Data and Decision Fusion with Uncertainty Quantification in ML-based Healthcare Decision Systems

Supervisors : Sana Sellami (sana.sellami@univ-amu.fr) and Laure Berti-Equille (laure.berti@ird.fr)

Laboratory : LIS (Laboratoire d'Informatique et Systèmes)

Doctoral school : École Doctorale en Mathématiques et Informatique de Marseille (ED184)

1 Scientific Context

COVID-19 pandemic highlights the acute need to develop fast, on-demand therapeutics against pathogens and health threats. Traditional approaches to drug development are expensive, too slow to react to pandemics like COVID-19. AI and ML tools on the other hand have the potential to accelerate and transform this effort, enabling rapid, large scale search and identification of effective candidates for therapeutics and potentially transform our healthcare system from drug discovery to patient diagnostics and monitoring. To translate this potential to success, transdisciplinary research in computational and life sciences is needed. The PhD thesis proposal is indeed in the context of AI for Healthcare.

Data and Decision Fusion. The aim of this thesis proposal is to design solutions for data and decision fusion in healthcare systems. Data fusion is the process of integrating multiple data sources to generate more accurate information than that provided by any individual data source. Data fusion resolves conflicts from different data sources [DN09, DBS15] by identiying the best values among the conflicting ones. Although the problem is a pretty old one [LPL+08], it still receives a lot of attention from academia [BBM18, MTSP20]. On the other hand, decision fusion aims to fuse the decisions of various classifiers and getting an effective outcome.

Traditional data fusion techniques are based on probabilistic models [MJYP20]. Recently, machine learning models are becoming essential to analyze data or to predict critical events such as a disease or a stroke. Recently, several lines of work have addressed data fusion and decision fusion for health prediction of COVID-19 patients [GIH+22, DNS+21, HLC+22, KVW+17]. For example, in [GIH+22], a decision fusion method that combines three classifiers (random forest, gradient boosting, and extreme gradient) is proposed in order to improve the prediction of the COVID-19 patient health for early monitoring and efficient treatment.

Uncertainty Quantification. However, the main challenges of designing AI-based solutions for critical healthcare decisions are related to the lack of reliable annotated data (and the need of manual annotation for training the ML models) and also to the uncertainty quantification [Gal16].

High-stake decision processes require both robust methods and the ability to quantify uncertainty of predictive machine learning approaches to minimize the risks and provide the required scientific rigor. Nevertheless, traditional machine learning methods such as deep learning have difficulties in explaining their outputs, in enforcing physical/medical constraints, and in handling small noisy data sets [CN20]. Medical records or health monitoring systems for instance, may offer limited or low-quality data, ground truth is regularly unknown, benchmark data sets are conventionally rare, and finally, their problems usually have unknown terms and parameters. Despite the progress of incorporating uncertainty quantification techniques into recent approches, they are still underused for various reasons [APH+20, DGK21, KYH+20]. First, they are still a developing field with many unclear concepts not yet understood by the machine learning community [PMZ+22, HW21]. Likewise, machine learning communities have relied on simple data sets to validate uncertainty quantification methods, and they can only handle low-dimensional problems [RT20]. In this context, this thesis aims to develop new uncertainty quantification strategies for scientific machine learning for biomedical decision systems in the line of recent contributions in this field [GT20, KWKP20].

2 Objectives

This research will be focused in designing methods for data and decision fusion with uncertainty quantification in collaboration with our colleagues of the intensive care service (ICS) of APHM-Hôpitaux de Marseille.

- In the first 6 months, the candidate has to review the state-of-the-art in the domain of data and decision fusion based on Machine Learning and Deep Learning methods that are particularly relevant for critical heath monitoring applications. A review of the litterature on uncertainty quantification will be completed as well.
- At the end of the first year, solutions for resolving data inconsistencices and evaluating data sources reliability will be proposed in order to intergate data from various monitoring devices. Data coming from different sources may be incomplete, erroneous or out-of-date.
- During the second year, machine learning based methods for data and decision fusion will be designed and tested over multimodal data obtained from the intensive care service. A set of baseline methods for uncertainty quantification will be implemented and tested over the data analysis pipelines.
- During the third year, a new method capturing all the uncertainties generated from the data collection, data integration and data fusion to the ML-based decision will be designed, tested and validated with real-world use cases from the APHM-Hôpitaux de Marseille services.

