Explaining decisions of graph convolutional neural networks for analyses of molecular subnetworks in cancer

Dissertation
for the award of the degree
"Doctor of Philosophy" Ph.D.
Division of Mathematics and Natural Sciences
at the Georg-August-Universität Göttingen

within the doctoral program
International Max Planck Research School for Genome Science
of the Georg-August University School of Science (GAUSS)

submitted by
Hryhorii Chereda
from Kyiv, Ukraine

Göttingen, 2021

Thesis Committee:

Prof. Dr. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Center Göttingen

Dr. Johannes Söding
Research Group Quantitative and Computational Biology
Max Planck Institute for Biophysical Chemistry

Prof. Dr. Frank Kramer
IT Infrastructure for Translational Medical Research
University of Augsburg

Members of the Examination Board:

1st Referee: Prof. Dr. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Center Göttingen

2nd Referee: Dr. Johannes Söding
Research Group Quantitative and Computational Biology
Max Planck Institute for Biophysical Chemistry

Further members of the Examination Board:

Prof. Dr. Frank Kramer
IT Infrastructure for Translational Medical Research
University of Augsburg

Prof. Dr. Burkhard Morgenstern
Department of Bioinformatics
Institute for Microbiology and Genetics
Georg August University Göttingen

Dr. Nico Posnien
Department of Developmental Biology
Johann-Friedrich-Blumenbach-Institute of Zoology and Anthropology
Georg August University Göttingen

Prof. Dr. Fabian Sinz
Research Group Machine Learning
Institute of Computer Science
Georg August University Göttingen

Date of oral examination: 24th of February 2022

Abstract

Contemporary deep learning approaches exhibit state-of-the-art performance in various areas. In healthcare, the application of deep learning remains limited since deep learning methods are often considered as non-interpretable black-box models. However, the *Machine Learning* (ML) community made recent elaborations to develop the methods of *eXplainable Artificial Intelligence* (XAI). The explanation methods explain single decisions of an ML model when a single data point is fed into the model's input. In a clinical setup, a data point can represent a single patient. Data point-specific explanations could possibly assist the need in personalized precision medicine decisions via explaining patient-specific predictions.

*Convolutional Neural Network*s (CNNs) as deep learning methods have been already applied to classify transformed into images gene expression profiles of patients. Gene expression data can be structured by a prior knowledge molecular network (encoded as a graph) representing connections between genes. Each vertex of a molecular network is assigned a gene expression value as an attribute. The set of the attributes creates a graph signal representing a patient. Emerging field of *geometric deep learning* deals with methods applicable to graph structured data and extends CNNs as *Graph Convolutional Neural Network*s (GCNNs) classifying graph signals.

Layer-wise Relevance Propagation (LRP) is a method to explain decisions of CNNs classifying image data. I extended the LRP method to make it available for GCNNs. *Graph Layer-wise Relevance Propagation* (GLRP) is presented as a new method to explain single decisions made by a GCNN model.

In this thesis, I present a novel methodology generating patient-specific molecular subnetworks as explanations for classification decisions of an ML approach utilizing prior knowledge of molecular networks. GCNN serves as a ML approach, and its decisions are explained by developed GLRP. A sanity check of the developed GLRP method was demonstrated on a hand-written digits dataset. The biological validation was performed by applying the developed methodology to gene expression data from *Human Umbilical Vein Endothelial Cells* (HUVEC) treated or not treated with tumor necrosis factor alpha. To show the utility of introduced methodology in the scopes of precision medicine, it was applied to a large breast cancer dataset. The generated patient-specific subnetworks largely agree with clinical knowledge and could assist precision medicine approaches by identifying common as well as novel, and potentially druggable, drivers of tumor progression.

Apart from generating patient-specific subnetworks, the developed methodology can be used as a general feature selection approach. The outcome of a feature selection approach is a subset of important for classification genes corresponding to a whole dataset. It is essential to sustain stability of selected feature subsets across different datasets with the same clinical endpoint since the selected genes are possible candidates for prognostic biomarkers.

I analysed the stability of feature selection performed by GCNN+LRP. I have implemented a graph convolutional layer of GCNN as a Keras layer so that the *SHapley Additive exPlanations* (SHAP) method could be also applied to a Keras version of a GCNN model. The stability of feature selection performed by GCNN+LRP was compared to the stability of GCNN+SHAP and other ML-based feature selection approaches. The GCNN+LRP approach shows the highest stability. GCNN+LRP subnetworks were compared to GCNN+SHAP subnetworks in terms of connectivity and permutation feature importance. While GCNN+SHAP subnetworks demonstrate higher permutation importance than GCNN+LRP subnetworks, a GCNN+LRP subnetwork of an individual patient is on average substantially more connected and, therefore more interpretable in the context of prior knowledge than a GCNN+SHAP subnetwork which consists mainly of single vertices.

Acknowledgements

First and foremost I want to express my gratitude to my supervisor Prof. Dr. Tim Beißbarth for his invaluable guidance, for the door always open for spontaneous meetings, for encouragement, and for sharing his knowledge. Tim's ideas and enthusiasm have been an additional source of motivation to me and helped to overcome some of the overwhelming periods on the course of my Ph.D. student life.

My gratitude also goes to the thesis committee members, Dr. Johannes Söding and Prof. Dr. Frank Kramer. I would like to thank Frank, without whom this thesis never would have happened, for guiding me into the field of deep learning on graphs at the beginning of my doctoral studies. Also, I want to express my deep thanks to Johannes for always being eager to meet and to give insightful advice, for constructive and inspiring discussions, and for being fast in replying to my emails.

Also, I am very grateful to current and former colleagues. I would like to thank Dorit Meyer, Doris Waldmann, and Yvonne Lamprecht for taking care of my administrative inquires and for always being helpful. I would also like to thank Dr. Daniela Großman for always being helpful. Especially I would like to acknowledge Dr. Martin Haubrock, Darius Wlochowitz, Dr. Jürgen Dönitz, Dr. Manuel Nietert, Liza Vinhoven, Dr. Maren Sitte, Halima Alachram, Ph.D., Karly Conrads, Niels Paul, Kevin Kornrumpf, Malte Sahrhage, Hazal Timucin for creating a great working environment.

Many many thanks go to Florian Auer and Júlia Perera Bel, Ph.D., who are my former office mates, closest collaborators, and friends, for always having time for questions, discussions, and for providing an exceptional scientific and interpersonal environment. I would also like to thank Dr. Andreas Leha, whose professionalism in resolving research problems I took as an example, for mentoring, for being genuinely interested in my research questions, and for motivating feedback. Special thanks go to Darius Wlochowitz for promoting my German language skills, constructive advice, and his friendship. Also I would like to thank Philip Stegmaier for providing scientific insights and fruitful discussions.

I would like to thank Dr. Kerstin Menck and Prof. Dr. med. Annalen Bleckmann. Their contributions and expert insights are crucial to the work presented in this thesis.

I would also like to thank Torsten Schöps for greatly taking care of the software and hardware infrastructure at our department and for being always responsive on my inquires.

Furthermore, I would like to thank Dr. Henriette Irmer — coordinator of the IMPRS-GS program — for maintaining an atmosphere of excellence and being extremely helpful.

Sincere thanks also goes to Nicole Seifert, Niels Paul, and Niklas Lück for proofreading this thesis.

I owe thanks to Michaela Bayerlová, Ph.D. for providing scientific advice and bioinformatics insights.

Finally, I would like to express my gratitude to my family and friends for their support.

Contents

List of Figures

Acronyms

CNN *Convolutional Neural Network*

DL *Deep Learning*

EU *European Union*

GCNN *Graph Convolutional Neural Network*

GNN *Graph Neural Network*

GSP *Graph Signal Processing*

HPRD *Human Protein Reference Database*

HUVEC *Human Umbilical Vein Endothelial Cells*

LIME *Local Interpretable Model-agnostic Explanations*

LRP *Layer-wise Relevance Propagation*

GLRP *Graph Layer-wise Relevance Propagation*

ML *Machine Learning*

MLP *Multi-layer Perceptron*

MTB *Molecular Tumor Board*

NGS *Next Generation Sequencing*

NN *Neural Network*

PPI *Protein-Protein Interaction*

PPM *Precision and Personalized medicine*

RGNN *Recurrent Graph Neural Network*

SHAP *SHapley Additive exPlanations*

TCGA *The Cancer Genome Atlas*

XAI *eXplainable Artificial Intelligence*

1 Introduction

After cardiovascular diseases, the second leading cause of death in *European Union* (EU) countries is cancer. Cancer caused 2.7 million cases and 1.3 million deaths in the EU in 2020 [59]. In 31 countries in Europe (EU-27 plus Iceland, Norway, Switzerland, and the United Kingdom) cancer incidence gradually increased by around 50 % from 2.1 to 3.1 million cases from 1995 to 2018 [28]. Cancer mortality increased by around 20 % from 1.2 million to 1.4 million cases [28] during the same time frame, but deaths in people younger than 65 years decreased. The increase in cancer deaths has been happening slower than increase in cancer incidence between 1995 and 2018. Hence, it reflects improvements in patient outcomes [28].

