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 D4.6
"End-user centred explanatory interfaces for the human-in-the-loop"

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) - Established by the European Commission, Unit HaDEA.A.3 – Health Research				
Project Coordinator	Jan BAUMBACH, UNIVERSITY OF HAMBURG (UHAM)				
Deliverable	D4.6 “End-user centred explanatory interfaces for the human-in-the-loop”				
Work Package	WP4 “Supervised Federated Machine Learning”				
Date of Delivery	**Contractual**	31/12/2022 (M48)		**Actual**	16/12/2022
Nature	Report		**Dissemination Level**	Public	
Lead Beneficiary	03 MUG				
Responsible Author(s)	Prof. Dr. Andreas Holzinger, MUG				
Keywords	Explainable AI, Graph Neural Networks, Counterfactuals, Human-in-the-loop, Causability				

Table of Content

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

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

WP 4 will contribute to theoretical and experimental research, design and development of federated interactive learning approaches following a “privacy by design and architecture”. Additionally, WP 4 will experiment and evaluate Explainable AI (xAI) approaches in order to make machine learning results transparent, re-traceable, reenactable and eventually understandable. xAI is a term that encompasses several methods like saliency/heatmap, rule-based, textual output, and causal reasoning, that provide information about the internal decision-making process of an AI algorithm to a human or AI agent. Different xAI methods exist in the literature, mostly for different Neural Network architectures, where the goal is either before, after or during the training of the AI algorithm with the available data to have some overview, control and actionable insight about it. Objective 5 is to design, develop and evaluate end-user-centered interfaces to enable a) the interaction of humans with the algorithms developed; and b) to enable to re-enact and to re-trace in order to explain and understand the results in the context of the medical problem (Task 5).
The milestone MS28 is achieved. The Prototype of the interface was evaluated and is currently publicly accessible on a server (http://rshiny.gwdg.de/apps/interactive_xai_for_gnn/).

2 Executive Summary

- Methodology (if applicable): An interactive user interface was developed to allow the user to manipulate graph data. The used data represents a (protein-protein-interaction) PPI network (see Pfeifer, B. et al. (2022)). Each of the used datasets consists of multiple patients, whereby each patient is represented by one graph. Before the initial training of the Graph Neural Network, the dataset is split into training and test sets after a stratified sampling procedure. Once a graph representing a particular patient is selected to be part of the training or test set, this fact remains unchanged for the rest of the experiment and is supported by the unique identifier that each graph has. Users of the interface can choose between various explainable AI methods for the node, edge (and features thereof) relevances respectively.
- Results (if applicable): The result is a user-centered interface prototype (Beinecke, J. et al. (2022)), which is currently hosted at http://rshiny.gwdg.de/apps/interactive_xai_for_gnn/. The interface allows user manipulation of any type of graph data that represents patient data; PPI networks are one of the possible types. At the same time, transparency through explainable AI methods is provided. For each of the graphs, the true label, predicted label, and confidence of the performance (also called ”assurance”) is presented, and at the same time the explanation for this prediction - for both correct and misclassifications. Different explainable AI methods can be selected to get an explanation of the systems’ classification decision for each of the patients.
- Distinctive features (if applicable): Human-in-the-loop approach through user manipulation of Graph Neural Networks (GNN). Users can find the most relevant proteins in a network while analysing how their changes affected the decision of the GNN. In the case the user uses the changes in the graphs to retrain the GNN, the prediction confidence (and thereby the decision of the GNN) will in general change for all patients. The same is expected for the results of the used xAI methods. In the case where the user wants to just check and reason how his/her actions affect the prediction confidence of the GNN of the previous iteration (trained before the changes), then just using the “predict” functionality is sufficient. The xAI methods results are expected to be also affected, but only for the changed graphs. Lastly, it is important to note that if the user changes only graphs that happen to be in the test set, then retraining will not affect the GNN and thereby the prediction confidence and xAI of the training set will remain intact. Please refer to the figures on page 11.
- Progress beyond the state-of-the-art (if applicable): The interface combines multiple explainable AI methods, thus the increased transparency and interpretability (to non-

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

technophilic domain experts) of the graph neural network and its predictions are beyond state-of-the-art.

