FeatureCloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Deliverable D4.7
"Integrated evaluation results for classifier ensembles in a federated approach"

Work Package WP4
"Supervised Federated Machine Learning"

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number: 826078				**Acronym: FeatureCloud**
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Christos Maramis, Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Jan Baumbach, University of Hamburg (UHAM)			
Deliverable	D4.7 - Integrated evaluation results for classifier ensembles in a federated approach			
Work Package	WP4 - Supervised Federated Machine Learning			
Date of Delivery	**Contractual**	30/06/2023 (M54)	**Actual**	29/06/2023 (M54)
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	03 MUG			
Responsible Author(s)	Andreas Holzinger (AH), Anna Saranti (AS), Bastian Pfeifer (BP), Alessa Angerschmid (AA), Heimo Müller (HM), Richard Röttger (RR), Anne-Christin Hauschild (ACH), Dominik Heider (DH), Sandra Clemens (SC)			
Keywords	Classifier Ensembles, Federated Learning, Graph Neural Networks			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	15/04/2023	First draft	AH (MUG), BP (MUG), AS (MUG), AA (MUG), SC (UMR)
V0.2	22/04/2023	Comments	HM (MUG), RR (SDU)
V0.3	07/05/2023	Revisions	AH (MUG), BP (MUG), AS (MUG)
V0.4	15/05/2023	Comments	DH (UMR), ACH (UMR)
V0.5	25/05/2023	Revisions	AH (MUG), BP (MUG), AS (MUG)
V0.6	01/06/2023	Comments	DH (UMR), SC (UMR)
V0.7	06/16/2023	Final draft ready	AH (MUG), BP (MUG), AS (MUG), SC (UMR)
V0.8	16/06/2023	Reviewed Version	Internal reviewers DH & ACH (UMR)
V0.9	28/06/2023	Final revision	AH (MUG), BP (MUG), AS (MUG)
V1.0	29/06/2023	Final version	AH (MUG), BP (MUG), AS (MUG)

Actual effort in person-months (PMs)

Contributor (name and institution)	Invested resources (deliverable)	Overview of contributions
Andreas Holzinger (MUG)	0.8 PM	Requirements, First draft, draft, final version
Anna Saranti (MUG)	0.2 PM	Requirements, First draft, draft, final version
Bastian Pfeifer (MUG)	0.2 PM*	Requirements, First draft, draft, final version
Alessa Angerschmid (MUG)	0.1 PM	Requirements, First draft, draft, final version
Sandra Clemens (UMR)	2.4 PM	Requirements, First draft, draft, final version
Domink Heider (UMR)	0.1 PM	Comments, review
Anne-Christin Hauschild (UMR)	0.1 PM*	Comments, review
Heimo Müller (MUG)	0.1 PM	Requirements, comments
Richard Röttger (SDU)	0.1 PM	Requirements, comments

*This person dedicated a certain amount of time to FeatureCloud, but received no salary from the FeatureCloud budget (e.g. Professor, Supervisor, Master/Bachelor student, Intern etc.).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Table of Contents

1 Table of acronyms and definitions 5
2 Objectives of the deliverable based on the Description of Action (DoA) 6
3 Executive Summary 6
4 Introduction (Challenge) 7
5 Methodology (Theoretical Aspects) 8
5.1 Explainable AI on Graph Neural Networks (GNNs) 8
5.2 Federated Learning with GNNs 9
5.3 Knowledge Graphs 12
5.4 Human-in-the-loop 13
6 Results 14
6.1 Knowledge Graph 14
6.2 Disease Subnetwork Detection 14
6.3 Explainability 14
6.4 Explainable Federated Learning with GNNs 15
6.5 Centralised Federated Learning with xAI and Human-in-the-Loop 17
7 Conclusion 20
8 References 21

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Table of acronyms and definitions

AI	Artificial Intelligence
CLARUS	InteraCtive ExpLainable PlAtform for GRaph NeUral NetworkS
D	Deliverable
FC	FeatureCloud
FL	Federated learning
GND	Gnome Design SRL
GNN	graph neural network
HFL	Horizontal FL
IID	independent, identically distributed
LRP	Layer-wise Relevance Propagation
MS	Milestone
MUG	Medizinische Universität Graz
Patients	In this deliverable, we use the term “patients” for all research subjects. In FeatureCloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus, to increase readability, we simply refer to them as “patients”.
PPI	protein-protein interaction
RI	Research Institute AG & Co. KG
SDU	Syddansk Universitet
UI	User interface
UHAM	Universität Hamburg
UMR	Philipps Universität Marburg
VFL	Vertical FL
WP	Work package
xAI	explainable Artificial Intelligence

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

2 Objectives of the deliverable based on the Description of Action (DoA)

The objective of this deliverable is based on the DoA, which incorporates mostly aspects from **Objective 5** “to design, develop and evaluate end-user centred interfaces to enable the interaction of humans with the algorithms developed, and to enable to re-enact and to re-trace in order to explain and understand the results in the context of a medical problem”.

In this technical report D4.7 “Integrated evaluation results for classifier ensembles in a federated approach”, MUG presents the work carried out within **Task 6 “Ensemble Learning”** to develop and evaluate classifier ensembles for federated learning, e.g., random forests and ensembles of classifier chains, conducted to achieve Objective 5, and to complete **MS29 “Evaluation of classifier ensembles for federated learning”**. In this report, we also describe CLARUS which is short for “An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks”. The evaluation of CLARUS can be found in D 4.8 “Explanation strategies, i.e. post-hoc vs. ante-hoc approaches”.

3 Executive Summary

MUG has achieved all goals of objective 5, task 6, fulfilled MS 29 and made all algorithms publicly available on GitHub. The main contributions are documented in the following scientific publications, all of which we made gold open access, freely available to the international research community:

In (Holzinger et al., 2021), we described explainable graph-neural networks and the human-in-the-loop with counterfactual explanations. This paper has been extremely well received by the international research community and is a highly cited paper in the Science Citation Index of the Web of Science. The visions and ideas that we presented have since been largely implemented within WP4 and we have them well documented in the following publications: In our work (Pfeifer et al., 2022), we mask deep neural network learning with a protein-protein interaction (PPI) network. Each patient represents a PPI graph whose nodes can be enriched with patient-specific multi-modal genomic features (e.g., mRNA, DNA methylation). Subsequently, the classification has been made explainable, i.e. those subnetworks are detected that were relevant for the classification (“disease subnetworks"). These subgraphs are the so-called “local spheres” as described in (Holzinger et al., 2021) and in D4.5 “Framework for Local Sphere privacy aware federated learning on graphs”. To ensure a representative basis for comparison with the above methodology, the detection of subnetworks was implemented using a random forest (Pfeifer et al., 2022a), because random forests are highly relevant in medicine due to their better re-traceability, hence interpretability and explainability. Here too, the learning process was masked by a knowledge graph. Consequently, in our next work (Pfeifer et al., 2023) we enabled federated learning using the methods mentioned above. Here, the knowledge graph is decomposed (partitioned) into relevant subnetworks using explainable AI methods (Holzinger et al., 2022a), based on which an ensemble classifier is constructed. This ensemble classifier can be efficiently learned in a federated way. At the same time the resulting classifier in the form of an ensemble is more interpretable, because the classification performance of specific parts of the graph can now be inspected more efficiently.

Finally, we have designed, developed and evaluated an end-user-centred interface to enable an expert end-user to re-enact and to re-trace the results (Beinecke et al., 2022). In this tool called “CLARUS” the domain expert can analyse the detected subnets, manipulate them (e.g., delete nodes and add nodes) in order to gain insight into the network behaviour, and could finally be integrated back into the ensemble classifier. The evaluation and usability study performed is reported in Deliverable D4.8 "Explanation strategies, i.e. post-hoc vs. ante-hoc approaches". A summary about the main results of task 6 has been published in (Holzinger et al., 2023).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

4 Introduction (Challenge)

One of the main challenges encountered in our task was to mask deep neural network learning with a protein-protein interaction (PPI) network. In the context of this task, “masking”' refers to incorporating a domain-knowledge graph (specifically, a PPI network) into a deep neural network for classification. The nodes and the edges of the PPI network are respectively added to the input layer of the neural network and used to enrich the features of the data processed by the deep neural network (Pfeifer et al., 2022b). This is important because features are key for (machine) learning, but also key for (human) understanding and explanation, therefore consolidated features are more accurate and robust and fosters trust in the results. Trustworthiness is ensured by both explainability and robustness (Holzinger, 2021).