Besides, cancer is an extremely costly disease. The total economic impact for Europe in 2018 was estimated to be 199 billion euro [28]. This amount includes direct costs within the health-care system, the costs of informal care provided by family and friends, and productivity losses caused by premature mortality and morbidity.

The economic burden of cancer is crucial for policy decisions made not only on the national, but also on European level. In February 2021, the European Commission presented "Europe's Beating Cancer Plan" — a political commitment to take action against cancer [19]. The commitment is focused on four key aspects where the EU can add the most value: (1) prevention; (2) early detection; (3) diagnosis and treatment; and (4) quality of life of cancer patients and survivors. These key aspects are heavily dependent on the fields of personalized medicine and precision medicine [19], taken together as *Precision and Personalized medicine* (PPM).

1.1 Precision and personalized medicine

For many decades, cancer treatment was mostly applied as *one-size-fits-all* approach. In this approach, the most common types of standard cancer treatments are surgery, radiation therapy, chemotherapy. Chemotherapy has been the standard of care in treating many different types of cancers, and oftentimes may be the only treatment that a patient receives. Unfortunately, *one-size-fits-all* approach ignores high heterogeneity of cancers. A high diversity of cells is present within one person's tumor, and there are no two patients' cancers that are exactly the same [79, 12]. Hence, different patients may have variable responses to standart treatments such as chemotherapy and radiation.

Using patient's genetic patterns, lifestyles, drug responses, and environmental and cultural factors, *precision medicine* refers to the classification of people into subpopulations (e.g. patient stratification). Patient stratification is commonly used as a key factor in treatment decisions [36]. In the context of oncology, *precision medicine* takes into account molecular features (such as gene expression) of a genetic profile of patient. At the present time, precision oncology is moving towards including other characteristics (proteome, epigenome, microbiome, lifestyle, diet) for patient stratification [9].

Personalized medicine is often used as a synonym for *precision medicine*, although the term *individualized medicine* refers to truly personalized medicine, where treatment is individually adapted to a patient. Truly personalized medicine is yet to be implemented in a clinical setting and currently only feasible in limited situations [25].

The field of PPM would not exist without the major accomplishment of sequencing the human genome and development of microarray and *Next Generation Sequencing* (NGS) technologies. Contemporary high-throughput sequencing technologies serve for collecting transcriptomic information to quantify expression levels of genes. Gene expression provides a snapshot of a molecular status of a specific tumor tissue and can potentially be used for identifying predictive gene signatures and discovering biomarkers in cancer prognosis [55].

1.1.1 Example of data-driven biomarkers utilized in precision oncology

The text of this section is based on the part of the "Background" section of [16]. Breast cancer is one of the paradigmatic examples of the utility of high-throughput data to derive prognostic molecular signatures (PAM50, MammaPrint, OncotypeDX) [63, 70] for patient stratification. Based on the expression of 50 genes the PAM50 classifier is widely used to divide breast cancers into four main molecular subtypes: luminal A, luminal B, triple-negative/basal-like, and HER2-enriched [26]. Two luminal subtypes are characterized by high hormone receptor expression and generally have a better prognosis. The basal-like breast cancers are a heterogeneous group of hormone receptor- and HER2-negative breast cancers that are highly proliferative and often metastasize early. Stratification of patients according to the likelihood of metastasis can be performed by MammaPrint and OncotypeDX, which are 70- and 21-gene expression signatures. Even though established molecular signatures for breast cancer have prognostic impact and are used in clinical practise, a recent study [45] concludes that existing gene signatures of breast cancer lack a sensible biological meaning. Therefore, biomarker discovery is full of technical and principal challenges.

1.2 Challenges of biomarker discovery

One of the tasks of clinical cancer research is to identify prognostic gene signatures that are able to predict clinical outcome [29]. The clinical outcome is usually presented as a classification task performed by an ML model, and the discriminative features are considered as

potential biomarkers. Predictive gene signature is a feature subset driving the classification result of an ML model.

Gene expression data typically contains many more features (e.g. genes) than data points (e.g. patients). Such kind of problems are called *high-dimensional* and imply plenty of challenges for feature selection, that can be referred to as *curse of dimensionality* phenomenon. The essence of this phenomenon is that for the same quantity of data points, the increase of dimensionality will lead to more and more scarse distribution of these data points in the feature space. To achieve the same level of coverage in a high dimensional space as in a low dimensional one, the number of data points would need to increase exponentially (section 1.4 in [8]). Since the amount of training data is limited, the sparsity of data points leads to overfitting. Thus, the feature selection for the ML model becomes unstable, which means that the selected features vary across different datasets with the same clinical endpoint.

The stability of a feature selection algorithm is essentially the robustness of the algorithm's feature preferences. If small changes in training data lead to large changes in selected feature subsets then the feature selection approach is unstable. The stability can be quantified by providing different samples from the same training data and measuring the changes among chosen feature subset. An estimate of feature selection algorithm's stability can address the question — *how much can we trust the algorithm?*[52]. Besides, it is crucial for biomarker discovery to guarantee the reproducibility of the given feature selection methods [41].

Correlation structures, inherently present in gene expression measurements, also influence the gene selection procedures [34]. Firstly, majority of genes correlate with the clinical endpoint: even random gene expression signatures are significantly associated with the clinical outcome [77]. Secondly, if a gene is a good predictor and highly correlated to the clinical outcome, then other genes highly correlated to that gene are good predictors as well. The correlations between single genes and the clinical endpoint fluctuate strongly when measured over different subsets of patients [23], thus affecting the lists of candidates for predictive biomarkers.

Besides the instability of gene signatures, the members of prognostic gene lists do not necessarily relate to disease mechanisms. Manjang and colleagues [45] concluded that no gene signatures of breast cancer have a sensible biological interpretation in scopes of cancer pathology and underlying molecular processes. Not only the predictive outcome is of value, but also the biological meaning of the gene signatures [22].

Stability of gene signatures, as well as their biological interpretability, can be improved by including prior knowledge of molecular networks (e.g. biological pathways) into ML approaches [29, 57]. Molecular networks represent molecular processes in a given biological system and are widely used by biologists to interpret the results of a statistical analysis. The nodes or vertices of such networks depict molecules, which can be genes, RNAs, proteins and metabolites. The edges represent interactions between molecules. To approximate the interactions between genes different molecular networks can be used. Within this thesis, I used a *Protein-Protein Interaction* (PPI) network, where edges are undirected interactions

between pairs of binding proteins.

ML methods benefit from prior knowledge of a molecular network since neighboring genes are not treated as independent. This benefit is based on the hypothesis that adjacent genes in a molecular network should have similar expression profile [29]. Consequently, decisions of ML methods are driven by predictive subnetworks, rather than single genes. Since these subnetworks have genes connected according to the edges of a molecular network, the interpretability of the selected genes is improved. Furthermore, these subnetworks can differ from one patient to another according to their expression profiles, but convenient feature selection approaches select important for classification features that are the same for all patients and correspond to a whole gene expression dataset. In order to derive a patient-specific molecular subnetwork driving a single decision of a ML method, one can apply techniques of XAI to *Deep Learning* (DL) and ML models working with graph-structured data. Presenting interpretable patient-specific subnetworks to clinicians and researchers could possibly extend treatment options and promote the development of *individualized medicine*.

1.3 Deep learning on graphs and explainable artificial intelligence

1.3.1 Convolutional neural network as a deep learning approach

In recent years DL was successfully applied in various areas. These methods demonstrated state-of-the-art performance in visual object recognition, object detection, speech recognition, natural language processing as well as in other domains such as drug discovery and genomics [40]. DL methods aim to automatically learn data representations needed for a machine learning task. In other words, DL methods allow a model to be fed with raw data so that the representations needed for performing predictions are learned by a learning procedure [40]. Usually these representations are composed of several processing layers, and in many applications feedforward neural network architectures with multiple hidden layers are used. The word "deep" refer to neural networks with at least two hidden layers.

One of the feedforward *Neural Network*s (NNs), that was much easier to train and generalized much better than fully connected NNs, is CNN. This DL method shows cutting edge results for Euclidean structured data, which can have different spatial dimensionality: 2D for images, 3D for video and 1D for signals and sequences. CNNs capture local spatial patterns in natural signals and merge them into high-level abstractions.