3 Introduction (Challenge)

The use of machine learning could potentially help to identify patterns or structures within a graph structure that a human would not be aware of. Thus, the challenge of this deliverable was to enable a human to explore and observe changes in the Graph Neural Network (GNN) model and/or its predictions through a guided manipulation of the dataset with the help of xAI methods. This is done by presenting different explainability methods and metrics to the user for the prediction of a particular graph and the overall changes after an initiated retraining from scratch. We expect users not to be necessarily computer or data scientists but domain experts in the medical or biochemical domain. Hence, domain experts are able to ask counterfactual (“what-if”) questions and observe the changes, based on their questions, in the AI decision and explanation. That means that the users combine their own previously acquired domain knowledge with the results provided by the GNN’s performance and the xAI methods to understand what the juxtaposition of the current dataset, GNN architecture and xAI method reveals about the decision of the GNN’s decision-making process. The main focus thereby is to present the network in such a form that the user (practitioner, biologist, doctor) is able to interact with it and understands the GNN’s internal way of patient data classification. Since these graphs can be extremely complex, the user can choose how many nodes are shown at once. Further, different people prefer different metrics or explanations, hence multiple Explainable AI methods for the network are included within the application. The user is able to change the explainer and the preferred method constantly.

4 Methodology

Three pre-selected datasets are provided to the user, those are the Kidney Renal Clear Cell Carcinoma (KIRC) dataset (a real-world dataset), a smaller subset of the KIRC dataset, as well as, a synthetic dataset. The user is able to choose the preferred dataset to investigate further and select the graph of an individual patient. The interface then allows the user to manipulate and explore the dataset, as well as retrain the GNN or receive predictions of the current graph that corresponds to the data of one particular patient.

5 Results

An interactive explainable Artificial Intelligence platform for Graph Neural Networks (GNNs) was developed. As already stated, pre-selected datasets are provided to the user. Moreover, the user is able to choose the preferred explainer of the GNN. The user is able to specify the graph visualisations by selecting how to sort and colour the nodes, by degree, by relevances from XAI models or alphabetically according to their labels. The results of this deliverable might enable an experienced domain expert to draw conclusions on which genes affect the decision-making process.

6 Open issues

The current state of the interface needs to be evaluated with a user study in order to verify its usability and causability. Moreover, the underlying GNNs of the network are trained locally (on the server) to ensure the functionality of the application, thus the training itself is not federated yet. However, in cooperation with the partners UHAM and GND, MUG works on a federated solution.

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

7 Deviations (if applicable)

No deviations.

8 Conclusion

The interface combines an explorative and interactive approach to allow domain experts e.g., doctors, biologists, practitioners and scientists to investigate networks and their parameters. Further, those experts are able to choose between different xAI methods and graph visualisations, while manipulating the network based on counterfactual (“what-if”) questions. The interface supports transparency and should allow the domain experts to find underlying patterns within the data. Thus, impact and relevance for the project are an experimental platform, which allows a user to view and manipulate graph data in a concise manner. The explanations help to foster the human understanding of the networks, as well as the GNNs. By asking counterfactual questions, the experts can identify important parameters and their influence/relevance on the whole network. For example, this could mean that domain experts are able to identify certain proteins or structures that cause patients to develop cancer.

9 References

Pfeifer, B. et al. (2022) ‘GNN-SubNet: Disease subnetwork detection with explainable graph neural networks.’, *Bioinformatics* 38. Supplement_2 (2022): ii120-ii126. Available at: https://doi.org/10.1093/bioinformatics/btac478.

Beinecke, J. et al. (2022) ‘Interactive explainable AI platform for graph neural networks’, *bioRxiv*. Available at: https://doi.org/10.1101/2022.11.21.517358.

Holzinger, Andreas et al. (2022) ‘Explainable AI methods - A brief overview.’, *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-031-04083-2_2.

