provided in Section II-B.

Goal #3: Given the complexity of the mathematical models used in the pipelines, it is important to ensure that their results can be correctly interpreted and explained. In the health domain, this requirement is mandatory since using black-box predictive models affects trust from clinicians and hampers the "right to explanation" for patients.

Goal #4: A further issue is how to improve the prediction results by introducing extra knowledge. Research papers like [9], [24] describe the importance of integrating knowledge within the learning pipelines to: i) reduce the amount of data needed, ii) make approaches more robust, and iii) create interpretable and explainable learning systems. As stated in [11], one of the opportunities in applying intelligent approaches and data integration techniques is to involve humans in the loop to conduct the labeling and the verification of data and results. In line with this suggestion, we aim to increase the reliability of the pipeline steps within the process by introducing the domain experts' knowledge.

Our solution for tackling these challenges involves implementing a hierarchical software workflow for the prediction system. This process begins with individual learning pipelines working on distinct data sources. Subsequently, the hyper model collects and merges the predictions of multiple models using a dedicated aggregation function. The primary goal of the hyper model is to use predictions from multiple models to make an informed final decision that produces an overall prediction result.

The proposed approach must not be confused with federated learning. Simple federated learning involves sharing the same data model among participants, resulting in identical predictions [28]. The proposed approach, instead, utilises a pre-defined learning pipeline for each data source, where the training datasets diverge due to differences in physical locations and the types and volume of data. In contrast to pure ensemble learning, which applies various models to the same subset of data [10], the proposed approach combines predictions obtained from different models trained on different data formats and volumes. This is done with the aim of improving the accuracy of predictions while providing insights into the decision-making process of the ensemble. Additionally, the proposed approach incorporates interpretability and explainability concepts to enhance the understanding of the ensemble's decision-making process.

B. Workflow of trustworthy ML systems

In this subsection, we present a detailed explanation of the steps involved in our approach to designing trustworthy ML systems, considering the challenges and constraints illustrated in the previous subsection. The proposed process is general and particularly relevant in the medical domain. It can be applied to any domain where data from different sources cannot be shared and where the mentioned challenges arise.

The workflow we propose is shown in Figure 1 and is designed to handle data coming from heterogeneous sources that differ in volume and type. In order to address this challenge, we apply a learning pipeline to each data source. Each pipeline includes several stages: data pre-processing, feature engineering, model training and evaluation. This approach is repeated for all data sources and results in a set of models optimised for their specific data source. To combine the predictions from the different models, we employ an aggregation function (i.e., the aggregation box in Figure 1) that takes into account the individual predictions (on the left of Figure 1). The aggregation step can go from simple data merging to format unification and to ensemble. This approach aims to ensure more robustness and accuracy of evaluations in the next steps of the workflow. To ensure the trustworthiness of predictions, we employ a stacked strategy to train a new model – the hyper model box in Figure 1 – that incorporates a new entity representing the explainability concept, starting from individual aggregated predictions. In the literature, several papers describe possible workflows of ML pipelines [2], [16]. The hyper model proposed here is based on the ML workflow defined in [2]. Nevertheless, the ML pipeline is extended with further steps needed to design explainability and interoperability. Moreover, we explicitly introduce the concept of domain knowledge as a dynamic entity that feeds the pipeline steps and evolves accordingly when new knowledge is brought out by the system, eventually yielding more accurate and advanced future predictions. The enhanced learning pipeline implementing the hyper model – see the hyper model box in Figure 1 – is composed of nine steps (the nine yellow enumerated boxes in the figure), one is new w.r.t. [2] (the green box) and four steps are extended with

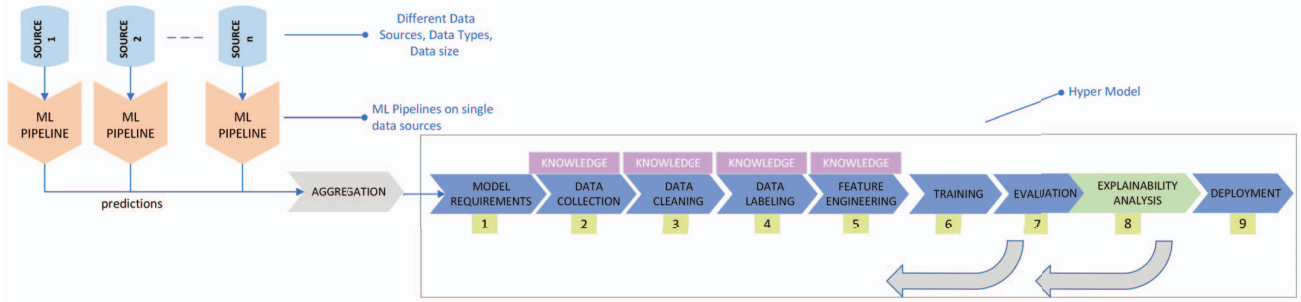

Fig. 1. Workflow for trustworthy ML predictions.

the domain knowledge. The idea is to integrate the domain knowledge by using expert rules and constraints, modelled using rule-based systems and decision trees. In medicine, expert rules can be used to constrain the predictions of the model based on specific medical guidelines. This knowledge evolves over time, embedding new discoveries obtained by data analysis or getting information from new clinical studies. For the sake of completeness and following the order in Figure 1, we briefly describe each step in the enhanced ML pipeline.