The architecture of CNN usually consists of three types of layers: convolutional layers, pooling layers, and fully connected layers. The convolutional and pooling layers exploit the Euclidean structure of the data preparing representations (features) for the fully connected layers. The convolution operation filters localized patterns present in data that are valuable for a prediction task. The learnable weights of a filter compose a feature to be detected in

input data. One convolutional layer has dozens of filters, so that multiple features are learned in parallel. In case of grid-structured data like images, convolution operation quantifies similarity between a feature of a filter and a local group of pixels that can form the same pattern at different locations on the image. After that, a nonlinear activation function is applied to quantified similarities. As a result, the feature map is created for each filter that consists of values indicating the locations and the strength of a detected feature. While convolutional layer detects local motifs of features, the pooling layer merges semantically similar features into one [40]. For example, shifting a filter by 1 row or 1 column on the image can give slightly different feature values for the same pattern. These feature values are merged usually by computing the maximum or an average of a local patch of features. Thus, the pooling layer performs a dimensionality reduction and promotes invariance to small shifts of a filter.

Stacking several convolutional and pooling layers enables the learning of hierarchical features, where higher-level features are obtained by composing lower-level ones. For instance, in images, local combinations of edges create motifs, motifs assemble into parts, and parts form objects.

Deep learning and CNNs have been already used in the field of bioinformatics [46] and cancer research [76]. Deep feed-forward NNs were utilized for gene expression classification in [72, 2]. In [44, 20, 51] gene expression data transformed into images were utilized by CNNs for tumor type classification. However, the gene expression data itself is not structured. Nonetheless, the gene expression data can be structured by a molecular network (encoded as a graph) representing connections between genes. Each vertex of a molecular network is assigned a gene expression value as an attribute. The set of the attributes create a graph signal. The patients, which can be classified by an ML method, are represented as graph signals (gene expression data) on a single graph. Figure 1.1 shows an example of a graph signal created from a gene expression profile of one patient. The gene expression profiles of patients form different graph signal patterns that can be learned by the means of *Geometric deep learning*.

1.3.2 Geometric deep learning and graph neural networks

Currently, deep learning is extending to Non-Euclidean data domains. The extension is based on generalization of CNNs to graphs and manifolds [11], [50]. While graphs include molecular networks in biology and networks in social sciences, manifolds themselves are 2D surfaces embedded into 3D space. *Geometric deep learning* is a term for the approaches that generalize deep learning models to Non-Euclidean domains. A CNN that is generalized to graphs and capable of learning graph signal patterns is called GCNN. In this thesis I focus on graph domain. A more detailed description of different aspects of *geometric deep learning* can be found [11, 10].

GCNNs belong to a larger class of methods, *Graph Neural Network*s (GNNs). Two types of graph neural networks, GCNNs, and *Recurrent Graph Neural Network*s (RGNNs) are

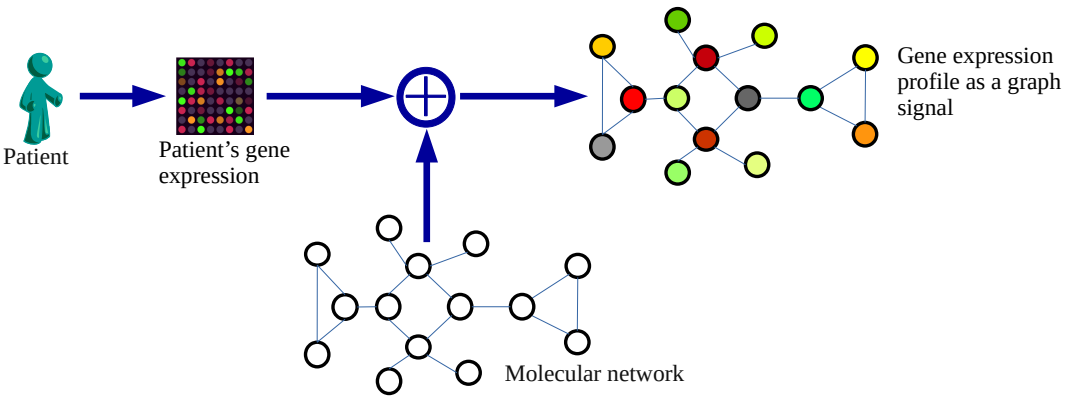

Figure 1.1: **Example of a graph signal**. Genes are mapped to the vertices of a molecular network. Gene expression values of a patient's profile are node attributes creating a graph signal.

trending within GNNs. Some other types of GNNs are present as well, and they differ by node featurization, ways of aggregating edge information (directed, undirected, edge label), utilizing attention mechanisms etc. A good overview on different GNNs is given in [80, 78].

One of the first works popularizing GNNs is from Scarselli and colleagues [65]. They developed recurrent GNN that learns representations of a graph and its nodes by propagating information from neighboring vertices and edges. Recurrent GNNs are based on a message passing approach used to compute states of nodes. The states of nodes are updated by exchanging neighborhood information recurrently. The resulting states are node representations: a node is mapped into a point in m-dimensional Euclidean space, creating a node embedding that can be useful for node classification problem. A good review of other methods for *graph representation learning* is given in [27].

In general, several graph analyses tasks can be performed by GNNs [80]. The following list is not complete:

1. **Node classification**. GNN infers a label of the node of interest, utilizing the node's attributes and its neighborhood. This task can be performed by GCNNs as well as RGNNs.

2. **Link prediction**. For a pair of nodes of interest, the task is to estimate the presence of the link between those vertices or to predict the connection strength. The representation of nodes (as output of a GNN) can be an input to a similarity function or feed-forward neural network.

3. **Graph classification/regression**. Having different graphs as possible inputs for a model, the task is to classify or regress those graphs. One of the examples is to predict properties of different molecules represented as graphs.

4. **Graph signal classification/regression**. The task of graph signal classification is conceptually similar to image classification, where pixel values of an image are assigned to the vertices of a regular grid. As described in Figure 1.1, features are connected by a graph that structures data. Feature values of a data point are assigned to the vertices of a graph creating a single graph signal. The graph is the same for every data point in a dataset.

In this thesis, I focus on graph signal classification as a task within *geometric deep learning*, where CNNs are generalized to graph-structured data.

1.3.2.1 Overview of GCNNs for graph signal classification

Generalizing CNNs to graphs has several requirements: (i) localized convolution filters on graphs, (ii) a graph coarsening procedure that groups together similar vertices for pooling operation, (iii) linear computational complexity w.r.t. to the input dimensionality (number of vertices in the graph). According to my knowledge, there are only three publications presenting GCNNs tailored to the aforementioned requirements.

One of the most prominent methods for graph signal classification is published by Defferard et al. [21]. This GCNN not only satisfies the requirements above, but also provides a possibility to select the size of the node's neighborhood participating in localized graph convolution. The graph convolution is based on spectral formulation and inspired by the emerging field of *Graph Signal Processing* (GSP) [69]. The convolutional filtering utilizes Chebyshev polynomials for computational stability. Order K of polynomials determines K-hop neighborhood around a node. Moreover, this GCNN was already used for classification of gene expression profiles structured by a PPI network in [61], [58].

The GCNN method of Levie and colleagues [42] is structurally similar to the method of Defferard et al. [21]. The first one utilizes parametric rational complex functions (Cayley polynomials) instead of Chebyshev polynomials to compute spectral filters. While being well-localized on the graph, the Cayley filters have narrower frequency bands with the same number of trainable parameters [42]. Levie and colleagues [42] show that the Cayley-based GCNNs slightly outperform GCNNs with Chebyshev-based filtering. This method is worth exploring in future works.

Deep learning technique from Zhang et al. [86] for graph classification implements convolution and pooling for graphs. Their approach is also suitable for graph signal classification. The drawback is that the convolution is always performed over a 1-hop neigboorhood of a graph's node. It is similar to the convolution in the method of Kipf and Welling [32], which is commonly used for node classification.

So far, there is a lack of methods that can perform graph signal classification and satisfy the described earlier requirements (i-iii). For the list of other RGNNs and GCNNs, that do not satisfy the mentioned requirements (i-iii) the reader could be referred to [80, TABLE III of].

Finally, I choose GCNN from Defferard et al. [21] for predictive modelling. The graph convolution of this method is less complex and more computationally efficient than the graph convolution designed by Levie et al. [42]. Also, GCNN from Defferard et al. [21] has an advantage over the method [86] by allowing a researcher to select the size of a node's neighborhood participating in localized graph convolution. Further, I am going to adapt explanation methods to GCNN to interpret its individual decisions, and the overview of the explanation methods is given in the next section.