Saporta, Adriel et al. (2022) ‘Benchmarking saliency methods for chest X-ray interpretation.’, *Nature Machine Intelligence* 4.10: 867-878. Available at: https://doi.org/10.1038/s42256-022-00536-x.

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

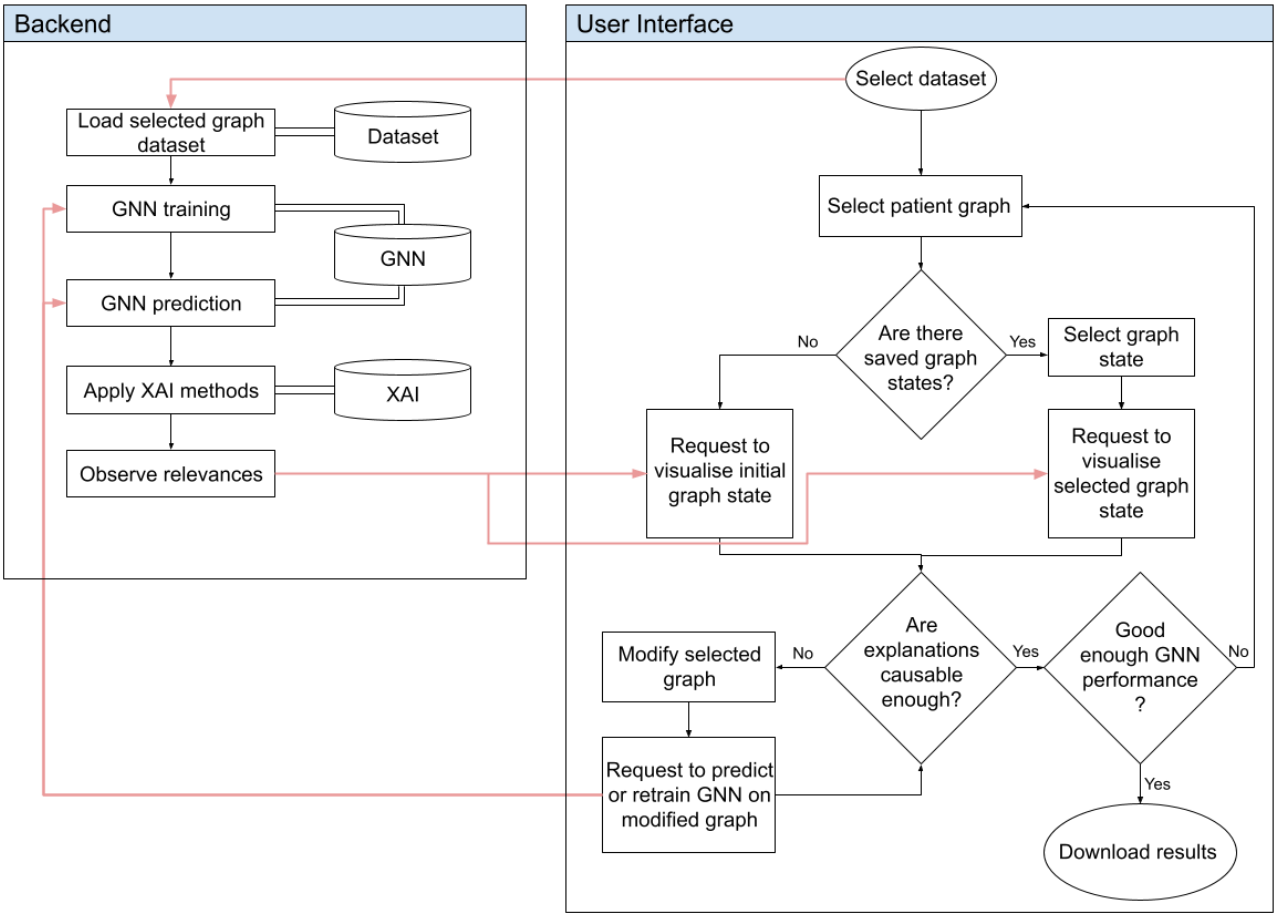
10 Table of acronyms and definitions

AI	Artificial Intelligence
concentris	concentris research management GmbH
GND	Gnome Design SRL
GNN	Graph Neural Network
KIRC	Kidney Renal Clear Cell Carcinoma
MS	Milestone
MUG	Medizinische Universitaet 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
SDU	Syddansk Universitet
UHAM	University of Hamburg
UMG	University Medical Center Göttingen
UMR	Philipps Universitaet Marburg
WP	Work package
xAI	Explainable Artificial Intelligence

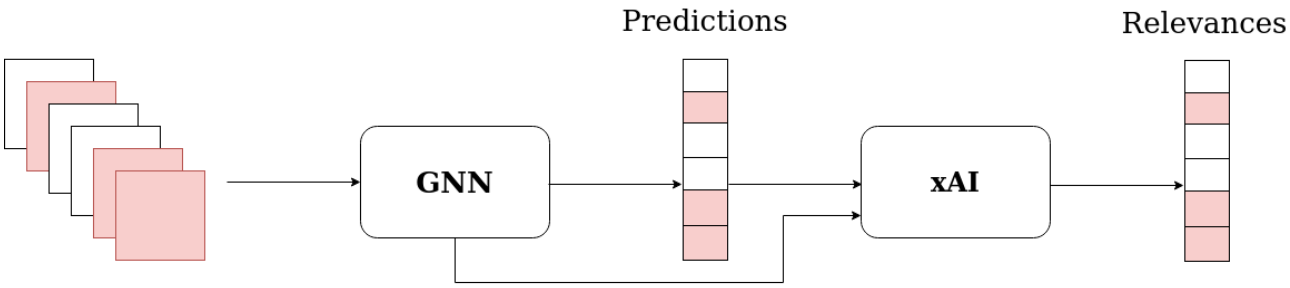
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

11 Other supporting documents / figures / tables (if applicable)

The following image shows the implementation of the xAI interface and a typical user workflow.

