# Towards a Collaborative Platform for Advanced Meta-Learning in Health care Predictive Analytics

Milan Vukicevic[1], Sandro Radovanovic[1], Joaquin Vanschoren[2], Giulio Napolitano[3], Boris Delibasic[1]

[1] University of Belgrade, Faculty of Organizational Sciences, Jove Ilica 154, Belgrade, Serbia
[2] Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, Netherlands
[3] Bonn University, Germany

Modern medical research and clinical practice are more dependent than ever on multi-factorial data sets originating from various sources, such as medical imaging, DNA analysis, patient health records and contextual factors. This data drives research, facilitates correct diagnoses and ultimately helps to develop and select the appropriate treatments. The volume and impact of this data has increased tremendously through technological developments such as high-throughput genomics and high-resolution medical imaging techniques. Additionally, the availability and popularity of different wearable health care devices has allowed the collection and monitoring of fine-grained personal health care data. The fusion and combination of these heterogeneous data sources has already led to many breakthroughs in health research and shows high potential for the development of methods that will push current reactive practices towards predictive, personalized and preventive health care. This potential is recognized and has led to the development of many platforms for the collection and statistical analysis of health care data (e.g. Apple Health, Microsoft Health Vault, Oracle Health Management, Philips HealthSuite, and EMC Health care Analytics). However, the heterogeneity of the data, privacy concerns, and the complexity and multiplicity of health care processes (e.g. diagnoses, therapy control, and risk prediction) creates significant challenges for data fusion, algorithm selection and tuning. These challenges leave a *gap between the actual and the potential data usage* in health care, which prevents a paradigm shift from delayed generalized medicine to predictive personalized medicine [1]. As such, a platform for collaborative and privacy-preserving sharing, analysis and evaluation of health care data would drastically facilitate the creation of advanced models on heterogeneous fused data, as well as ensure the reproducibility of results, and provide a solid basis for the development of algorithm ranking and selection methods based on collaborative meta-learning.

In this work we present an extensions of the OpenML platform that will be addressed in our future work in order to meet the needs of meta-learning in health care predictive analytics: privacy preserving sharing of data, workflows and evaluations, reproducibility of the results, and rich meta-data spaces about both data and algorithms.

**OpenML.org** [2] is a collaboration platform which is designed to organize datasets, machine learning workflows, models and their evaluations. Currently, OpenML is not fully distributed but can be installed on local instances which can communicate with the main OpenML database using mirroring techniques. The downside of this approach is that code (machine learning workflows), datasets, experiments (models and evaluations) are physically kept on local instances, so users cannot communicate and share. We plan to turn OpenML into a fully distributed machine learning platform, which will be accessible from different data mining and machine learning platforms such as RapidMiner, R, WEKA, KNIME, or similar. Such a distributed platform would allow the ease of sharing data and knowledge. Currently, regulations and privacy concerns often prevent hospitals to learn from each other's approaches (e.g. machine learning workflows), reproduce work done by others (data version control, preprocessing and statistical analysis), and build models collaboratively.

On the other hand, meta-data such as type of the hospital, percentage of readmitted patients or indicator of emergency treatment, as well as the learned models and their evaluations can be shared and have great potential for the development of a cutting edge meta-learning system for ranking, selection and tuning of machine learning algorithms.

The success of meta-learning systems is greatly influenced by the size of problem (data) and algorithm spaces, but also by the quality of the data and algorithm descriptions (meta-features). Thus, we plan to employ domain knowledge provided by expert and formal sources (e.g. ontologies) in order to extend the meta-feature space for meta-learning in health care applications. For example, in meta-analyses of gene expression microarray data, the type of chip is very important in predicting algorithm performance. Further, in fused data sources it would be useful to know which type of data contributed to the performance (electronic health records, laboratory tests, data from wearables etc.). In contrast to data descriptions, algorithm descriptions are much less analyzed and applied in the meta-learning process. Recent results [3] showed that descriptions on the level of algorithm parts (e.g. initialization type and internal evaluation measures in clustering algorithms), could improve quality of meta-learning predictions, and additionally identify which algorithm parts really influenced the overall performance. Hence, we will include component based algorithm definitions as meta-features and allow their usage as predictors in meta-learning systems. The development of such a collaborative meta-learning system would address different challenging tasks in health care predictive analytics like early diagnostics and risk detection, hospital re-admission prediction, automated therapy control or similar with many potential stakeholders: patients, doctors, hospitals, insurance companies, among others.

## Acknowledgement

This research was supported by SNSF Joint Research project (SCOPES), ID: IZ73Z0–152415.

## References


[1] Olga Golubnitschaja, Judita Kinkorova, and Vincenzo Costigliola. Predictive, preventive and personalised medicine as the hardcore of horizon 2020: Epma position paper. *EPMA J*, 5(1):6, 2014.

[2] Joaquin Vanschoren, Jan N van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

[3] Milan Vukicevic, Sandro Radovanovic, Boris Delibasic, and Milija Suknovic. Extending meta-learning framework for clustering gene expression data with component based algorithm design and internal evaluation measures. *International Journal of Data Mining and Bioinformatics*, "In Press".