1.3.3 Explainable artificial intelligence

Deep NNs consist of several nonlinear processing layers modelling complex interactions between the input and output variables. This complexity does not allow to directly understand the mechanism by which a model makes its decision. Thus NN is a black-box ML model that does not provide interpretable insights on its internal machinery.

Explanations supporting the decisions proposed by NNs are crucial, e.g., in precision medicine [5], where experts in the clinical domain require more information from the model than a simple binary outcome [75, 82]. The EU's recent General Data Protection Regulation (GDPR) restricted automated decision making produced by algorithms [53]. Article 13 of GDPR [53] states that clinics should provide patients with "meaningful information about the logic involved". Article 22 of [53] indicates that a patient shall have the right not to be subject to an automated decision unless the patient gives a consent with it (paragraph 2.c). Therefore, the explainability of deep neural networks becomes an imperative for clinical applications [16].

Explainability refers to post-hoc explainability, where explanation techniques are used to convert a non-interpretable ML model into an explainable one [5]. Nonetheless, explainability of a model is not the same as the interpretability. Interpretability comes from the design of the model itself (e.g linear models). Taking the difference between explainability and interpretability into account, XAI can be defined as follows [5]: *Given an audience, an explainable artificial intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

In this thesis the post-hoc explanation methods are considered. These techniques explain individual decisions of ML models when a single data point is fed as input. An explanation method generates relevance measure (could be also called influence, and importance score [5]) for each feature value of a specific data point. The relevance measure shows how much a particular feature influences the classifier's individual decision. I used as a classifier the GCNN method published by Defferard et al. [21]. This method utilizes a prior knowledge of a molecular network, which was in particular a PPI network used in this thesis. GCNN performs patient classification based on patient's gene expression profiles. Since genes are mapped to the vertices of a PPI network, the patients are represented as graph signals. For a single patient, an explanation method computes the relevance values for input features (genes) that are also the vertices of the PPI network. Out of the genes with the highest rele-

vance, the patient-specific subnetwork can be constructed explaining an individual decision of a GCNN (Figure 1.2).

1.3.3.1 Overview of explanation methods

The text of this section is based on the on the part of the "Background" section of [16]. Explanation methods that explain individual decisions of complex ML models in terms of input variables usually use one of two available approaches [48]: *functional* or *message passing*. The idea of *functional* approach is to produce explanations out of local analysis of a particular prediction. The *message passing* approach infers explanations by running a backward pass in a computational graph, which generates a prediction as its output. The first group of methods exploiting *functional* approach includes the sensitivity analysis, Taylor series expansion, as well as the model agnostic approaches *Local Interpretable Model-agnostic Explanations* (LIME) [62] and SHAP [43]. The second group of methods favour *message passing* approach: [84, 4, 68, 71]. This group also includes the method of integrated gradients [73], which can be easily implemented. This method cumulates non-local as well as local to a particular data point gradients and satisfies properties that a reasonable explanation method should have [73].

The LRP method [4] combines *functional* and *message passing* approaches to produce relevance values for each input feature. The relevances are generated for each data point (within this thesis each patient) individually. A version of LRP utilizing a special propagation rule [49] can be interpreted as a part of the framework of deep Taylor decomposition [48] and as an explanation method called excitation backpropagation [85].

LRP exhibited promising results on image data and has been applied in cancer research to discover prognostic biomarkers. Klauschen et al. [33] applied LRP for visual scoring of Tumor-Infiltrating Lymphocytes (TIL) on Hematoxylin&Eosin breast cancer images. Binder et al. [6] utilized LRP to identify spatial regions (cancer cell, stroma, TILs) on morphological tumor images. Identified spatial regions explained predictions of molecular tumor properties (like protein expression).

There are also some explanation methods tailored to GNNs. In [81, 67, 56] the authors provided explanation methods that are applicable only for GCNN of Kipf and Welling [32]. This GCNN implements graph convolution as a simplified version of the graph convolution proposed in [21], which is the method that I use for predictive modelling. Ying et al. [83] suggested the model-agnostic GNNExplainer. This approach is suitable for explaining node classification, link prediction, and graph classification. But Ying et al. did not consider an application of their approach to classify graph signals [15, 61]. The GNN-LRP method [66] provides explanations in form of scored sequences of edges on the input graph (i.e. relevant walks). This sequence of edges represents a path extracted from the input to the output of GNN and brings insights for GNN's decision strategy. Such explanations are useful for graph classification tasks, where a data point is an individual graph. In this thesis, patients are represented as graph signals on a single graph, so that this method is not applicable.

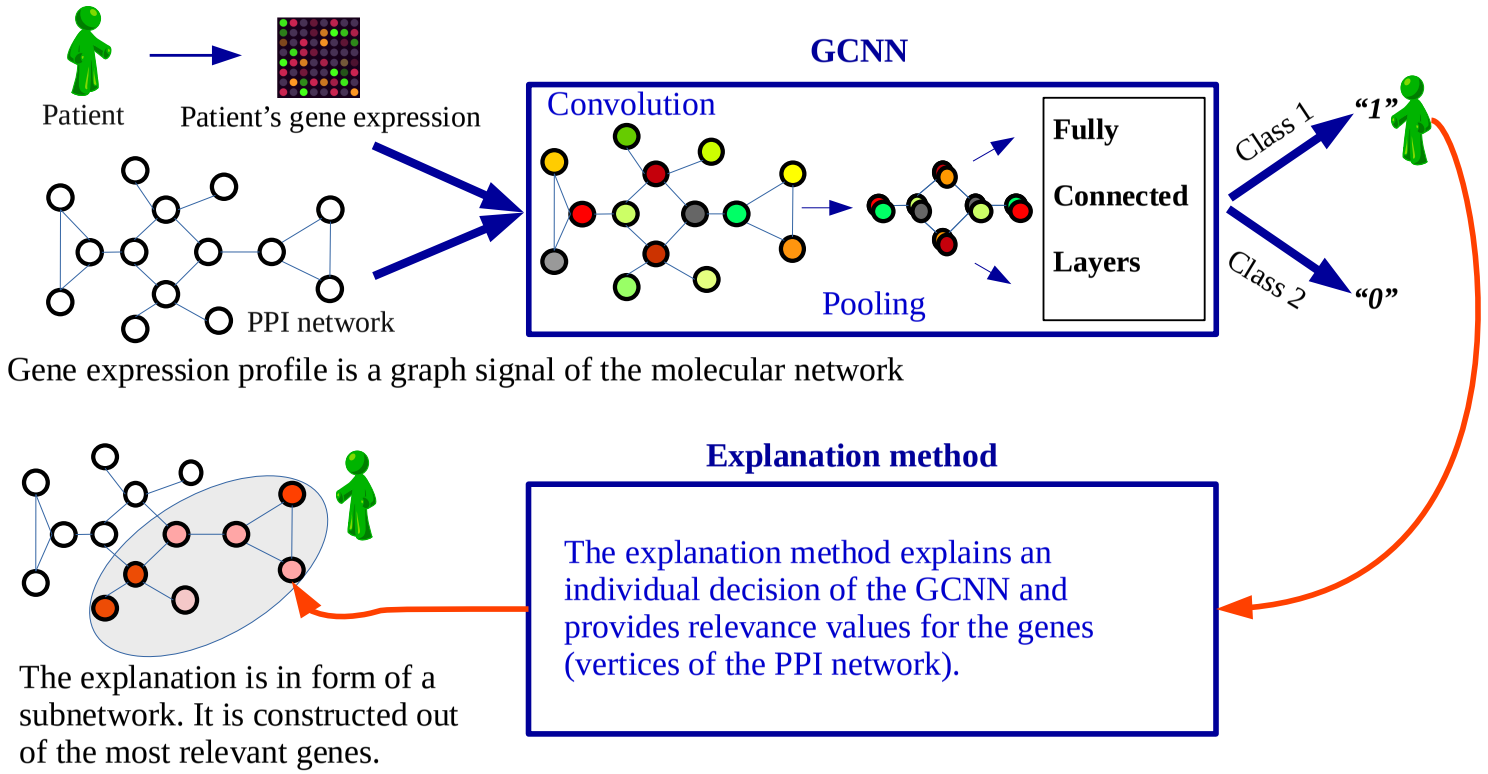

Figure 1.2: **Explaining a patient-specific decision of a GCNN**. A data point is represented by a gene expression profile of a patient. Gene expression values are assigned to the nodes of the molecular network (as attributes) so that the patient is represented as a graph signal. Trained GCNN classifies the patient. The explanation method is applied to the trained GCNN as a post-hoc processing step, assigning relevance values to the input features (genes, as vertices of the molecular network). Most relevant genes constitute a patient-specific subnetwork. This figure is inspired by [16, Figure 1 of].