The first general challenge was that even the most powerful learning methods suffer from the fact that it is difficult to retrace, to interpret and thus to explain to a human expert (both the domain expert and the machine learning expert) why a certain result was obtained.

The second general challenge was that the currently best algorithms lack robustness. Even the smallest perturbations in the input data can dramatically affect the output, leading to completely different results. This is of great importance in virtually all critical domains where we suffer from poor data quality, i.e., where we do not have available the independent, identically distributed (IID) data we would need for ideal learning. Therefore, robust and explainable, hence trustworthy solutions are essential for medical applications (Holzinger et al., 2022b).

Here, the human-in-the-loop can be of great help, because the human expert can bring in (sometimes - of course not always) previous knowledge, experience and common sense (Holzinger et al., 2023), (Mosqueira-Rey et al., 2023), (Wu et al., 2022), (Saranti et al., 2022), (Holzinger et al., 2019a), (Holzinger et al., 2019b), (Lage et al., 2018), (Holzinger, 2016), (Kieseberg et al., 2015).

These challenges have serious consequences for EU citizens and organisations, because AI - as currently known from the daily press - is a much-discussed topic in the European Union. Although AI systems currently remain under relatively controlled environments, it is expected that AI will be used on a much larger scale in the coming years. Therefore, the European Commission has committed to establishing principles for trustworthy AI and safe use of AI in the digital society (Hamon et al., 2020). Consequently, after our first step in making the classification robust, in our next step our classification has been made explainable, i.e. those subnetworks are detected that were relevant for the classification. We call them “disease subnetworks” and the subgraphs “local spheres", see our previous reports and see (Holzinger et al., 2021) and (Malle et al., 2017).

In order to guarantee a representative baseline comparison to the above methodology, the subnetwork detection was realised by means of a random forest (Pfeifer et al., 2022a). Here, too, the learning process is masked by a knowledge graph. Random forests are particularly relevant in medicine due to their good interpretability.

In the work of Pfeifer et al., 2023, we enabled federated learning with the methods mentioned above. Here, the knowledge graph is divided into relevant subnetworks using explainable AI, based on which an ensemble classifier is constructed. This ensemble classifier can be efficiently learned in a federated way. In addition, a user interface was developed (Beinecke et al., 2022) that allows a domain expert to analyse and manipulate the detected subnetworks (delete and add nodes) and finally reintegrate them into the ensemble classifier.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

5 Methodology (Theoretical Aspects)

In our task we followed four central topics from (Holzinger et al., 2021): (i) explainable AI on graph neural networks (GNNs), (ii) federated learning and multi-modality, (iii) knowledge graphs, and (iv) human-AI interaction to enable to re-enact and to re-trace to explain and understand the results in the context of a medical problem. Consequently, we focused on all of these topics on the medical problem of precision medicine (Holzinger et al., 2014), (Rost et al., 2016), (Johnson et al., 2021).

5.1 Explainable AI on Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) have revolutionised neural network frameworks with their ability to function with point-cloud data sets forming graphs, aka networks, and proteins' biological functions are defined by geometrical network structures (Sverrisson et al., 2021). GNN's employ learnable functions to extrapolate features and patterns from the graph structure, thereby facilitating various tasks such as node classification, graph classification, and link prediction (Wu et al., 2020). GNNs have shown significant success, particularly in their ability to incorporate domain-knowledge graphs, thereby enhancing the interpretability and explainability of Deep Learning generally (Holzinger et al., 2021). Federated applications of GNNs appear to be organically emerging in several and diverse areas, including distributed sensor networks for traffic surveillance, distributed social media applications, and inter-hospital collaborations for addressing complex medical challenges. In the current era, characterised by an abundance of big data, there are twin challenges: escalating graph dataset sizes and increasingly complex GNN architectures.

These challenges necessitate 1) efficient and 2) privacy-conscious mechanisms for both information exchange and computation. An intriguing facet of the current digital medical landscape is that the entities involved in communication, be they servers or clients, can themselves be represented as graphs. It has been demonstrated that GNN architectures can support federation in this context (Liu et al., 2022), (Fu et al., 2022), (Zhang et al., 2022a). Consequently, the full potential of GNNs extends beyond their primary role in data analysis, and includes significant contributions to enhancing federated solutions.

Unfortunately, as with neural networks in general, GNN results are not easy to understand and interpret. To address this shortcoming, intensive work is currently being done worldwide on GNN explainability methods. Examples of these are GNNexplainer, PGExplainer, and GNN-LRP, which we carefully examined in the progress of our work:

GNNExplainer (Ying et al., 2019) provides local explanations for predictions of any graph-based model. This can be used for both node classification and graph classification.

PGExplainer (Luo et al., 2020) is a parameterized modification of the GNNexplainer. Unlike the GNNexplainer, the PGExplainer provides model-level explanations that we find useful for graph classification tasks.

GNN-LRP (Schnake et al., 2020) is derived from higher-order Taylor expansions based on layer-wise relevance propagation (LRP), which is a standard methodology for many applications (Lapuschkin et al., 2016). GNN-LRP explains predictions by extracting paths from the input to the output of the GNN model that contribute most to the prediction. These paths correspond to walks on the input graph. GNN-LRP was developed for node-level explanations and has been modified to work in a special arrangement for graph classification (Chereda et al., 2021). Of particular interest is the work presented using a method called CF-Explainer (Lucic et al., 2022). Here, explanatory factors can be revealed using counterfactuals.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

GCExplainer (Magister et al., 2021) is the first GNN explainer that recognizes the learned concepts of a GNN. The main idea is to perform clustering after the last aggregation layer and assume that each of the clusters corresponds to a human recognizable concept. The user thus has the possibility to parameterize the explanation process by the number of clusters and the size of the neighbourhood of the explained component. This approach involves the human-in-the-loop and thus has been shown to achieve good concept purity and completeness (robustness) (Bodén et al., 2021), (Holzinger et al., 2019a), (Lage et al., 2018).

Moreover, it forms the basis for recent work that makes GNNs explainable by design (explainability) by first learning the concepts and then making a concept-based prediction based on that learning. This is beneficial in that the opaque reasoning of Graph Neural Networks leads to loss of confidence. Existing explanatory tools for graph networks attempt to address this problem by providing post-hoc explanations (see our report D4.8), but they fail to make the model itself more interpretable. To address this gap, a concept encoder module is used for concept discovery for graph networks. This approach makes graph networks explainable by first discovering graph concepts and then using them to solve the task (Magister et al., 2022).

For medicine, something like this is ultra-important because such methodology can facilitate the discovery of disease-causing regions in networks and help uncover a subset of candidate features organised in disease-relevant network modules that are otherwise difficult or impossible to find.

This is exactly where the human-in-the-loop concept helps, as interaction with explanations and the incorporation of conceptual knowledge can further improve the learning algorithm.

5.2 Federated Learning with GNNs

Federated learning (FL) is an ML approach in which the training data is distributed across multiple devices or locations, and the local model training process is performed locally on each device or location (Malle et al., 2017).

FL is of course useful in scenarios where the data is sensitive, private, or subject to regulatory constraints, such as medical records or financial transactions. Instead of centralising the data and running the model training process on a single server or cloud platform, federated learning allows the data to remain on individual devices or locations, and only the model updates are transmitted for aggregation. This preserves the privacy and security of the data and reduces the risk of data breaches or leaks.

FL should not be mixed up with purely decentralised learning, where local models do not automatically contribute to each other apart from manually sampling the models and updating the hyperparameters (Bellavista et al., 2021); and also not with collaborative learning in various forms, where the goal is to share information about internal model building between the involved parties in a peer-to-peer manner but keep the local training data confidential.

The local model training process influences and is influenced by a global collaborative training that involves the aggregation of predefined entities (weights, performance indicators, embeddings, Explainable AI information) of the local models for the construction of an aggregated global model that is shared and used among the clients.

The characteristics of this model are based on the exchanged information, the aggregation function (which can be fixed or learned), the global loss function, and the topology of the inter-client network with or without a central server. In Figure 1, two main approaches for distributed collaborative Federated Learning with GNNs are presented (see also Table 1).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

On the one hand, there is the centralised approach with one central server and several clients that contain local GNNs trained in isolation each with their local datasets. The central server is there to aggregate predefined entities of the local GNN models and send them back to the local clients so that they can use an aggregated GNN.

