SeFNet: Linking Tabular Datasets with Semantic Feature Nets

Katarzyna Woźnica, Piotr Wilczyński, Przemysław Biecek

Warsaw University of Technology, Poland

Motivation

- Heterogeneity of tabular datasets — Tabular data dominate ML applications but differ greatly in their feature sets, making it difficult to establish meaningful relationships between datasets.
- Lack of semantic connections — Current ML and meta-learning methods struggle to transfer knowledge across datasets due to missing semantic information linking variables with similar meaning.
- Need for domain-informed automation — Automated Data Science and Informed ML require structured semantic context to effectively reuse prior knowledge and collaborate with domain experts.
- Healthcare as an ideal testbed — Medical data are rich but fragmented; integrating ontologies to link related features can unlock more generalizable and interpretable ML models.

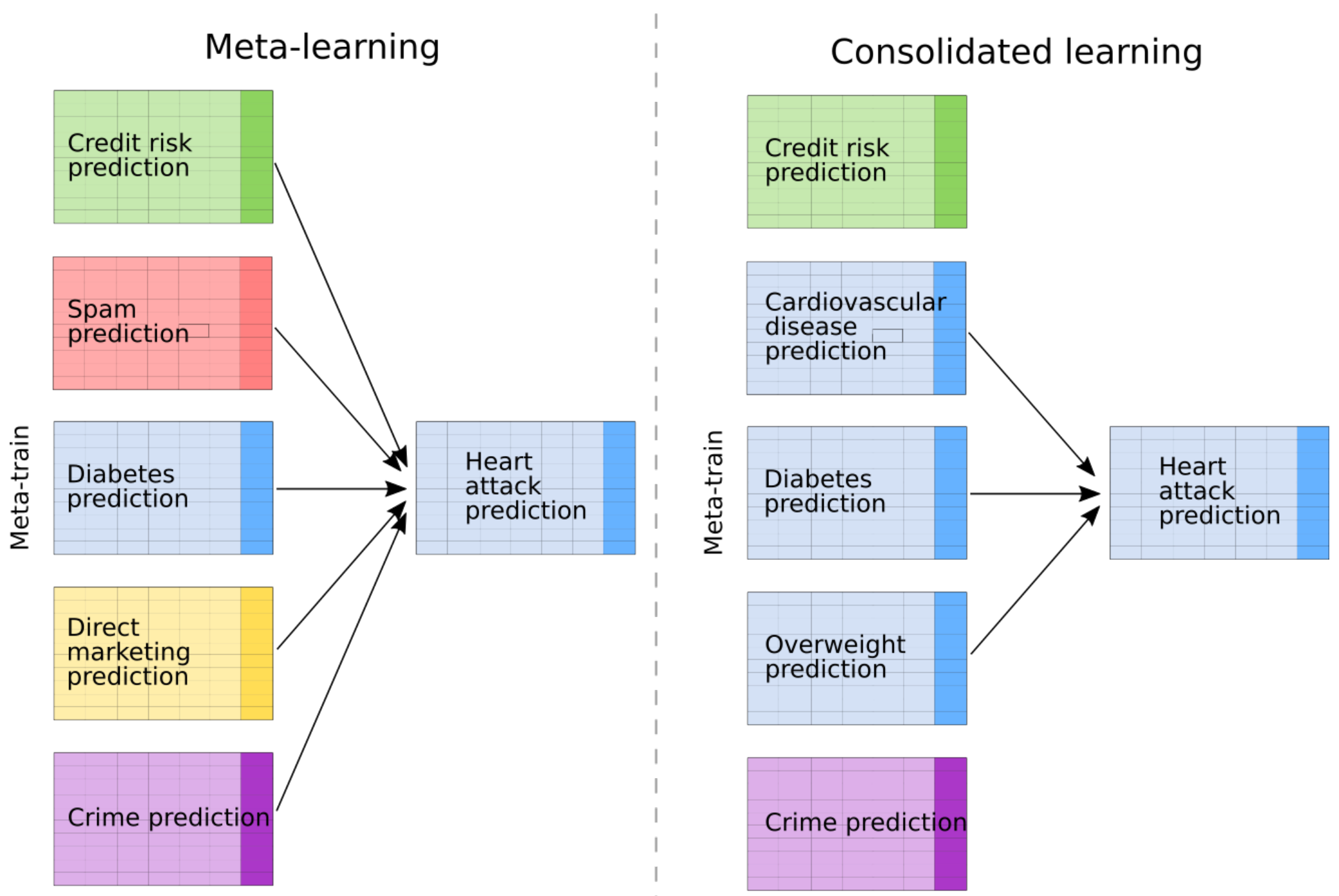

Figure 1: In current ML and Automated Data Science we treat tabular datasets as isolated entities.

Semantic Feature Net (SeFNet)

SeFNet is a methodology for connecting features across different tabular datasets using their semantic meaning. It links dataset variables to ontology concepts, creating a network that reveals relationships between features within and across datasets.