Since the LRP explanation method has shown great results in explaining the decisions of CNNs [49], and have already been used for biomarkers identification (on image data), I adapted LRP to the graph convolution of the geometric deep learning approach from [21]. The graph coarsening and pooling of the method of Defferrard et al. [21] is implemented in an efficient way. The vertices of a graph (and coarsened versions of it) are rearranged as a binary tree structure, so that the pooling of a graph signal is analogical to pooling of 1D signal. Therefore, there is no need to adapt the LRP method to the procedure of graph coarsening.

1.4 Aims and organization of the thesis

1.4.1 Primary research aim

The overall aim of the thesis was to develop a methodology that allows for extracting patient-specific molecular subnetworks as explanations for classification decisions of a ML approach utilizing prior knowledge of molecular networks. GCNN [21] served as a ML approach, and its decisions were explained by the developed GLRP method. The GLRP method is an adaptation of LRP to convolutional layers of the GCNN method. Here and further in this thesis, GCNN+LRP and GLRP are used as synonyms. The GCNN+LRP approach generates patient-specific explanations in form of subnetworks. These subnetworks can be visualized and interpreted in a biomedical context on the individual patient level and possibly extend treatment options. The biologically interpretable subnetworks could be useful for precision medicine approaches such as for example the molecular tumorboard. The primary research aim can be decomposed into the following objectives:

- Development of an ML-based methodology that utilizes prior knowledge of a molecular network and allows for extracting patient-specific molecular subnetworks (Figure 1.2)
  - Verification of GCNN's applicability [21] to predictive modeling with gene expression data
  - Development of the GLRP explanation method for GCNN [21] as an adaptation of LRP to the GCNN's graph convolutional layers
- Implementation of the designed methodology as an open-source software
- Computational validation of the developed methodology generating patient-specific subnetworks
  - Validation of biological meaning in subnetwork' genes prioritized by GLRP
  - Verification of concordance between subnetwork' genes and clinical knowledge from the perspective of precision oncology

Apart from generating patient-specific subnetworks, one can utilize the GLRP method as a general feature selection approach. Moreover, other explanation methods (see Figure

1.2), such as SHAP can be applied to GCNN to explain its individual decisions. Thus, the GCNN+SHAP approach can also deliver patient-specific subnetworks and be used as a feature selection technique.

1.4.2 Secondary research aim

Another aim of the thesis was to investigate the stability of feature selection based on the proposed methodology (see previous section) as well as to compare the GCNN+LRP and GCNN+SHAP approaches. The objectives of the secondary research aim are the following:

- Estimation and comparison of the stability of feature selection performed by GCNN+LRP, GCNN+SHAP, and other ML-based approaches. They include the "glmgraph" method [14] utilizing molecular network information, and methods that do not use any prior knowledge: Random Forest, *Multi-layer Perceptron* (MLP)+LRP, and MLP+SHAP
  - Presentation of Keras [18] compatible graph convolutional layer of the GCNN method [21] to provide the applicability of SHAP to GCNN
  - Application of the SHAP explanation method to GCNN models implemented as Keras *Sequential* instances
- Comparison and analyses of the subnetworks generated by GCNN+LRP and GCNN+SHAP
  - Estimation of how important the patient-specific subnetworks are for driving classification results
  - Comparison of the connectivity of the subnetworks

1.4.3 Organization of the thesis

This thesis is organized as follows: the first publication Chapter 2 shows the utility of GCNN for classifying breast cancer patients using their gene expression profiles structured by a PPI network. This step is necessary and included into the *Primary research aim* (section 1.4.1). The ML task is to predict the occurrence of metastatic events. The aim of the first publication was to compare the classification performance of GCNN with other ML methods. The performance of GCNN is comparable to that of other methods which approves the application of GCNN to high-dimensional gene expression data.

The second publication Chapter 3 introduces a novel approach for generating patient-specific molecular subnetworks, presents a new GLRP method, and covers the objectives of the *Primary research aim* (section 1.4.1). The GLRP method explains individual decisions of a GCNN performing multinomial classification tasks on graph-structured data. A sanity check of the developed GLRP method was demonstrated on a hand-written digits dataset. The biological validation was performed by applying the developed methodology to gene

expression data from *Human Umbilical Vein Endothelial Cells* (HUVEC)s treated or not treated with tumor necrosis factor alpha. To show the utility of introduced methodology in the scopes of *precision medicine*, it was applied to the third, large breast cancer dataset. The generated patient-specific subnetworks potentially enabled the extension of possible drug targets and treatment options.

The third publication Chapter 4 utilizes the developed GLRP method to estimate its stability as a feature selection approach, compares it to GCNN+SHAP and other ML-based feature selection approaches, and investigates the properties of molecular subnetworks w.r.t permutation feature importance and connectivity. The analyses of the GCNN+LRP approach was done according to the objectives of section 1.4.2 *Secondary research aim* and could identify its advantages over other feature selection techniques. The results of the aforementioned publications are discussed in Chapter 5 and the thesis is wrapped up in the conclusion remarks in Chapter 6.

2 Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer

Reference
Chereda H, Bleckmann A, Kramer F, Leha A, Beissbarth T. Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer. Stud Health Technol Inform. 2019 Sep 3;267:181-186. doi: 10.3233/SHTI190824

Original Contribution
The conceptualization of the study was conceived by Hryhorii Chereda, Frank Kramer, and Tim Beißbarth. Computational experiments were performed by Hryhorii Chereda. Hryhorii Chereda generated the figure, prepared the results and wrote the manuscript. Annalen Bleckmann, Tim Beißbarth, Andreas Leha, and Frank Kramer performed review and editing.

Utilizing molecular network information via Graph Convolutional Neural Networks to predict metastatic event in breast cancer

Hryhorii CHEREDA[a], Annalen BLECKMANN[a,b,c], Frank KRAMER[d], Andreas LEHA[a] and Tim BEISSBARTH[a,1]
[a] *Medical Bioinformatics, University Medical Center Göttingen*
[b] *Hematology & Medical Oncology, University Medical Center Göttingen*
[c] *Internal Medicine-A (Hematology, Oncology, Hemostaseology and Pulmonology), University Hospital Muenster*
[d] *IT Infrastructure for Translational Medical Research, University of Augsburg*

Abstract. Gene expression data is commonly available in cancer research and provides a snapshot of the molecular status of a specific tumor tissue. This high-dimensional data can be analyzed for diagnoses, prognoses, and to suggest treatment options. Machine learning based methods are widely used for such analysis. Recently, a set of deep learning techniques was successfully applied in different domains including bioinformatics. One of these prominent techniques are convolutional neural networks (CNN). Currently, CNNs are extending to non-Euclidean domains like graphs. Molecular networks are commonly represented as graphs detailing interactions between molecules. Gene expression data can be assigned to the vertices of these graphs, and the edges can depict interactions, regulations and signal flow. In other words, gene expression data can be structured by utilizing molecular network information as prior knowledge. Here, we applied graph CNN to gene expression data of breast cancer patients to predict the occurrence of metastatic events. To structure the data we utilized a protein-protein interaction network. We show that the graph CNN exploiting the prior knowledge is able to provide classification improvements for the prediction of metastatic events compared to existing methods.

Keywords. Gene expression data, classification, CNN, prior knowledge, molecular network.

1. Introduction

Technologies as microarray gene-expression profiling and next-generation sequencing are becoming more and more available and play a significant role in cancer prognosis, for example in discovering individual biomarkers [1]. Furthermore, high-throughput technologies produce huge amounts of data that can be used for assessment of metastatic events. At the moment, deep learning techniques are well known to show prominent results in many research fields with big and complex data.

In recent years deep learning was applied to a wide range of problems in various areas. Deep learning methods are aimed at the automatic learning of data representa-

[1] Corresponding Author, Tim Beißbarth, University Medical Center Göttingen, Medical Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany; E-mail: tim.beissbarth@ams.med.uni-goettingen.de.

tions (features) needed for machine learning task. These methods demonstrated state-of-the-art performance in visual object recognition, object detection, speech recognition as well as other domains such as drug discovery and genomics [2]. One of the most popular methods of deep learning are Convolutional Neural Networks (CNN). They show cutting edge results for data that are spatially structured. Different classes of such data have different spatial dimensionality: 2D for images, 3D for video and 1D for signals and sequences. The main property of CNNs is a capability of capturing local spatial patterns in natural signals and merging them into high-level abstractions.