In this case, the central server has the responsibility of containing information about the clients (their identifier and other meta-information) and the aggregation; if the server is compromised by hardware failure or attack, then the federation cannot be accomplished. Clients can communicate with each other in principle, but not in the context of federation.

The second approach is decentralised, where no central server exists, nor is there a client that has different functionality from the others. All clients can exchange information with each other in general; nonetheless, the topology of their network is not always fully-connected. Each of the clients must contain information about the topology of the client network and ideally know when a client is compromised, deleted or added to the network; this can be achieved with self-attention during the global training (Liu et al., 2022). The decentralised approach is not going to be affected as much if one of the clients has a hardware failure or is compromised, but that means that this deficiency might also be unnoticed for a longer period.

Both those approaches have advantages and disadvantages with respect to privacy and non-IID-distributed data among the clients, which can be summarised in the following table:

Table 1. Advantages and disadvantages of centralised vs. decentralised Federated Learning for GNNs (Liu et al., 2022).

Scenario	Sub-Scenario	Advantages	Disadvantages
Centralized FedGNNs	Server-Side GNNs Training	• The server has more flexibility in FL aggregation to relieve non-IID problem.	• Difficult to prove the convergence. • The server requires high computation costs when the inter-client graph is large. • Imprecise inter-client graph deteriorates performance.
	Client-Side GNNs Training	• Relieve the non-IID problem between clients.	• Shared inter-client graph may leak privacy. • Imprecise inter-client graph deteriorates performance.
Decentralized FedGNNs	Decentralized FL Aggregation	• Relieve the non-IID problem with personalized local models in clients. • Do not require a central server.	• Shared models may leak privacy between neighbors. • High communication cost between clients. • Clients with higher centrality are vulnerable. • Need to re-train the model when new clients join.

Federated GNN use cases consist - as also described in section (4) - of the distribution of nodes, features (Hu et al., 2018) or subgraphs; this also depends on the task that is accomplished. It has been known for some time that features for one modality are learned better when multiple modalities are present at the time of feature learning. In multimodal learning, information is from multiple sources. Often, several different modalities contribute to a result. We are motivated by (Holzinger et al., 2019c), (Acosta et al., 2022), (Ektefaie et al., 2023), which brings us directly to knowledge graphs.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

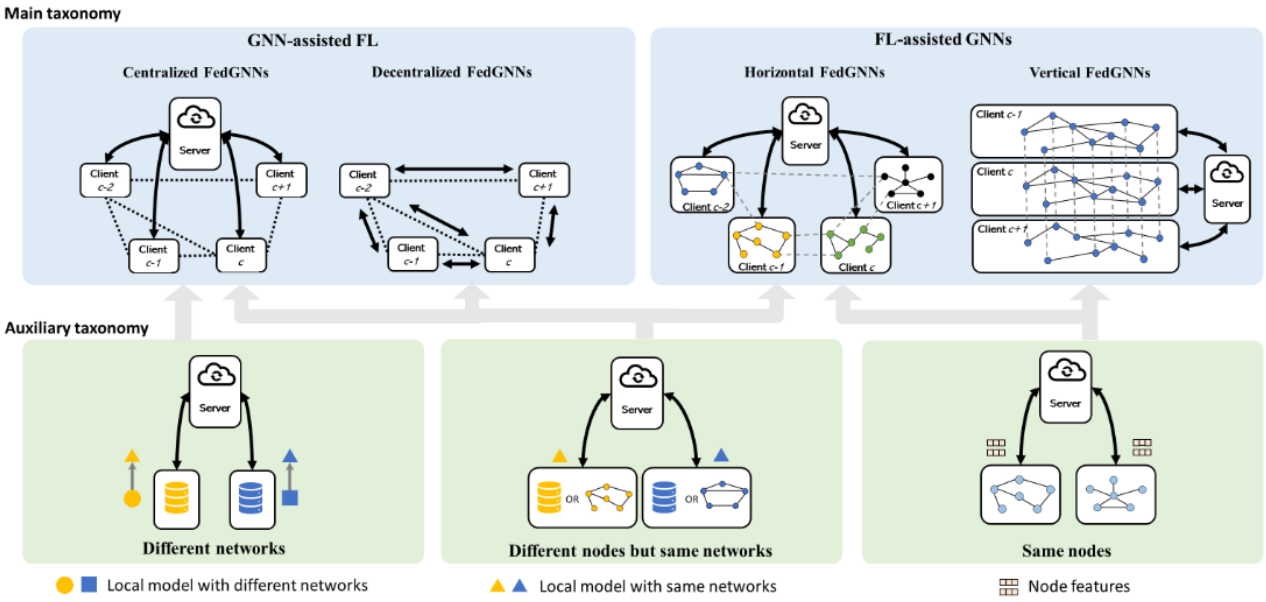

Figure 1. Federated centralised vs. federated decentralised GNNs (Liu et al., 2022).

Federation itself has evolved to be a broad topic; although the main principles are firm, different implementations realise the same goals. What is similar in all instantiations is that there is data isolation to some degree and that the information being exchanged should be minimal and privacy-preserved (i.e. encrypted). Furthermore, the IDD scenario is rather the exception than the norm; several frameworks need to simulate it before the actual deployment (Ortega et al., 2018). Nonetheless, collaboration has proven to be fruitful in most cases, since no one dataset contains all representative information about a task and ML solutions lack the ability of systematic generalisation and out-of-distribution (OOD) prediction even when trained with rich and diverse datasets.

In the more concrete case of Federated GNN, there are mainly three possibilities (He et al., 2021), as also shown in Figure 2. In the graph-level FL, each client has its graph dataset and potentially also a GNN. In the subgraph-level FL, each of the clients has one part of the graph and in the node-level FL nodes of one graph are distributed among clients.

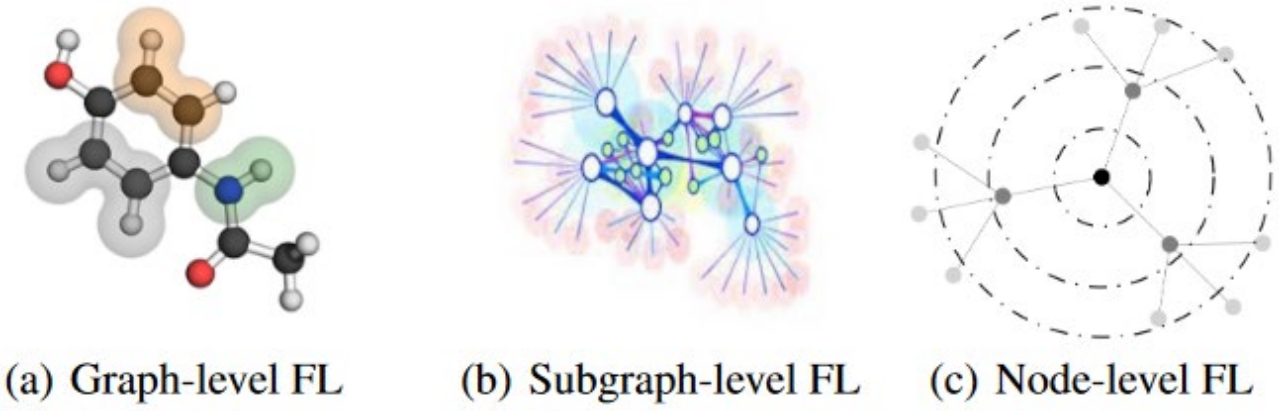

Figure 2. Three settings of GNN federation (He et al., 2021)

This is following the principles of Horizontal FL (HFL) and Vertical FL (VFL). In the first case, the features of the graphs of all clients are quite similar, but their sample characteristics (data distribution) differ substantially. The opposite occurs in the second case. Both of them are viable scenarios of FL and need to be addressed either with centralised or decentralised FL. In the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

centralised strategy, it is typical that there are several synchronous or asynchronous events containing parts of the dataset (as shown in Figure 1, and one server is responsible for the federation (which is also called aggregation). In the decentralised case, many clients exchange information with each other; this is more robust as far as privacy attacks are concerned but has substantial communication and organisational overhead.

Regardless of the client-server topology, there is an inter-client graph that can either be known a priori or can be discovered through self-attention mechanisms (this can also be helpful in the case where new clients enter the client-server topology). There are several ways to implement federation for all four combinations of FL possibilities, which in some centralised cases even needs alternating local and global optimization. Each of them has different convergence guarantees (if any) and individual countermeasures for the expected privacy attacks.