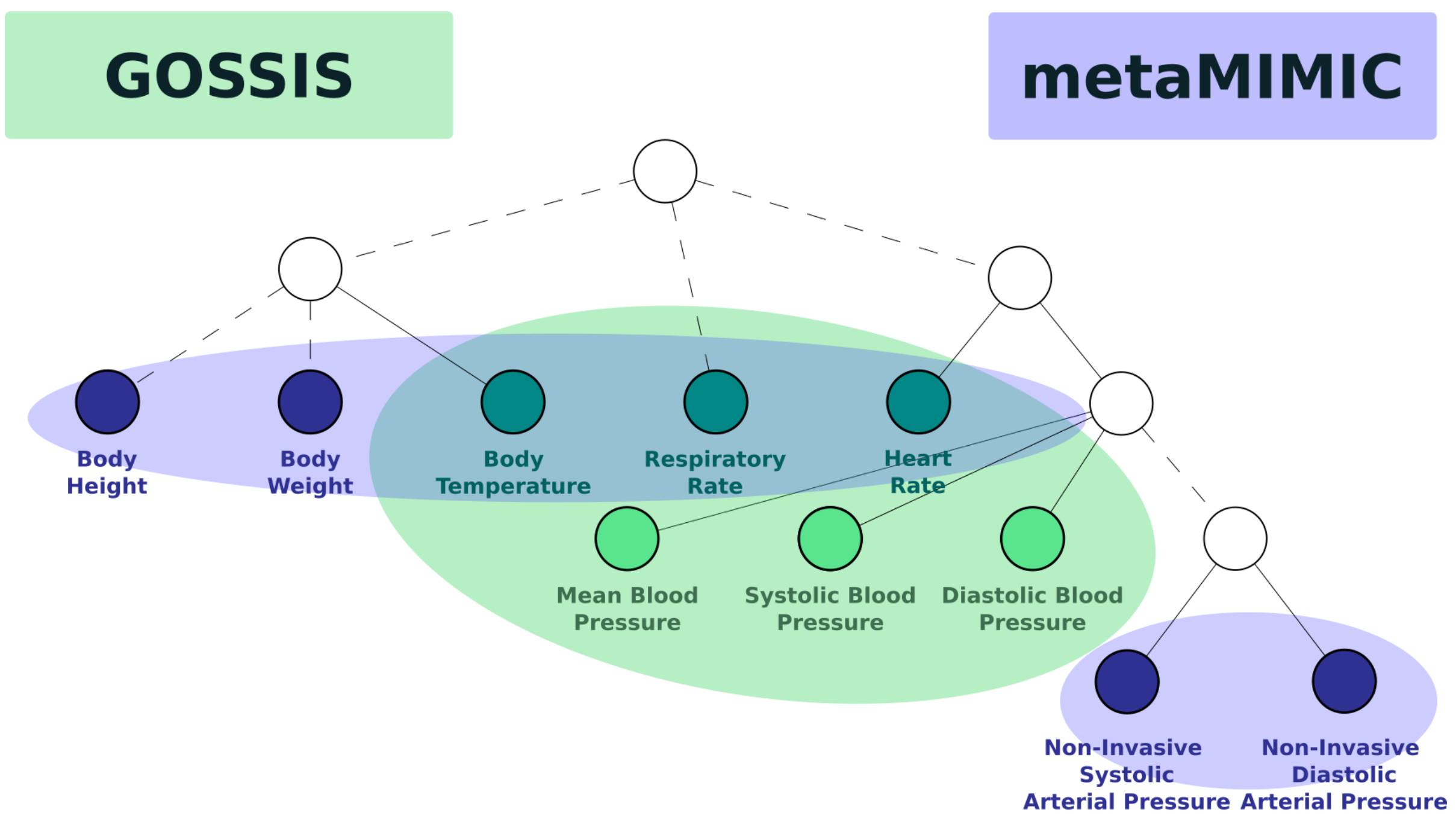

Figure 2: An example of subset of SeFNet for the two tabular datasets: *GOSSIS* and *metaMIMIC*. This resource encodes the structure of the relations between their features. Features are mapped on terms from the SNOMED-CT ontology.

To define SeFNet-based network and adapt it to a specific application, we need to specify three essential components:

1. **a set of tabular datasets** from a selected domain. These datasets serve as the basis for extracting and structuring features.
2. **an ontology** that covers the relevant concepts from the considered domain and the datasets. The choice of ontology is the responsibility of domain experts.
3. **a semantic similarity measure** consistent with the selected ontology.

See the full paper for more details

SeFNet Healthcare

In SeFNet-Healthcare, we collect 16 datasets, which can be divided into two groups due to the method of data collection and the specificity of features in each dataset: (1) datasets based on survey data (Survey), (2) group of datasets where predominant electronic health records (EHR).