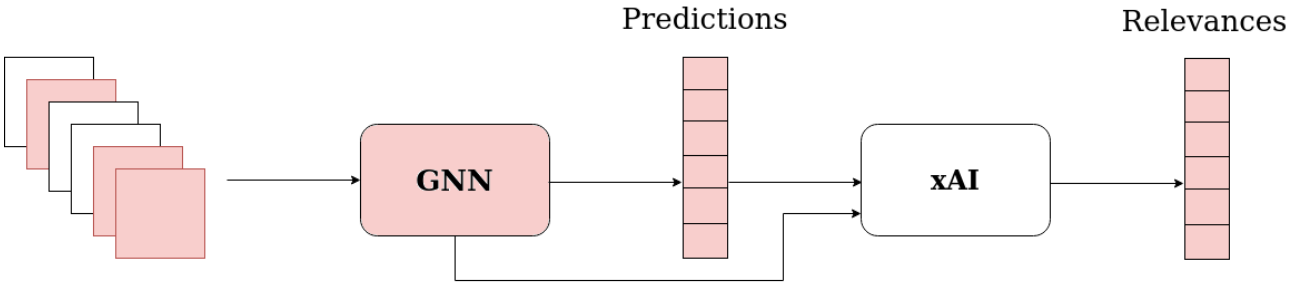


Prediction procedure. The graphs that were changed by the actions of the user are marked with pink colour. After the “predict” procedure is activated (through the pressing of the “predict” button) the GNN remains unchanged but the predictions of the changed graphs (in general) change - and therefore their relevances also. The predictions of the unchanged graphs remain intact. This is shown in the following figure.



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

Retrain procedure. The graphs that were changed by the actions of the user are marked with pink colour. After the “retrain” procedure is activated (through the pressing of the “retrain” button) the GNN changes internally, and the predictions of all graphs change - and therefore the relevances also. This scenario presupposes that at least one of the changed graphs belongs to the training set.



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