What is more, the client-server topology or the inter-client graph (in the decentralised case), have also a graph structure. Although the basic aggregation procedure (both in centralised and decentralised versions) is the averaging weighted by the number of samples in each client (McMahan et al., 2017), there are other types of more sophisticated federations. Inspired by the ideas of learned aggregation functions of GNNs themselves, more sophisticated handling of weights and biases were invented, as seen in Figure 3.

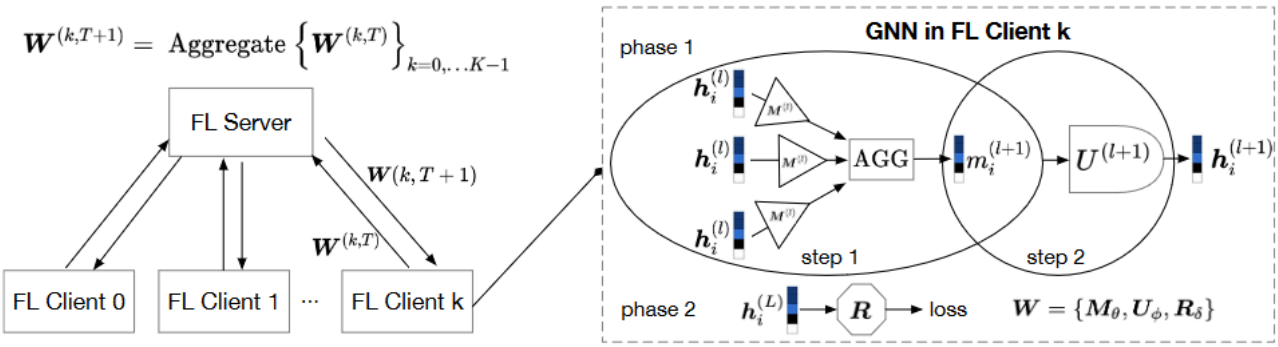

Figure 3. FedGraphGNN learned aggregation procedure for federation, as presented in (He et al., 2021).

The whole aggregation functionality can be solved by a GNN that takes as input the topology along with weights, gradients or even embeddings (provided they've been sent encrypted) as node and edge features, and returns new parameters in each federation round. Whether the topology is known beforehand or is changing, GNN-assisted FL is an emerging area of research (Liu et al., 2022). In many real-world application cases, the assumption is that clients that are "close" have similar data, thereby their local GNNs will probably also have similar parameters. In this case, the result of the trained GNN is practically the federation function, which goes beyond FedAVg (McMahan et al., 2017) and (Asad et al., 2020) possibilities.

5.3 Knowledge Graphs

Knowledge graphs (KG) are a type of database that represents knowledge in a structured, interconnected format, using a graph-based data model. It typically consists of a set of nodes (also called entities) that represent concepts or things, and a set of edges (also called relationships or properties) that connect the nodes and represent the connections or interactions between them. Many phenomena from nature can be represented in graph structures, whether at the molecular level (e.g. protein-protein interaction) or at the macroscopic level (e.g. social networks) and various methods from network science (Dehmer et al., 2017) and computational topology (Holzinger, 2014) can be applied. Some of the most successful application areas of machine learning and knowledge extraction in recent years can be seen as learning with graph representations (Veličković, 2023).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

In a knowledge graph, each node and edge can have additional attributes or metadata associated with it, providing additional information or context about the node or edge. This metadata can include labels, descriptions, categories, or other semantic information. Knowledge graphs are often used to represent information from diverse sources and domains in a multi-modal manner. They can be used to represent both factual knowledge (such as the properties of objects or events) and conceptual knowledge (such as the relationships between abstract concepts). Knowledge graphs are also used as a foundation for various applications, such as natural language processing, semantic search, recommendation systems, and data integration. They enable efficient querying and reasoning about complex, heterogeneous data, as well as support the development of intelligent agents that can reason and learn from the knowledge represented in the graph (Hamilton et al., 2018). KG's are very useful for explainability and explainable AI methods based on counterfactual queries to the trained GNN models are very promising.

5.4 Human-in-the-loop

Theoretical aspects of CLARUS:

CLARUS is currently not yet federated, but has the potential that every client has this CLARUS platform and this would then fulfil the federated human-in-the-loop.

Human-in-the-Loop (Holzinger, 2016) refers to the process of involving a human expert interactively in the machine learning (ML) process to provide feedback, guidance, or even corrections to the model. The human is an integral part of the ML pipeline, interacting with the model/algorithm to improve its performance and ensuring that it aligns with the desired goals and values. This approach is useful in scenarios where the data is complex, ambiguous, or subject to change, and where the model's performance can benefit from human expertise or even from the experts' subjective judgement. This is because sometimes - of course not always - the human expert has domain knowledge, experience and contextual understanding, in German ”Hausverstand”' - what the best AI algorithms are lacking today. An additional benefit is that the human-in-the-loop approach can also improve the transparency, interpretability, and fairness of machine learning models, as it allows for human oversight and intervention in cases where the model produces biased or undesirable results. However, the human-in-the-loop approach, on the other hand, has drawbacks as it can be time-consuming, expensive, and potentially introduce bias or subjectivity into the modelling process, so it is important to carefully design and evaluate the interaction between the human and the model.

The aforementioned explainable AI methods can facilitate the discovery of disease-causing regions within the networks thereby contributing to uncovering a subset of candidate features organised in disease-relevant network modules. Such methods can be used for validation of their applicability to the biomedical domain, fostering better decision-making through interacting with explanations via the human-in-the-loop approach, and there are numerous successful examples (Teso & Kersting, 2019), (Brueckert et al., 2020), (Schramowski et al., 2020), (Baur et al., 2020), (Bodén et al., 2021).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6 Results

6.1 Knowledge Graph

GNNs provide a crucial benefit of enabling the integration of knowledge graphs (Ji et al., 2021). This implies that both ontologies and PPI networks can be effectively incorporated into the algorithmic pipeline, as highlighted in previous research (Staab & Studer, 2010), (Kulmanov et al., 2020), (Liu et al., 2009), (Jeanquartier et al., 2015).
This also enables the integration of human experience, conceptual knowledge, and contextual understanding into machine learning architectures, which is a notable advantage. This so-called "human-in-the-loop" or "expert-in-the-loop" approach can, in some cases, lead to more robust, reliable, and interpretable results (Holzinger, 2016), (Holzinger et al., 2019), (Hudec et al., 2021). It is worth noting that the inclusion of a domain expert does not guarantee success in every instance. However, the incorporation of such expertise can contribute to the attainment of the most critical goals of the AI community, namely, the development of robust, explainable and trustworthy solutions (Holzinger et al., 2022b). These objectives are essential in ensuring the practical and ethical applications of AI in various fields and are meanwhile mandatory e.g. in the European Union. In WP4 knowledge graphs are a key element to make machine learning more interpretable.

6.2 Disease Subnetwork Detection

In a publication about GNNSubNet (Peifer et al., 2022b), we presented a novel method for disease subnetwork detection using protein-protein interaction (PPI) networks and explainable GNNs. Our method leveraged the PPI knowledge to enable more reliable and biologically meaningful learning trajectories compared to classical deep learning approaches. The nodes of the induced PPI network are enriched by biological features from various modalities, such as gene expression and DNA methylation. We applied our proposed method to patients with kidney cancer and demonstrated its ability to detect disease subnetworks. The developed methodology is implemented within our GNN-SubNet Python package, freely available on GitHub (https://github.com/pievos101/GNN-SubNet). In addition, we enhance ensemble learning based on the detected networks. This makes the classifier more robust, but also more interpretable (Pfeifer et al., 2023). Ensemble-learning with GNNs is implemented within our Ensemble-GNN Python package (https://github.com/pievos101/Ensemble-GNN).

Moreover, as a reliable baseline, in terms of classification performance and overlay interpretability, we have developed the software package DFNET (https://github.com/pievos101/DFNET) (Pfeifer et al., 2022a), which implements a network-guided random forest to derive an ensemble classifier based on any induced knowledge-graph. However, in a federated case, a local random forest would need to share the exact split values of its nodes. This is of much concern and was one of the reasons why we further developed federated solutions based on deep GNNs. The shared parameters among clients in that deep learning setting are more secure in terms of privacy.