ID	Dataset	Origin	Cat.	No.Feat.	No.Ann.
1	Cardiovascular Study	[Kaggle]	Survey	16	15
2	Diagnosis of COVID-19 (Subset)	[Kaggle]	EHR	19	18
3	Diabetes Health Indicators	[Kaggle]	Survey	22	21
4	Diabetes 130 US	[UCI,OpenML,Kaggle]	EHR	49	38
5	GOSSIS-1-eICU Model Ready	[PhysioNet]	EHR	68	60
6	Stroke Prediction	[Kaggle]	Survey	11	11
7	Heart Disease Indicators	[Kaggle]	Survey	22	21
8	Heart Disease (Comprehensive)	[OpenML]	EHR	12	11
9	HCV data	[UCI,OpenML,Kaggle]	EHR	13	13
10	Hepatitis	[UCI,Kaggle]	EHR	20	19
11	HiRID Preprocessed	[PhysioNet]	EHR	18	17
12	Pima Indians Diabetes	[OpenML,Kaggle]	EHR	9	8
13	ILPD	[UCI,OpenML,Kaggle]	EHR	11	11
14	Breast Cancer	[UCI,OpenML]	EHR	10	9
15	metaMIMIC	[Paper]	EHR	184	175
16	Thyroid Disease	[UCI,OpenML,Kaggle]	EHR	30	27

We employ the SNOMED-CT ontology to describe medical and demographic concepts encoded in variables. We annotate 216 different features included in the selected datasets, and up to 92% of them are annotated. Then we apply the proposed Dataset Ontology-based Semantic Similarity (DOSS) measure to the SeFNet-Healthcare database.