Model Requirements is guided by the requests of the systems' stakeholders. At the end of this step, all stakeholders should have a clear vision of the problem to be solved, what parts could be implemented by exploiting existing machine learning techniques, and a list of possible models to use to solve the problem.

In *Data Collection*, the stakeholders check the data and possibly look for extra data to be used in the analysis.

Data Cleaning is the process of removing or fixing incorrect, inaccurate, not well-formatted, duplicate, or incomplete information within the dataset.

In *Data Labeling*, ground truth labels are assigned to each element of the dataset. This is a very important step because it allows supervised learning methods to work at their best.

Feature Engineering refers to all the activities necessary to extract, select and transform initial data into informative features, aiming at simplifying and speed up data transformation and making learning techniques work well on the various tasks.

During the *Training*, models are trained and iteratively tuned on the pre-processed data.

At the end of the training, the *Evaluation* step is executed. All the stakeholders must evaluate the outputs of the model on different testing datasets according to pre-defined metrics. This step is very important as it allows for demonstrating (in a testing environment) that the chosen model meets the stakeholders' needs and other regulatory or ethics requirements of the system.

The subsequent stage, named *Explainability Analysis*, is necessary to comply with the explainability requirement. Here, two high-level alternatives can be chosen accordingly to the previous machine-learning method used in the pipeline. If the technique is *interpretable by design*, this stage focuses on measuring the provided interpretation's effectiveness. If the employed prediction method is a black-box one[1], then an explainability method must be used in this stage. Even though the type of explainer depends (see [6], [7], [15], [18] for a broader discussion), by the kind of data (i.e., images, graph, tabular, sequences, and temporal), the level of desired explanation (i.e., local or global), and by the desired kind of explanation (i.e., factual, counterfactual, prototypical, etc.), it must be noticed that the explainability can be achieved through a post-hoc model differently to the interpretability where the prediction model itself provides. For example, in the case of translational biomedical studies (such as those related to CVD), it is important to consider -omics networks. Here the reference predictor is based on Graph Neural Networks (GNN) as is discussed in [20], and the kind of explainability is of counterfactual type (the counterfactual explainers for GNN are mainly based on heuristic perturbation approaches as in [1], [4], [26]). Moreover, several evaluation frameworks were presented in the literature for some specific domains [21], [27], but we are far from having a "Swiss Army knife" for any circumstance. Despite the ongoing discussion and the efforts made to clarify the area (see [23]), there is no well-established process to integrate and evaluate the explainability methods into a broader software process. Once the *Explainability Analysis* successfully terminates, the *Deployment* will be executed.

To better explain how interpretability can be achieved, we refer to the CVD case. Let us assume that the outcome of all the (sub-)models might be merged into a further fuzzy interpretable hyper model, giving the final prediction of the CVD risk. The fuzzy rules of the hyper model will be based on antecedent variables corresponding to the predictions of the sub-models. Linguistic variables generation and model calibration (e.g., fuzzy sets parameters, consequent coefficients) will be both data-driven [14] and refined on the extended knowledge base. Hence, the hyper model will be the result of a holistic approach to CVD, able to take advantage of the domain experts' knowledge – clinical knowledge – to semantically link the prediction results obtained by different models.

[1]e.g., Artificial Neural Networks and its evolutions.

III. Conclusion and Future Work

Various studies in the field of cardiovascular diseases emphasised the importance of forecasting the cardiovascular risk using a variety of data sources, which requires the application of various ML techniques. This is actually a general characteristics of the health domain that leads to a new challenge in terms of system organisation and outcome combination to obtain a final prediction result. Additionally, in the health domain, it is also crucial to ensure interpretability or explainability of all predictions, given the heterogeneity of data sources, and their limited accessibility due to ethical and organisational concerns. In this study, we presented a comprehensive workflow that involves the use of multiple pipelines, with the number of pipelines matching the number of heterogeneous data sources. Each pipeline follows a learning process that takes a specific data source as input and provides a prediction as output. The outputs generated by all pipelines are then aggregated and used to train a hyper model, which is designed to provide highly accurate and robust predictions, implementing explainability and interoperability. We also incorporated medical knowledge into the learning pipelines to ensure prediction reliability. Such integration aims at enhancing forecast accuracy, by constraining the predictions on the basis of specific medical knowledge and guidelines. The approach presented here for the health domain can be also applied in other settings, provided that the domain expert knowledge is available.