6.3 Explainability

The classification of Part 1 has been made explainable, i.e. those subnetworks detected that were relevant for the classification ("disease subnetworks") - subgraphs aka "local spheres". For this purpose, we have developed a modified version of the GNNexplainer (Ying et al., 2019) to compute global explanations. This is realised by sampling patient-specific input graphs while optimising a single-node mask. From these values, edge weights are calculated and assigned to the edges of the PPI knowledge graph. Finally, a weighted community detection algorithm infers the relevant subnetworks.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Furthermore, model-agnostic counterfactual explanations and their associated counterfactual paths can be generated using our cpath software library (https://github.com/pievos101/cpath). The underlying algorithm subsequently samples nodes from the induced knowledge graph and permutes/shuffles the associated features. The procedure is repeated until a certain number of outcome classes swap to the other class. The path to that swap is stored. Multiple such paths build a counterfactual graph, which can be explored to better understand the black-box model in place. A scientific paper is in progress; the corresponding software, however, is already available from GitHub under an MIT licence.

6.4 Explainable Federated Learning with GNNs

In recent work (Pfeifer et al., 2022c), we have enabled federated learning with the methods mentioned above.

During this project, we have learned that biomedical knowledge graphs can still be too large to be explored by a domain expert. To address this shortcoming we have developed a novel approach where the knowledge graph is divided into relevant subnetworks using explainable AI, based on which an ensemble classifier is constructed. This ensemble classifier can be efficiently learned in a federated way, and at the same time is more interpretable.

The main idea of the ensemble federation is depicted in Figure 4 and Figure 5. Each client contains several graphs and each of those graphs represents a patient. The values of the nodes and edges are different in general (as depicted by the different colours of the nodes in the upper part of Figure 4), but the structure of the graphs is the same. Those graphs can be classified by a GNN and the GNN-SubNet method (Pfeifer et al., 2022b) can compute a set of relevant subgraphs for this classification. GNN-SubNet concentrates on providing the relevant structure or topology only; therefore the subgraphs are depicted with white in the middle of Figure 4. The concrete values of the nodes and edges are transferred in a third step though from the original graphs (upper part of Figure 4) to the concrete subgraphs that have the topology of the relevant subgraphs and values overtaken from the original graph (lower part of Figure 4). By creating a new dataset for each discovered relevant subgraph where its structure is repeated and the values are taken from the original graph of all the patients in the client, a separate GNN is trained. The predictions of all those GNNs are input to a majority vote procedure that - in its non-federated version - has an acceptable local performance.

The federation is depicted in Figure 5 and follows a decentralised strategy. The clients use local GNNs of their peers in the inter-client network, that were created with similar logic but were trained with graphs having different topologies - since the relevant subgraphs for each client are expected to vary in general. There is no exchange of the discovered relevant topologies of each client, only the GNN parameters are transferred - which is as far as privacy is concerned less revealing. The majority vote over all those GNNs provided a better performance over each client's test set, but not over a test set that was isolated from all clients, as shown in (Pfeifer et al., 2023). The scenario of non-IDD data has to be simulated in future work, by including imbalanced distribution of data and potentially explicitly defining different feature distributions in the clients (Ortega et al., 2018).

Lastly, the discovered relevant topologies can also be subject to changes driven by human users through the CLARUS UI, changing the ensemble subgraphs of the local GNNs, and by that the whole federation process.

The described method for ensemble-based federated learning on graphs is implemented within our Python package Ensemble-GNN, freely available on GitHub. Moreover, the Ensemble-GNN software is integrated within the FeatureCloud platform (see https://github.com/pievos101/fc-ensemble-gnn).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

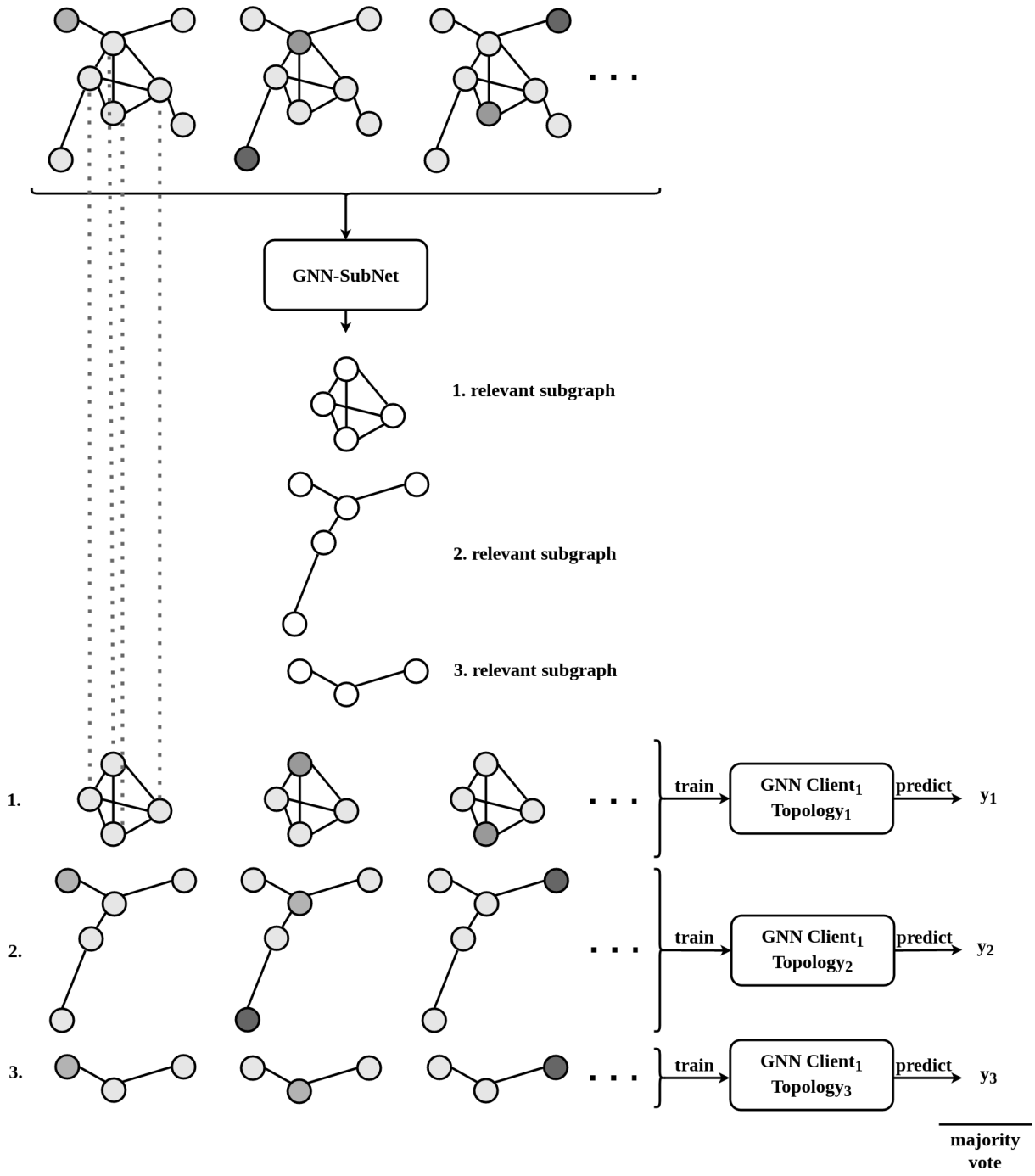

Figure 4. The use of GNN-SubNet in one client, containing a set of graphs for classification. This method extracts a list of relevant subgraph structures (topologies) and uses them by filling the corresponding values of nodes and edges from the original graphs. The newly created datasets are used to train local GNNs and make predictions which are aggregated by majority voting (Holzinger et al., 2023).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