References

[APH+20] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul W. Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *CoRR*, abs/2011.06225, 2020.

[BBM18] Laure Berti-Équille, Angela Bonifati, and Tova Milo. Machine learning to data management: A round trip. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*, pages 1735–1738. IEEE Computer Society, 2018.

[CN20] Joã o Caldeira and Brian Nord. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology*, 2(1):015002, dec 2020.

[DBS15] Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. Data fusion: Resolving conflicts from multiple sources. *CoRR*, abs/1503.00310, 2015.

[DGK21] Eoin Delaney, Derek Greene, and Mark T. Keane. Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions. *CoRR*, abs/2107.09734, 2021.

[DN09] Xin Luna Dong and Felix Naumann. Data fusion - resolving data conflicts for integration. *Proc. VLDB Endow.*, 2(2):1654–1655, 2009.

[DNS+21] Weiping Ding, Janmenjoy Nayak, H. Swapnarekha, Ajith Abraham, Bighnaraj Naik, and Danilo Pelusi. Fusion of intelligent learning for COVID-19: A state-of-the-art review and analysis on real medical data. *Neurocomputing*, 457:40–66, 2021.

[Gal16] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[GIH+22] Abdu Gumaei, Walaa N. Ismail, Md. Rafiul Hassan, Mohammad Mehedi Hassan, Ebtsam Mohamed, Abdullah Alelaiwi, and Giancarlo Fortino. A decision-level fusion method for COVID-19 patient health prediction. *Big Data Res.*, 27:100287, 2022.

[GT20] Biraja Ghoshal and Allan Tucker. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *CoRR*, abs/2003.10769, 2020.

[HLC+22] Zhongwei Huang, Haijun Lei, Guoliang Chen, Haimei Li, Chuandong Li, Wenwen Gao, Yue Chen, Yaofa Wang, Haibo Xu, Guolin Ma, and Baiying Lei. Multi-center sparse learning and decision fusion for automatic COVID-19 diagnosis. *Appl. Soft Comput.*, 115:108088, 2022.

[HW21] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, mar 2021.

[KVW+17] Rachel C. King, Emma Villeneuve, Ruth White, Robert Simon Sherratt, William Holderbaum, and William S. Harwin. Application of data fusion techniques and technologies for wearable health monitoring. *Medical engineering & physics*, 42:1–12, 2017.

[KWKP20] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics Data Analysis*, 142:106816, 2020.

[KYH+20] Georgios Kissas, Yibo Yang, Eileen Hwuang, Walter R. Witschey, John A. Detre, and Paris Perdikaris. Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4d flow mri data using physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 358:112623, 2020.

[LPL+08] Hyun Lee, Kyungseo Park, Byoungyong Lee, Jae Sung Choi, and Ramez Elmasri. Issues in data fusion for healthcare monitoring. In Fillia Makedon and Lynne Baillie, editors, *Proceedings of the 1st ACM International Conference on Pervasive Technologies Related to Assistive Environments, PETRA 2008, Athens, Greece, July 16-18, 2008*, volume 282 of *ACM International Conference Proceeding Series*, page 3. ACM, 2008.

[MJYP20] Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Inf. Fusion*, 57:115–129, 2020.

[MTSP20] Muhammad Muzammal, Romana Talat, Ali Hassan Sodhro, and Sandeep Pirbhulal. A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks. *Inf. Fusion*, 53:155–164, 2020.

[PMZ+22] Apostolos F. Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *CoRR*, abs/2201.07766, 2022.

[RT20] Rahul Rahaman and Alexandre H. Thiery. Uncertainty quantification and deep ensembles. *CoRR*, abs/2007.08792, 2020.