Future work will involve applying our suggested design to an actual dataset, specifically the UK Biobank database, which houses patient-level information on more than 500,000 people. The database has a longitudinal design with substantial prognostic data on cardiovascular outcomes, observations at numerous time periods, and multi-level data comprising anthropometric, biochemical, imaging, and genetic data. The follow-up period is long (15 years). By using this dataset, we hope to illustrate the usefulness of our suggested strategy in a practical setting and offer insights into the integration of diverse data sources for research on cardiovascular disease.

References

[1] C. Abrate and F. Bonchi. Counterfactual graphs for explainable classification of brain networks. In *Proc. of the 27th ACM SIGKDD Conf. on Knowl. Disc. & Data Mining*, pages 2495–2504, 2021.

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*, pages 291–300, 2019.

[3] Natalie Arnold and Wolfgang Koenig. Polygenic risk score: Clinically useful tool for prediction of cardiovascular disease and benefit from lipid-lowering therapy? *Cardiovasc. Drugs Ther.*, 35(3):627–635, 2021.

[4] M. Bajaj, L. Chu, Z.Y. Xue, J. Pei, L. Wang, P.C.H Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. *Advances in Neural Inf. Proc. Sys.*, 34, 2021.

[5] Kevin M Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2):114–126, 2022.

[6] David L Buckeridge. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*. Springer, 2021.

[7] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *J. Artif. Int. Res.*, 70:245–317, 2021.

[8] Sarah Costantino, Francesco Paneni, and Francesco Cosentino. Ageing, metabolism and cardiovascular disease. *J. Physiol.*, 594(8):2061–2073, 2016.

[9] Changyu Deng, Xunbi Ji, Colton Rainey, Jianyu Zhang, and Wei Lu. Integrating machine learning with human knowledge. *iScience*, 23(11):101656, 2020.

[10] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Front. Comput. Sci.*, 14(2):241–258, apr 2020.

[11] Xin Luna Dong and Theodoros Rekatsinas. Data integration and machine learning: A natural synergy. *Proc. VLDB Endow.*, 11(12):2094–2097, aug 2018.

[12] Ben Ali et al. Implementing machine learning in interventional cardiology: The benefits are worth the trouble. *Frontiers in Cardiovascular Medicine*, 8, 2021.

[13] Jabbour et al. Combining chest x-rays and electronic health record (ehr) data using machine learning to diagnose acute respiratory failure. *Journal of the American Medical Informatics Association*, 29(6):1060–1068, 2022.

[14] Caro Fuchs, Simone Spolaor, Marco S. Nobile, and Uzay Kaymak. pyfume: a python package for fuzzy model estimation. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2020.

[15] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[16] Mark Haakman, Luís Cruz, Hennie Huijgens, and Arie van Deursen. Ai lifecycle models need to be revised: An exploratory study in fintech. *Empirical Softw. Engg.*, 26(5), 2021.

[17] Andreas Holzinger, Benjamin Haibe-Kains, and Igor Jurisica. Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, 46, 12 2019.

[18] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.

[19] Chayakrit Krittanawong, Hafeez Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp Johnson, Rachel Pinotti, Hongju Zhang, Scott Kaplin, Bharat Narasimhan, Takeshi Kitai, Usman Baber, Jonathan Halperin, and W.H. Tang. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, 10, 09 2020.

[20] L. Madeddu and G. Stilo. *Deep Learning methods in Network Biology*. 2022.

[21] Mario Alfonso Prado-Romero and Giovanni Stilo. Gretel: Graph counterfactual explanation evaluation framework. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, page 4389–4393, 2022.

[22] Giorgio Quer, Ramy Arnaout, Michael Henne, and Rima Arnaout. Machine learning and the future of cardiovascular care: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 77(3):300–313, 2021.

[23] Chakkrit Kla Tantithamthavorn and Jirayus Jiarpakdee. Explainable ai for software engineering. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1–2, 2021.

[24] Laura von Rueden et al. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[25] Zachary J Ward, Sara N Bleich, Angie L Cradock, Jessica L Barrett, Catherine M Giles, Chasmine Flax, Michael W Long, and Steven L Gortmaker. Projected u.s. state-level prevalence of adult obesity and severe obesity. *N. Engl. J. Med.*, 381(25):2440–2450, December 2019.

[26] G. P. Wellawatte, A. Seshadri, and A. D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical science*, 13(13):3697–3705, 2022.

[27] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022.

[28] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.