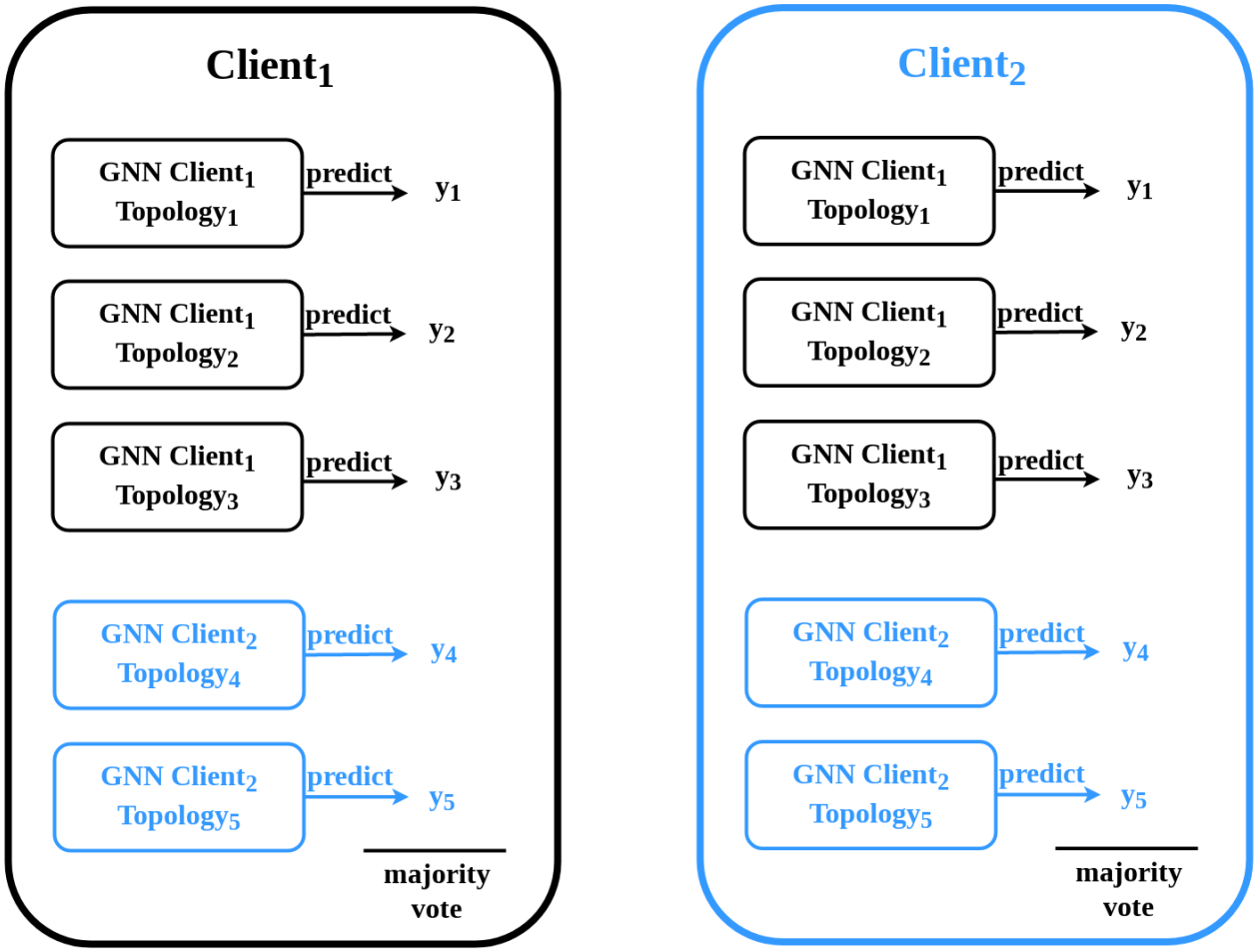

Figure 5. Depiction of the federated learning of Ensemble-GNN. The late fusion of exchanged GNN's predictions through voting is the way the federation is driven by the result of the xAI method (Holzinger et al., 2023).

6.5 Centralised Federated Learning with xAI and Human-in-the-Loop

The main goal of the federation is to create an AI model that distils knowledge from diverse datasets, that share some pre-defined commonalities either in the feature space (Horizontal FL) or in the data space (Vertical FL). One characteristic example in the medical domain is the existence of medical records for different patients in different hospitals. Provided that the data are gathered to diagnose the same disease, one can fairly assume that all hospitals gather similar information. Nevertheless, one cannot completely exclude a situation where a hospital or a doctor decides to gather more or different information than the others. This creates a situation where the number of input features of the AI model is not the same among all hospitals. Furthermore, the number of samples (corresponding to one case or one patient) of each hospital is different; particularly for the cases where a small dataset is gathered one would like to have a more powerful AI model, which is not possible unless one has enough data.

Due to privacy constraints, it is not possible to share information between the hospitals in general. Due to the General Data Protection Regulation (GDPR), personal data cannot be shared. Nonetheless, it is less risky to send information about the AI models, such as weights, gradients, or embeddings although it has been shown that one can extract valuable information even from those, in the case of a successful attack (Geiping et al. 2020), (Hatamizadeh et al., 2023), (Chen et al.,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2022). Since there are ways to counteract those attacks (Huang et al., 2021), (Zhang et al., 2022b), (Eloul et al., 2022), this research work concentrates on the simulation of the information exchange containing GNN weights, gradients and embeddings, and the aspects of privacy are handled by colleagues in the FeatureCloud Project (WP8 “Testing and Evaluation in Clinical Translation”) - as they are taken over for the other AI models in the FeatureCloud App Store as well.

The basic model of centralised federation applied in the Counterfactuals platform is presented in Figure 6. It is composed of several clients, each of them operating independently from each other, and one central server. Each of the clients contains its dataset and its own GNN model, although there must be some pre-agreement between them as far as the similarities and differences of the datasets are concerned. It is expected that the size of the graph and the number of the features of each dataset can also vary; for this task which encompasses graph classification, different graph sizes are not a problem - as long as the type of nodes and edges is the same, meaning that the features have to be equal (up to their values). The platform does not support heterogeneous graphs yet; this is a prerequisite for federating graph datasets of different types of nodes and edges. For the basic federation scheme to work, all clients and the server have to have the same GNN architecture.

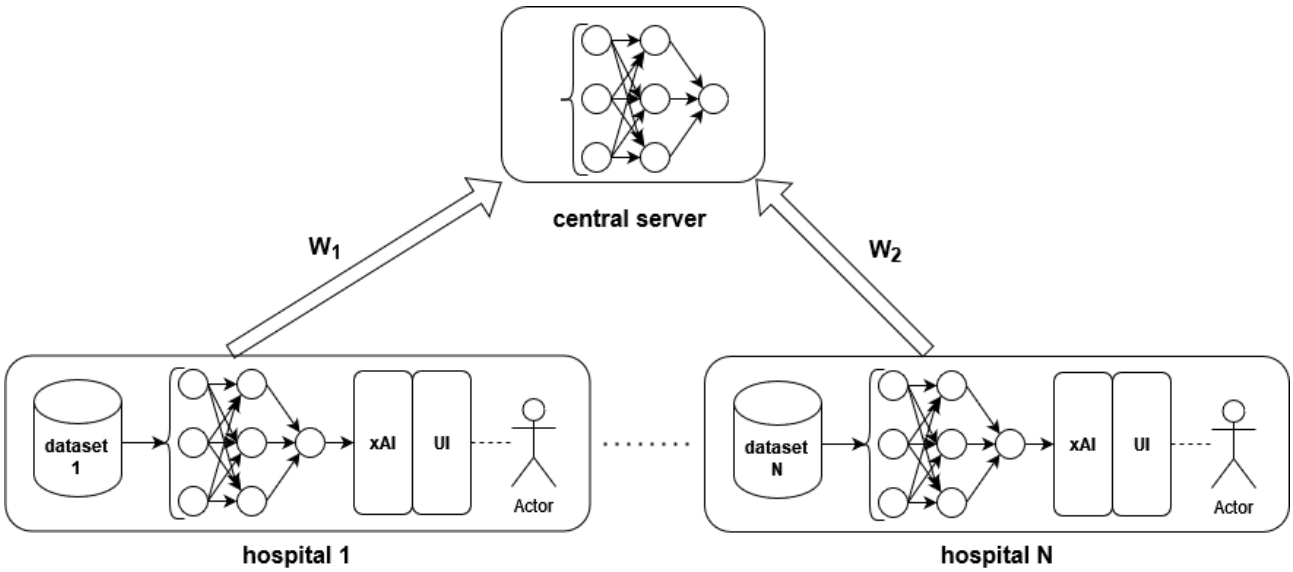

Figure 6. Federated Learning Overview: Each hospital has its own dataset with different characteristics, but also some similarities with the others.