The usual CNN architecture consists of three types of layers: convolutional layers, pooling layers, and fully connected layers. The first two layers utilize the Euclidean structure of the data preparing informative features for the fully connected neural network layers. For grid-structured data as images, the convolution layer performs filtering operation to extract highly correlated local groups of pixels forming the same pattern in different parts of the image. A nonlinear function is applied to each output of filtering. As a result, the feature map is created per each filter, consisting of the feature values based on the same pattern. As for the pooling layer, since the slightly shifted position by 1 row or 1 column can give slightly different feature values for the same pattern it merges the feature values into one [2]. Usually this operation is performed by computing the maximum of a local patch of features. In such a way, the dimensionality reduction and the gain of invariance to small shifts are performed.

Deep learning and CNNs are already used in the field of bioinformatics [3]. As an example, CNNs were applied to gene expression data for tumor type classification [4]. One should notice that in Lyu and Haque [4] the gene expression data were transformed into images and then CNNs were applied to them. In general, gene expression data do not have any spatial structure, and the number of genes is much higher than the number of patients that might lead to poor classification performance on the test set. Thus to deal with this problem, still approaches are needed that utilize prior knowledge based on known interactions in molecular networks. Here we demonstrate that the classification performance can be improved by a combination of deep learning and prior biological knowledge.

Nowadays, deep learning is extending to Non-Euclidean domains. This extension is based on generalization of CNNs [5] to graphs and manifolds. We applied graph CNN [6] to gene expression data structured by a molecular network representing the connection between genes. In other words, since each vertex of a molecular network is assigned a gene expression value, we are performing a graph-signal classification task. Recently, quite similar methodology was applied to classify breast cancer subtypes utilizing gene expression data structured by protein-protein interaction network [7]. Breast cancer is one of the three most common cancers in industrialized countries [8]. Patients often develop metastases that limit survival, as there has not been any curative therapy for them [9]. We show that graph CNN outperforms more classical machine learning methods at the prediction of metastatic events in breast cancer.

2. Materials and Methods

2.1. Breast Cancer Data

We used the breast cancer patient data previously studied and preprocessed in research [10]. The data consist of 10 public microarray datasets measured on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A arrays. The datasets have accession num-

bers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, GSE6532 and are available from the Gene Expression Omnibus (GEO) [11] data repository. The RMA probe-summary algorithm [12] was used to process each dataset after which they were combined together on the basis of HG-U133A array probe names and quantile normalization was applied over all datasets. In the case of few probes mapped to one gene the probe with the highest average value was taken. In the end, we ended up with 12179 genes per each patient. Further, patients with and without metastatic events were selected to formulate two classes for the prediction task: 393 patients with metastasis within the first 5 years, 576 patients without metastasis having the last follow up between 5 and 10 years.

2.2. Protein-Protein Interaction Network

We used the Human Protein Reference Database (HPRD) protein-protein interaction (PPI) network [13] to structure the gene expression data. This PPI network consists of binary interactions between pairs of proteins and can be represented as an undirected graph. One should notice that this graph is not connected. The genes from gene expression data can be mapped to the vertices of the PPI network. In such a way, the resulting PPI graph has 7168 vertices (genes) matched, and it has 207 connected components. The main connected component has 6888 vertices, and each of the 206 other components has from 1 to 4 vertices. The graph CNN requires graph to be connected so all the machine learning methods had 6888 genes as an input.

2.3. Problem formulation

Initially, the problem is formulated as a binary classification of gene expression data $X \in R^{m \times n}$ to target variable $Y \in R^m$ representing the occurrence of metastatic event. m is a number of samples (patients) and n is a number of features (genes). Additionally, we incorporate the information of the molecular network which is represented as a undirected graph $G = (V, E, A)$, where V and Ecorrespond to the sets of vertices and edges respectively. A is an adjacency matrix. The number of vertices is equal to the number of genes n. A row x of gene expression matrix X contains data from one patient and can be mapped to the vertices of the graph G. The values of x are interpreted as a graph signal.

2.4. Graph Convolutional Neural Network and Multilayer Perceptron

The graph CNN [6] captures localized patterns of a graph signal via convolution and pooling operations performed on a graph. The convolution operation bases on the spectral graph theory utilizing the convolution theorem and graph Fourier transform. The graph convolutional filter can be approximated by a parameterized expansion of Chebyshev polynomials of graph frequencies [6]. Such filter of polynomial degree k localizes the signal pattern in K-hop neighboring nodes. For the pooling operation, the graph is coarsened exploiting a graph clustering technique. We applied the graph CNN with following hyperparameters for learning. Two convolutional layers were used with 32 convolutional filters and polynomial degree 8 per each layer. Maximum pooling of size 2 applies to both of the convolutional layers. Two fully connected layers have 512 and 128 units consequently. ReLU (rectified linear unit) activation function was used and cross entropy loss was minimized. Application of usual CNN is not straightforward for gene expression data since it is not spatially ordered. Therefore, we applied deep

Multilayer Perceptron implemented in Keras [14], on the same set of genes but without prior knowledge structuring the data. The hyperparameters of our deep neural network are the following: 4 hidden layers and each of them consist of 1024 units with ELU (exponential linear unit) activation function. Cross entropy loss was minimized.

2.5. Random Forest and Lasso Logistic Regression

Random Forest and lasso penalized Logistic Regression were used as baseline methods. Random Forest is a tree-based ensemble machine learning technique combining bagging and random subspace method. It is widely used for high-dimensional data analysis, and considered as a standard tool for class prediction and gene selection with microarray data [15]. Logistic regression with lasso regularization is another classical method for classification of high-dimensional data. Lasso penalty allows shrinking of some coefficients to zero so that the variable selection is automatically performed. For the both baseline methods we utilized RandomForestClassifier and LogisticRegression classes implemented in Scikit-learn package [16].

3. Results

3.1. Our approach

Our approach is to structure gene expression data by applying it to prior knowledge on molecular interactions and to feed this structured data as input for the graph CNN deep learning method. The workflow for predicting metastasis events is shown in Figure 1.

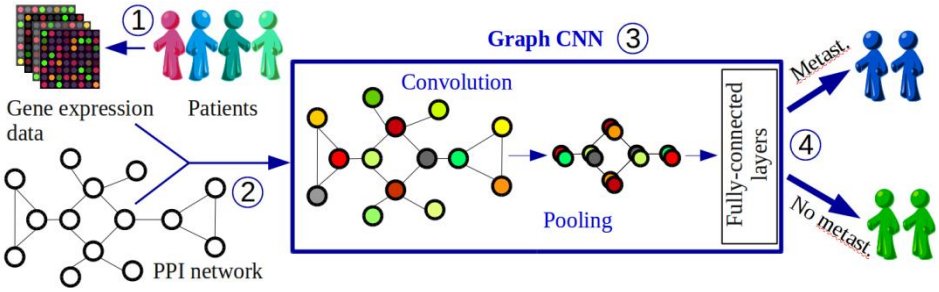

Figure 1. The schema of suggested workflow: 1. Patients' microarray data is preprocessed. 2. Genes are mapped to the vertices of PPI network. 3. The graph CNN processes gene expression data as graph signals. 4. The graph CNN predicts whether the patient is getting metastases during the first 5 years or not.

The endpoint is to predict the occurrence of a metastatic event for a patient. In other words, to classify patients into 2 groups, metastatic and non-metastatic. The first group corresponds to patients with metastasis within the first 5 years and the second concerns patients who are metastasis-free within first 5 years. Graph CNNs were developed recently, and according to the knowledge of the authors the approach described in the paper was not used for metastatic event prediction.

3.2. Comparison of machine learning methods

We compared the graph CNN approach with Multilayer Perceptron, Random Forest and Lasso Logistic Regression. The performance was assessed by 10-fold cross validation. For each of the data splits the model was trained on 9-folds and the classification was evaluated using 10th fold as a validation set. For training, the input was standar-

dized and the validation sets were scaled according to the means and standard deviations on the training set. For each machine learning algorithm the hyperparameters were the same. For each data split the graph CNN and Multilayer Perceptron were trained on the same number of epochs. In this paper we used the most common metrics: area under ROC curve (AUC), accuracy and F1-weighted score. The metrics were averaged over folds and the standard errors of their means were calculated (Table 1).

Table 1. Performance comparison of machine learning methods on metastatic event prediction.

Method	100*AUC	Accuracy, %	F1-weighted, %
Graph CNN	82.16±1.25	76.18±1.36	75.86±1.35
Random Forest	81.40±1.76	74.74±1.67	74.00±1.82
Multilayer perceptron	81.01±1.84	73.92±1.48	73.64±1.54
Lasso Logistic Regression	80.95±1.61	74.74±1.27	74.53±1.27

The graph CNN demonstrates higher values for all three metrics estimating the quality of metastatic event prediction. In such a way we show that utilization of prior knowledge into graph CNN is beneficial in comparison to standard machine learning methods for discriminating classes of patients with or without metastases within 5 years after treatment.