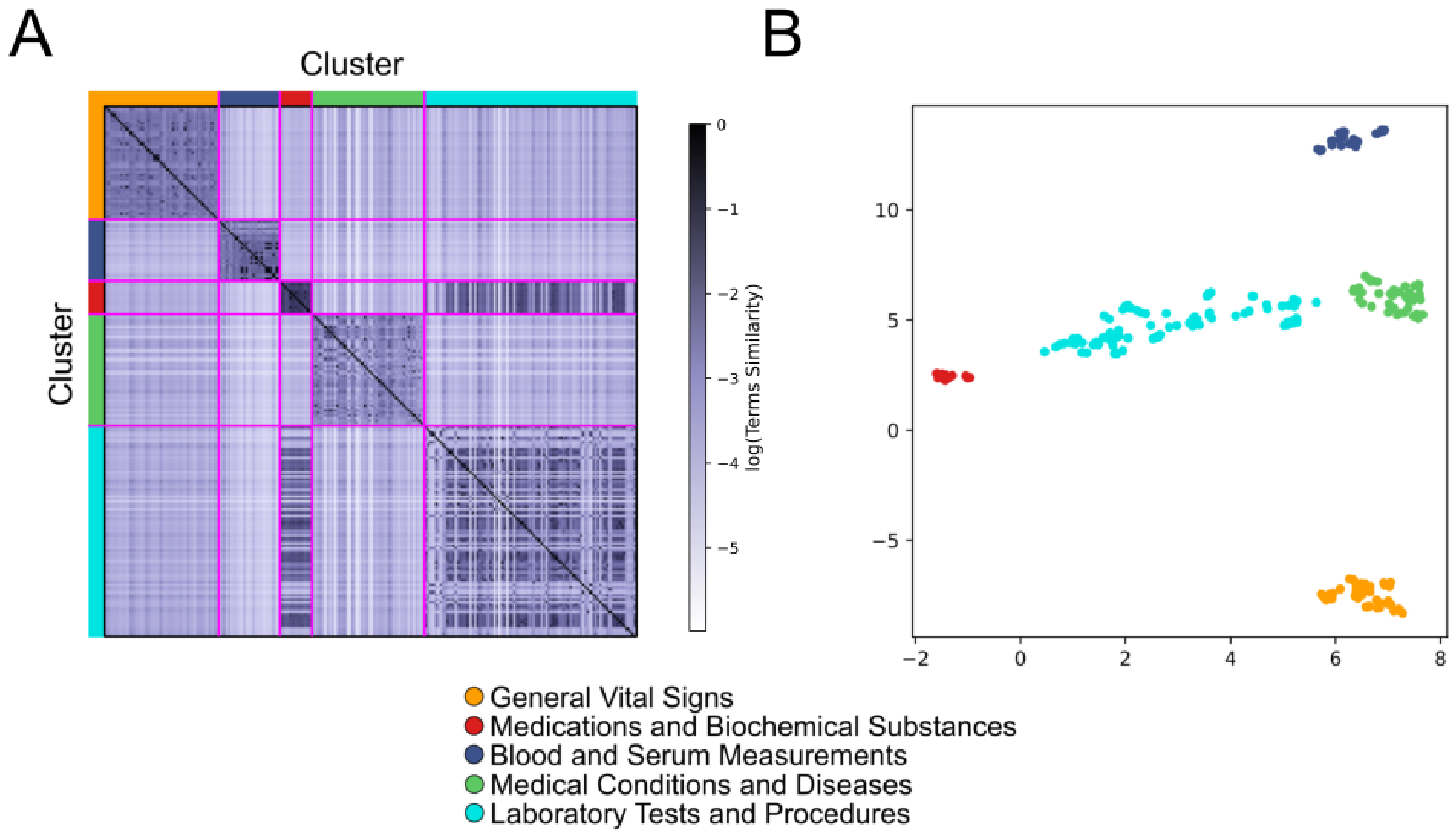

Figure 3: Similarity between the annotated features occurring in SeFNet-Healthcare. In panel A, we show the similarity matrix between each pair of features. The features' order corresponds to the clusters' belonging, illustrated in panel B. We can distinguish five groups of features named based on the high-level concepts.