As also seen in Figure 6, it is expected that in the first round, all clients train their own local GNN, each of them with their dataset. The weights of the first GNN are described by **W_1**, the ones of the second **W_2** and so on. At some particular point which can either be a) periodic or b) asynchronous, the weights of one or more of those GNNs are sent to the server. There, they will be averaged, as described in the research work (McMahan et al., 2017), and the resulting weights **W** are going to be sent back to each of the clients. The GNN of each client can use the weights **W** and replace its weights with it, or not.

Typically after the adoption of **W** weights one detects lower performance in each of the individual client GNNs, which is expected since those weights are not tailored to the distribution of the local dataset. Nevertheless, the adoption of those weights prepares the GNN for adequate generalization in cases where similar patients/diseases (whatever the graph represents) as the ones presented to other clients, occur in the future - even if there are not as many as the ones the other models from other clients have been trained to. One example scenario would be as follows: a patient coming to hospital 1 that has similar characteristics with several patients of the other hospitals will highly likely

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

not have a good diagnosis with the first GNN, if he/she is different from the patients already registered in hospital 1. Even re-training with this one new data sample in the dataset will not be enough to influence the weights of the first GNN towards a direction where this (more or less) outlier will need for producing a correct prediction in the output. What is more, there are cases reported where the central GNN model weights **W** were proven to be better - as far as performance goes - than the local GNN model (He et al., 2021). The reasons for that are currently unknown, they could be probably uncovered through xAI methods and they are a very interesting direction for future work.

One fundamental difference of the approach with the use of the xAI Counterfactuals platform is that the model parameters change only after retraining; this is decided by a human user and it does not occur because some new patient or new disease information has entered or was removed from the local dataset. Those two processes both occur asynchronously. Two strategies can be developed: either the weights are gathered periodically (even if for some local clients they haven't changed at all - one can ”catch” it with a request) or each time a client retrains, after the retrain is finished, the weights can be sent to the server and a new average can be computed.

The scenario of different hospitals needs to be simulated; to achieve this, the dataset that we already work with (PPI) needs to be split in a way that simulates non-identically independent (non-IDD) distributed data (Ortega et al., 2018). That means that the balance of each of the client datasets needs to differ. Since the task is a binary classification, it must be ensured that there are local datasets containing f.e. 70% of their data belonging to class 0 and 30% of their data belonging to class 1, but also other ones having f.e. 60% of their data belonging to class 1 and 40% of their data belonging to class 0. The logic of creating this imbalance and the number of clients should be configurable. This is the most fundamental type of non-IDD simulation for the data distribution; in a future step, there is the need to synthetically generate non-IDD topology data (Liu et al., 2022).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

7 Conclusion

MUG has successfully accomplished the objectives set forth, meeting expectations for Objective 5, Task 6, and delivering on MS 29. All developed algorithms have been shared openly on GitHub for the benefit of the international research community.

Critical contributions have included the description and development of explainable graph-neural networks and the human-in-the-loop concept involving counterfactual explanations. These developments have been recognized and frequently cited by the international research community, leading to widespread implementation of these visions and ideas within related work.

Innovative work has also been conducted in the field of deep neural network learning, specifically in relation to protein-protein interaction (PPI) networks. By masking the learning process with a PPI network, patient-specific multi-modal genomic features could be included, thereby improving the classification process. This process was rendered explainable by detecting relevant subnetworks for the classification, also known as "disease subnetworks".

For a more robust comparison, subnetwork detection was also performed using a random forest methodology. This approach was chosen for its importance in the medical field due to its interpretability and explainability, and the learning process was masked by a knowledge graph.

Building upon this work, federated deep learning on graphs was enabled by decomposing the knowledge graph into relevant subnetworks. This approach allowed the construction of an ensemble classifier that could be learned in a federated manner. The resulting ensemble classifier offers greater interpretability by allowing for more efficient inspection of the classification performance of specific parts of the graph.

Finally, an end-user-focused interface, known as "CLARUS," was designed and developed. This tool allows expert users to re-enact and re-trace results, analyse detected subnets, and manipulate them to gain insights into network behaviour. The usability and evaluation study of this tool has provided further valuable insights, and it has been integrated back into the ensemble classifier. All major outcomes and findings have been summarised and shared openly with the international research community.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

8 References

- Acosta, J.N., Falcone, G.J., Rajpurkar, P. & Topol, E.J. (2022). Multimodal biomedical AI. Nature Medicine, 28, (9), 1773-1784, doi: s41591-022-01981-2.
- Asad, M., Moustafa, A., & Ito, T. (2020). FedOpt: Towards communication efficiency and privacy preservation in federated learning. Applied Sciences, 10(8), 2864.
- Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). eXplainable cooperative machine learning with NOVA. KI-Künstliche Intelligenz, 34, 143-164.
- Beinecke, J., Saranti, A., Angerschmid, A., Pfeifer, B., Klemt, V., Holzinger, A. & Hauschild, A.-C. (2022). CLARUS: An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks. bioRxiv, 2022.11. 21.517358, doi: 10.1101/2022.11.21.517358.
- Bellavista, P., Foschini, L. & Mora, A. (2021). Decentralised learning in federated deployment environments: A system-level survey. ACM Computing Surveys (CSUR), 54 (1), 1--38, doi: 10.1145/3429252.
- Bodén, A.C.S., Molin, J., Garvin, S., West, R.A., Lundström, C. & Treanor, D. (2021). The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice. Histopathology, 79 (2), 210--218, doi: 10.1111/his.14356.
- Bruckert, S., Finzel, B., & Schmid, U. (2020). The next generation of medical decision support: a roadmap toward transparent expert companions. Frontiers in artificial intelligence, 3, 507973.
- Chen, J., Huang, G., Zheng, H., Yu, S., Jiang, W., & Cui, C. (2022). Graph-fraudster: Adversarial attacks on graph neural network-based vertical federated learning. IEEE Transactions on Computational Social Systems.
- Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., Kramer, F., Leha, A. & Beißbarth, T. (2021). Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. Genome medicine, 13, 1-16, doi: 10.1186/s13073-021-00845-7.
- Dehmer, M., Emmert-Streib, F., & Shi, Y. (2017). Quantitative graph theory: a new branch of graph theory and network science. Information Sciences, 418, 575-580.
- Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M. & Zitnik, M. (2023). Multimodal learning with graphs. Nature Machine Intelligence, 1-11, doi: 10.1038/s42256-023-00624-6.
- Eloul, S., Silavong, F., Kamthe, S., Georgiadis, A., & Moran, S. J. (2022). Enhancing Privacy against Inversion Attacks in Federated Learning by using Mixing Gradients Strategies. preprint arXiv: 2204.12495.
- Fu, X., Zhang, B., Dong, Y., Chen, C. & Li, J. (2022). Federated graph machine learning: A survey of concepts, techniques, and applications. ACM SIGKDD Explorations Newsletter, 24, (2), 32--47, doi: 10.1145/3575637.3575644.
- Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems, 33, 16937-16947.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Hamilton, W., Bajaj, P., Zitnik, M., Jurafsky, D., & Leskovec, J. (2018). Embedding logical queries on knowledge graphs. Advances in neural information processing systems, 31.
- Hamon, R., Junklewitz, H. & Sanchez, I. (2020). Robustness and Explainability of Artificial Intelligence - From technical to policy solutions, Luxembourg, Publications Office of the European Union, doi: 10.2760/57493.
- Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., ... & Roth, H. R. (2023). Do gradient inversion attacks make federated learning unsafe? IEEE Transactions on Medical Imaging.
- He, C., Balasubramanian, K., Ceyani, E., Yang, C., Xie, H., Sun, L., ... & Avestimehr, S. (2021). Fedgraphnn: A federated learning benchmark system for graph neural networks. In ICLR 2021 Workshop on Distributed and Private Machine Learning (DPML).
- Holzinger, A. (2014). On topological data mining. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges, BMC Bioinformatics, 15, 331-356.
- Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3, (2), 119-131, doi: 10.1007/s40708-016-0042-6.
- Holzinger, A. (2021). The Next Frontier: AI We Can Really Trust. In: Kamp, Michael (ed.) Proceedings of the ECML PKDD 2021, CCIS 1524. Cham: Springer Nature, pp. 427--440, doi: 10.1007/978-3-030-93736-2_33.
- Holzinger, A., Dehmer, M. & Jurisica, I. (2014). Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. Springer/Nature BMC Bioinformatics, 15, (S6), I1, doi: 10.1186/1471-2105-15-S6-I1.
- Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I. & Díaz-Rodríguez, N. (2022b). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, (3), 263--278, doi: 10.1016/j.inffus.2021.10.007.
- Holzinger, A., Haibe-Kains, B. & Jurisica, I. (2019c). 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, (13), 2722-2730, doi: 10.1007/s00259-019-04382-9.
- Holzinger, A., Malle, B., Saranti, A. & Pfeifer, B. (2021). Towards Multi-Modal Causability with Graph Neural Networks enabling Information Fusion for explainable AI. Information Fusion, 71, (7), 28-37, doi: 10.1016/j.inffus.2021.01.008.
- Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C.-M. & Palade, V. (2019b). A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. Creative Mathematics and Informatics, 28, (2), 121--134, doi: 10.37193/CMI.2019.02.04.1708.01104.
- Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G.C., Pintea, C.-M. & Palade, V. (2019a). Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence, 49, (7), 2401-2414, doi: 10.1007/s10489-018-1361-5.
- Holzinger, A., Saranti, A., Hauschild, A.-C., Beinecke, J., Heider, D., Roettger, R., Mueller, H. & Baumbach, J., Pfeifer, B. (2023). Human-in-the-Loop Integration with Domain-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Knowledge Graphs for Explainable Federated Deep Learning. Springer Lecture Notes in Computer Science (LNCS) Volume 14065. 1--26.