4. Discussion

We demonstrated that the graph CNN applied to graph-structured data predicts metastatic event better than other classical methods that are trained on the same set of features (gene set) and that do not incorporate any prior knowledge. We predicted the occurrence of metastatic events in a breast cancer data set. We have shown that even under the limitations of available data (from deep learning perspective) graph CNN could still outperform other methods. It is well known for breast cancer that molecular subtypes show metastatic differences [10] and thus molecular subtypes influence metastasis-free survival. However, additional confounding factors (e.g. age) may exist that mask the association between input and output variables. The consideration of such confounding factors may be additional future work to consider to evaluate the practical value of such a classifier. Turning event times into a binary endpoint might lead to information loss. One could adapt our method to predict metastasis-free survival.

To structure the gene expression data we utilized only the main connected component of the PPI graph. The majority of other vertices are just single nodes of the PPI graph, thus the prior knowledge of the molecular network does not structure them. In future work we consider utilization of the rest of genes that were not mapped to the main connected component as additional input units of fully-connected layer of graph CNN. The authors in Rhee et al [7] applied the graph CNN to RNA-seq gene-expression data structured by the PPI network extracted from STRING database [17] to predict breast cancer subtypes. The STRING PPI network contains weights on pairs of proteins that interact with each other. 4303 genes were selected. It was also shown that graph CNN could outperform the baseline machine learning methods for the specified classification task. In our case, we have 6888 genes and a binary topology which lead to the hypothesis that graph CNN is able to capture meaningful data representation even if edges do not have weights. For future work we are planning to check how the weighted graph of STRING PPI would improve the classification performance and compare the two methods.

5. Conclusion

In this study we showed that graph CNN applied to microarray gene expression data structured by PPI network outperforms other machine learning methods that do not use any prior knowledge.

6. Conflict of Interest

The authors declare no conflict of interest.

7. Acknowledgements

This work was funded by the German Ministry of Education and Research (BMBF) e:Med projects *MyPathSem* (031L0024A), *Her2Low* (031A429C) and *MMML-Demonstrators* (031A428B).

References

[1] J. Perera-Bel, A. Leha, T. Beißbarth, Bioinformatic Methods and Resources for Biomarker Discovery, Validation, Development, and Integration: Applications in Precision Medicine, In S. Badve, G. Kumar, *Predictive Biomarkers in Oncology*, Springer, Switzerland, 2019. doi:10.1007/978-3-319-95228-4_11

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, *Nature* **521** (2015), 436-444. doi:10.1038/nature14539

[3] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, *Briefings in Bioinformatics* **18** (2017), 851–869. doi:10.1093/bib/bbw068

[4] B. Lyu, A. Haque, Deep Learning Based Tumor Type Classification Using Gene Expression Data, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 89-96. doi:10.1145/3233547.3233588

[5] F. Monti et al, Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5115-5124. doi: 10.1109/CVPR.2017.576

[6] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, *Advances in Neural Information Processing Systems* (2016), 3844–3852.

[7] S. Rhee, S. Seo, S. Kim, Hybrid Approach of Relation Network and Localized Graph Convolutional Filtering for Breast Cancer Subtype Classification. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization* (2018), 3527–3534. doi:10.24963/ijcai.2018/490

[8] J. Ferlay et al, Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012, *International Journal of Cancer* **136** (2015), 359-386. doi: 10.1002/ijc.29210

[9] F. Bray et al, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer Journal for Clinicians* **68** (2018), 394-424. doi: 10.3322/caac.21492

[10] Bayerlová et al, Ror2 Signaling and Its Relevance in Breast Cancer Progression, *Frontiers in Oncology* **7** (2017). doi:10.3389/fonc.2017.00135.

[11] T. Barrett et al, NCBI GEO: archive for functional genomics data sets – update, *Nucleic Acids Res* **41** (2013), 991–995. doi:10.1093/nar/gks1193

[12] R. A. Irizarry et al, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** (2003), 249–264. doi:10.1093/biostatistics/4.2.249

[13] T. S. K. Prasad et al, Human Protein Reference Database - 2009 Update, *Nucleic Acids Research* **37** (2009), 767-772. doi:10.1093/nar/gkn892

[14] F. Chollet, Keras, *GitHub* (2015). https://github.com/fchollet/keras

[15] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7** (2006). doi:10.1186/1471-2105-7-3

[16] F. Pedregosa et al, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.

[17] D. Szklarczyk et al, String v10: protein–protein interaction networks, integrated over the tree of life, *Nucleic acids research* **43** (2014), 447–452. doi:10.1093/nar/gku1003

3 Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer

Reference

Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, Auer F, Kramer F, Leha A, Beißbarth T. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. Genome Med. 2021 Mar 11;13(1):42. doi: 10.1186/s13073-021-00845-7

Original Contribution

HC and TB designed the study. HC developed and implemented the approach and performed the computational experiments. AB, FK, and PS provided major contributions to the study design. AB and KM provided clinical insights as well as JPB and PS provided biological insights, performing analyses of patient-specific subnetworks. FA developed the web-site to visualize the subnetworks. TB and AL provided machine learning insights. HC, TB, AL, KM, and PS wrote the manuscript. All authors read and approved the final manuscript.

Chereda *et al. Genome Medicine* (2021) 13:42
https://doi.org/10.1186/s13073-021-00845-7

Genome Medicine

RESEARCH Open Access

Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer

Hryhorii Chereda[1], Annalen Bleckmann[2], Kerstin Menck[2], Júlia Perera-Bel[3], Philip Stegmaier[4], Florian Auer[5], Frank Kramer[5], Andreas Leha[6] and Tim Beißbarth[1,7*]

Abstract

Background: Contemporary deep learning approaches show cutting-edge performance in a variety of complex prediction tasks. Nonetheless, the application of deep learning in healthcare remains limited since deep learning methods are often considered as non-interpretable black-box models. However, the machine learning community made recent elaborations on interpretability methods explaining data point-specific decisions of deep learning techniques. We believe that such explanations can assist the need in personalized precision medicine decisions via explaining patient-specific predictions.

Methods: Layer-wise Relevance Propagation (LRP) is a technique to explain decisions of deep learning methods. It is widely used to interpret Convolutional Neural Networks (CNNs) applied on image data. Recently, CNNs started to extend towards non-Euclidean domains like graphs. Molecular networks are commonly represented as graphs detailing interactions between molecules. Gene expression data can be assigned to the vertices of these graphs. In other words, gene expression data can be structured by utilizing molecular network information as prior knowledge. Graph-CNNs can be applied to structured gene expression data, for example, to predict metastatic events in breast cancer. Therefore, there is a need for explanations showing which part of a molecular network is relevant for predicting an event, e.g., distant metastasis in cancer, for each individual patient.

Results: We extended the procedure of LRP to make it available for Graph-CNN and tested its applicability on a large breast cancer dataset. We present Graph Layer-wise Relevance Propagation (GLRP) as a new method to explain the decisions made by Graph-CNNs. We demonstrate a sanity check of the developed GLRP on a hand-written digits dataset and then apply the method on gene expression data. We show that GLRP provides patient-specific molecular subnetworks that largely agree with clinical knowledge and identify common as well as novel, and potentially druggable, drivers of tumor progression.

Conclusions: The developed method could be potentially highly useful on interpreting classification results in the context of different omics data and prior knowledge molecular networks on the individual patient level, as for example in precision medicine approaches or a molecular tumor board.

Keywords: Gene expression data, Explainable AI, Personalized medicine, Precision medicine, Classification of cancer, Deep learning, Prior knowledge, Molecular networks

*Correspondence: tim.beissbarth@bioinf.med.uni-goettingen.de
[1]Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[7]Campus-Institute Data Science (CIDAS), University of Göttingen, Göttingen, Germany
Full list of author information is available at the end of the article

Background

Gene expression profiling by microarrays or next-generation sequencing has played a significant role in identifying predictive gene signatures and discovering individual biomarkers in cancer prognosis [1]. High-throughput sequencing produces huge amounts of gene expression data that can potentially be used for deriving clinical predictors (e.g., predicting occurrence of metastases) and identifying novel drug targets. Breast cancer is one of the paradigmatic examples of the utility of high-throughput data to derive prognostic molecular signatures (PAM50, MammaPrint, OncotypeDX) [2, 3] that predict clinical outcome. Based on the expression of 50 genes, the PAM50 classifier is widely used to divide breast cancers into four main molecular subtypes: luminal A, luminal B, triple-negative/basal-like, and HER2-enriched [4]. While the two luminal subtypes are characterized by high hormone receptor expression and generally have a better prognosis, the basal-like breast cancers are a heterogeneous group of hormone receptor- and HER2-negative breast cancers that are highly proliferative and often metastasize early. MammaPrint and OncotypeDX are 70- and 21-gene expression signatures that stratify patients according to the likelihood of metastasis. Although molecular signatures have prognostic impact, a more complete analysis of the molecular characteristics in the individual patient is required for personalized breast cancer therapy [2]. We hypothesize that molecular signatures can differ from one patient to another due to the heterogeneity of breast cancers. Such molecular signatures can be depicted as patient-specific subnetworks that are parts of a molecular network representing background knowledge about biological mechanisms. Presenting interpretable patient-specific subnetworks to clinicians and researchers enables better interpretability of the data for further medical and pharmaceutical insights, and possibly, for extended treatment options.