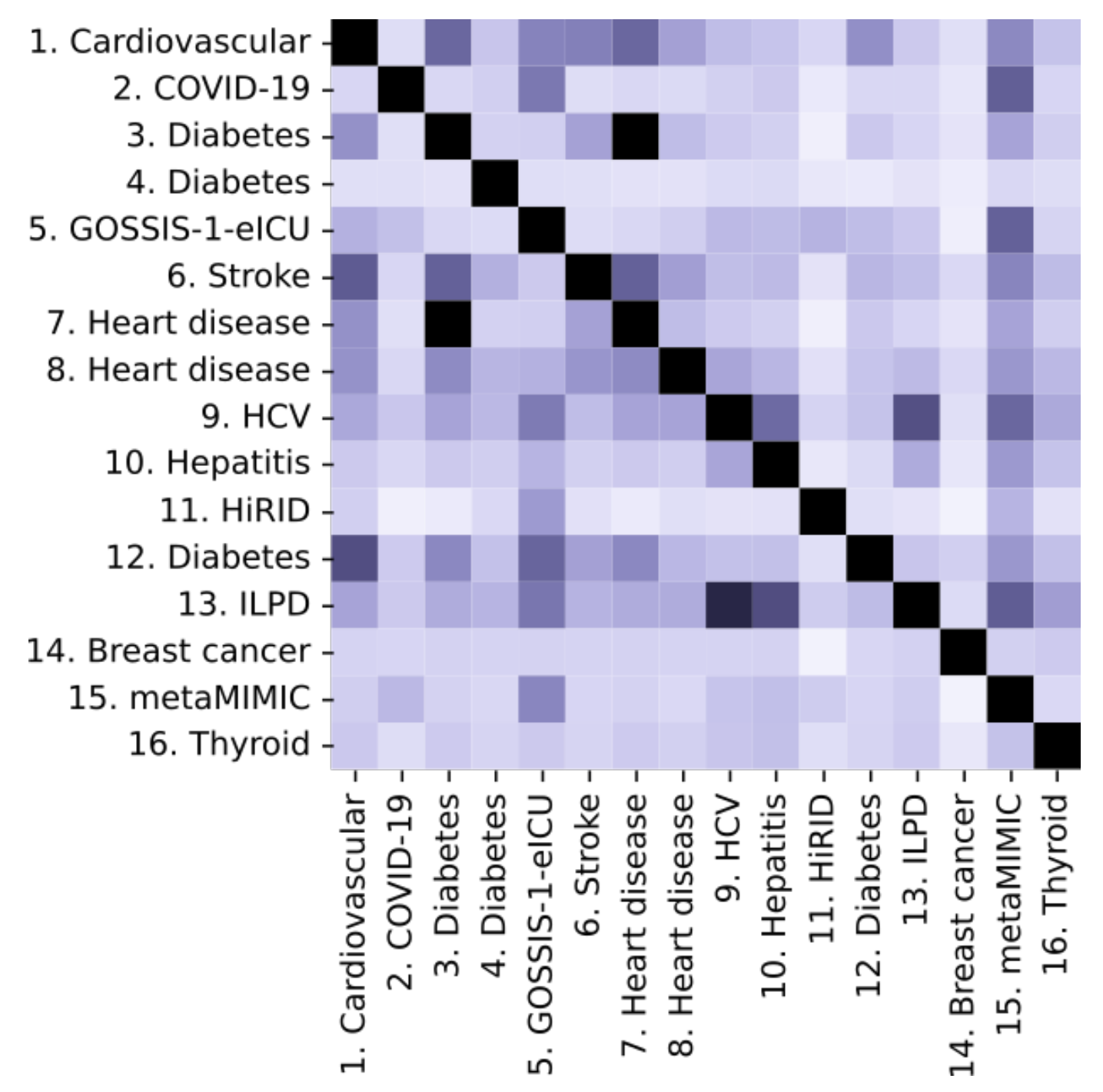

Figure 4: Matrix of DOSS values between any two annotated datasets in SeFNet-Healthcare. The matrix is not symmetrical since the defined DOSS measure does not have this property.

Application of SeFNet

- Assists data scientists by linking datasets through shared semantics, supporting collaboration with domain experts and exploration of related experiments.
- Improves meta-learning by introducing ontology-based dataset similarity, enabling more effective knowledge and hyperparameter transfer.
- Enhances explainable ML by embedding domain knowledge, allowing interpretable models and explanations at different levels of abstraction.

Contact

✉ woznicakatarzyna22@gmail.com