- Holzinger, A., Saranti, A., Molnar, C., Biecek, P. & Samek, W. (2022a). Explainable AI Methods - A Brief Overview. XXAI - Lecture Notes in Artificial Intelligence LNAI 13200. Cham: Springer, pp. 13--38, doi: 10.1007/978-3-031-04083-2_2.
- Hu, Y., Niu, D., Yang, J. & Zhou, S. (2018). Stochastic Distributed Optimization for Machine Learning from Decentralised Features. arXiv: 1812.06415, 1-10.
- Huang, Y., Gupta, S., Song, Z., Li, K., & Arora, S. (2021). Evaluating gradient inversion attacks and defenses in federated learning. Advances in Neural Information Processing Systems, 34, 7232-7241.
- Hudec, M., Mináriková, E., Mesiar, R., Saranti, A., & Holzinger, A. (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. Knowledge-Based Systems, 220, 106916.
- Jeanquartier, F., Jean-Quartier, C., & Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. BMC bioinformatics, 16(1), 1-16.
- Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), 494-514.
- Johnson, K.B., Wei, W.Q., Weeraratne, D., Frisse, M.E., Misulis, K., Rhee, K., Zhao, J. & Snowdon, J.L. (2021). Precision medicine, AI, and the future of personalized health care. Clinical and Translational Science, 14, (1), 86--93, doi: 10.1111/cts.12884.
- Kieseberg, P., Frühwirt, P., Weippl, E. & Holzinger, A. (2015). Witnesses for the Doctor in the Loop. In: Guo, Yike, Friston, Karl, Aldo, Faisal, Hill, Sean & Peng, Hanchuan (eds.) Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Cham, Heidelberg, Berlin: Springer, pp. 369-378, doi: 10.1007/978-3-319-23344-4_36.
- Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2020). Machine learning with biomedical ontologies. biorxiv, 2020-05.
- Lage, I., Ross, A., Gershman, S.J., Kim, B. & Doshi-Velez, F. (2018). Human-in-the-loop interpretability prior. Advances in Neural Information Processing Systems NeurIPS 2018. Montreal. 10159--10168.
- Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R. & Samek, W. (2016). The LRP toolbox for artificial neural networks. The Journal of Machine Learning Research (JMLR), 17, (1), 3938--3942.
- Liu, G., Wong, L., & Chua, H. N. (2009). Complex discovery from weighted PPI networks. Bioinformatics, 25(15), 1891-1897.
- Liu, R., Xing, P., Deng, Z., Li, A., Guan, C. & Yu, H. (2022). Federated graph neural networks: Overview, techniques and challenges. arXiv:2202.07256, 1--16, doi: 10.48550/arXiv.2202.07256.
- Lucic, A., Ter Hoeve, M.A., Tolomei, G., De Rijke, M. & Silvestri, F. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. International Conference on Artificial Intelligence and Statistics, (2022). PMLR.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H. & Zhang, X. (2020). Parameterized explainer for graph neural network. Advances in neural information processing systems, 33, 19620-19631.
- Magister, L. C., Kazhdan, D., Singh, V., & Liò, P. (2021). Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. arXiv preprint arXiv: 2107.11889.
- Magister, L.C., Barbiero, P., Kazhdan, D., Siciliano, F., Ciravegna, G., Silvestri, F., Jamnik, M. & Lio, P. (2022). Encoding concepts in graph neural networks. arXiv: 2207.13586.
- Malle, B., Giuliani, N., Kieseberg, P. & Holzinger, A. (2017). The More the Merrier - Federated Learning from Local Sphere Recommendations. Machine Learning and Knowledge Extraction, Lecture Notes in Computer Science LNCS 10410 Cham: Springer, pp. 367--374, doi: 10.1007/978-3-319-66808-6_24.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273-1282).
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56, (4), 3005-3054, doi: 10.1007/s10462-022-10246-w.
- Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., & Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. Proceedings of the IEEE, 106(5), 808-828.
- Pfeifer, B., Baniecki, H., Saranti, A., Biecek, P. & Holzinger, A. (2022a). Multi-omics disease module detection with an explainable Greedy Decision Forest. Scientific reports, 12, (1), 1--15, doi: 10.1038/s41598-022-21417-8.
- Pfeifer, B., Chereda, H., Martin, R., Saranti, A., Angerschmid, A., Clemens, S., Hauschild, A.-C., Beissbarth, T., Holzinger, A. & Heider, D. (2023). Ensemble-GNN: federated ensemble learning with graph neural networks for disease module discovery and classification. bioRxiv, 2023.03.22.533772, doi: 10.1101/2023.03.22.533772.
- Pfeifer, B., Holzinger, A., & Schimek, M. G. (2022c). Robust Random Forest-Based All-Relevant Feature Ranks for Trustworthy AI. Studies in Health Technology and Informatics, 294, 137-138.
- Pfeifer, B., Saranti, A. & Holzinger, A. (2022b). GNN-SubNet: disease subnetwork detection with explainable Graph Neural Networks. Bioinformatics, 38, (S-2), ii120-ii126, doi: 10.1093/bioinformatics/btac478.
- Rost, B., Radivojac, P. & Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. FEBS letters, 590, (15), 2327-2341.
- Saranti, A., Hudec, M., Minarikova, E., Takac, Z., Großschedl, U., Koch, C., Pfeifer, B., Angerschmid, A. & Holzinger, A. (2022). Actionable explainable AI (AxAI): a practical example with aggregation functions for adaptive classification and textual explanations for interpretable Machine Learning. Machine Learning and Knowledge Extraction, 4, (4), 924--953, doi: 10.3390/make4040047.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K.T., Müller, K.-R. & Montavon, G. (2020). XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. arXiv: 2006.03589.
- Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., ... & Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), 476-486.
- Staab, S., & Studer, R. (Eds.). (2010). Handbook on ontologies. Springer Science & Business Media.
- Sverrisson, F., Feydy, J., Correia, B.E. & Bronstein, M.M. Fast end-to-end learning on protein surfaces. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2021).
- Teso, S., & Kersting, K. (2019). Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 239-245).
- Veličković, P. (2023). Everything is connected: Graph neural networks. Current Opinion in Structural Biology, 79, 102538.
- Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. & He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 135, (10), 364--381, doi: 10.1016/j.future.2022.05.014.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. & Philip, S.Y. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 1-21, doi: 10.1109/TNNLS.2020.2978386.
- Ying, Z., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In: Wallach, Hanna, Larochelle, Hugo, Beygelzimer, Alina, D'alche-Buc, Florence, Fox, Emily & Garnett, Roman (eds.) Advances in neural information processing systems. Vancouver. 9244--9255.
- Zhang, H., Wu, B., Yuan, X., Pan, S., Tong, H. & Pei, J. (2022a). Trustworthy graph neural networks: Aspects, methods and trends. arXiv:2205.07424, 1--36, doi: 10.48550/arXiv.2205.07424.
- Zhang, R., Guo, S., Wang, J., Xie, X., & Tao, D. (2022b). A Survey on Gradient Inversion: Attacks, Defenses and Future Directions. Preprint arXiv: 2206.07284.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.