From a machine learning (ML) perspective, the prediction of a clinical outcome is a classification task, and molecular signatures can be identified as discriminative features. One drawback is that the search for molecular signatures is based on high-dimensional gene expression datasets, where the number of genes is much higher than the number of patients. The "curse of dimensionality" leads to instability in the feature selection process across different datasets. Stability can be improved including prior knowledge of molecular networks (e.g., pathways) into ML approaches [5]. ML methods benefit from pathway knowledge since neighboring genes are not treated as independent but instead similarities among adjacent genes, which should have similar expression profiles, are captured [6].

The essence of our classification task is to predict an occurrence of distant metastasis based on gene expression data structured by a molecular network (encoded as a graph) representing connections between genes. The patients are represented as graph signals (gene expression data) on a single graph. Since each vertex of a molecular network has a corresponding gene expression value as an attribute, we perform a graph signal classification task. Patients' gene expression profiles create different graph signal patterns that can be learned by the means of deep learning.

In recent years, deep learning has been widely applied on image data using convolutional neural networks (CNNs). The CNNs exploit the grid-like structure of images and cannot directly process data structured in non-Euclidean domains. Examples of non-Euclidean data domains include networks in social sciences and molecular networks in biology. Recently, deep learning methods extended to domains like graphs and manifolds [7]. Graph-CNN [8] learns graph signal patterns and can be applied to our graph signal classification task.

Deep neural networks are able to model complex interactions between the input and output variables. This complexity does not allow to track what role a particular input feature plays in the output; thus, a neural network itself as a black-box ML model does not provide interpretable insights.

On the other hand, decisions proposed by neural networks have to be explained before they can be taken into account in the clinical domain [9]. The European Union's recent General Data Protection Regulation (GDPR) restricted automated decision making produced by algorithms [10]. Article 13 of [10] specifies that clinics should provide patients with "meaningful information about the logic involved". Article 22 of [10] states that a patient shall have the right not to be subject to an automated decision unless the patient gives a consent with it (paragraph 2.c). Therefore, the explainability of deep neural networks becomes an imperative for clinical applications.

Explanation methods aim at making classification decisions of complex ML models interpretable in terms of input variables. These methods use one of two available approaches [11]: functional or message passing. The first group of methods produces explanations out of local analysis of a prediction. It includes the sensitivity analysis, Taylor series expansion, and the model agnostic approaches LIME [12] and SHAP [13]. The second group [14, 15] provides explanations by running a backward pass in a computational graph, which generates a prediction as its output. The Layer-Wise Relevance Propagation (LRP) method [15] combines through the framework of deep Taylor decomposition [11] functional and message passing approaches to generate relevances of each input feature. For a fixed input feature, the relevance shows how much this feature influences the classifier's decision.

The relevances are generated for each data point (in our application each patient) individually.

In image data, LRP exhibited promising results and has been applied in cancer research to identify prognostic biomarkers: Klauschen et al. [16] applied LRP for visual scoring of tumor-infiltrating lymphocytes (TIL) on hematoxylin and eosin breast cancer images. Binder et al. [17] used LRP to identify spatial regions (cancer cell, stroma, TILs) on morphological tumor images that explained predictions of molecular tumor properties (like protein expression).

There are also some interpretation methods specialized for Graph Neural Networks (GNN). In [18–20], the authors provided explanation methods that are exclusively based on and crafted only for Graph Convolutional Network [21] utilizing a convolutional architecture which is a simplified version of that of Graph-CNN [8] we use. Ying et al. [22] suggested the model-agnostic GNNExplainer that is suitable for node classification, link prediction, and graph classification, but the authors did not consider an application of their approach to graph signal classification [23, 24], which is the problem at hand. The GNN-LRP method [25] proposes explanations in the form of scored sequences of edges on the input graph (i.e., relevant walks). Such a sequence represents a path extracted from the input to the output of GNN that brings insights for GNN's decision strategy. This is useful especially for graph classification tasks, where each data point is represented as an individual graph. In our task, patients are represented as graph signals on a single graph, so that this method is not applicable.

Hence, there is still a lack of methods explaining individualized predictions in the context of graph signal classification task. Here, we adapted an existing LRP technique to graph convolutional layers of Graph-CNN [8] incorporating prior knowledge of a molecular network. Our approach generates explanations in the form of relevant subgraphs for each data point and allows to provide interpretable molecular subnetworks that are individual for each patient. According to the knowledge of the authors, an explanation method that benefits from prior knowledge and provides patient-specific subnetworks has not been shown before. The novelty of our work consists of two parts. First, we present the Graph Layer-wise Relevance Propagation (GLRP) method delivering data point-specific explanations for Graph-CNN [8]. Second, we train Graph-CNN on a large breast cancer dataset to predict an occurrence of distant metastasis and show how patient-specific molecular subnetworks assist in personalized precision medicine decisions: We interpret the classifier's predictions by patient-specific subnetworks that explain the differential clinical outcome and identify therapeutic vulnerabilities.

Methods

Gene expression data and molecular network

Protein-protein interaction network

We used the Human Protein Reference Database (HPRD) protein-protein interaction (PPI) network [26] as the molecular network to structure the gene expression data. The database contains protein-protein interaction information based on yeast two-hybrid analysis, in vitro and in vivo methods. The PPI network is an undirected graph with binary interactions between pairs of proteins. The graph is not connected.

Breast cancer data

We applied our methods to a large breast cancer patient dataset that we previously studied and preprocessed [27]. That data is compiled out of 10 public microarray datasets measured on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A arrays. The datasets are available from the Gene Expression Omnibus (GEO) [28] data repository (accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, GSE6532). The RMA probe-summary algorithm [29] was used to process each of the datasets, and only samples with metadata on metastasis-free survival were selected and combined together on the basis of HG-U133A array probe names. Quantile normalization was applied over all datasets. In the case of several probes mapping to one gene, only the probe with the highest average value was considered. After pre-processing the dataset contained 12,179 genes in 969 patients. The patients were assigned to one of two classes: 393 patients with distant metastasis within the first 5 years and 576 patients without metastasis having the last follow-up between 5 and 10 years. Breast cancer molecular subtypes for the patient samples were predicted in [27] utilizing *genefu* R-package [30].

After mapping of 12,179 genes to the vertices of the PPI, the resulting PPI graph consisted of 7168 vertices (mapped genes) in 207 connected components. The main connected component had 6888 vertices, and each of the other 206 components had from 1 to 4 vertices. For further analyses, we utilized only the main connected component since the Graph-CNN requires the graph to be connected. The preprocessed data is provided in [31].

Expression data of HUVECs before and after TNFα stimulation

For validation purposes, we analyzed gene expression data from human umbilical vein endothelial cells (HUVECs) treated or not treated with tumor necrosis factor alpha TNFα [32]. The data, provided by the same authors (GEO database series: GSE144803), containing 39 sample pairs (treated and untreated), were suitable for a binary classification task and balanced. The expression data were

quantile normalized and mapped to vertices of HPRD PPIs resulting in 7798 genes in the main connected component.

Problem formulation

We focus on explaining classifier decisions of Graph-CNN adapting existing LRP approaches for graph convolutional layers. LRP should be applied as a postprocessing step to a model already trained for the ML task. The task is formulated as a binary classification of gene expression data $X \in R^{n \times m}$ to a target variable $Y \in \{0,1\}^n$. n is the number of data points (patients) and m is the number of features (genes). The information of the molecular network is presented as an undirected weighted graph $G = (V, E, A)$, where V and E denote the sets of vertices and edges respectively and A denotes the adjacency matrix. The Graph-CNN was designed to work with weighted graphs. We define weighted adjacency matrix A of dimensionality $m \times m$ since in general molecular networks can be weighted. For the unweighted HPRD PPI network, the matrix A has only "0s" and "1s" as its elements. A row x of the gene expression matrix X contains data from one data point (patient) and can be mapped to the vertices of the graph G. In such a way, values of x are interpreted as a graph signal.

A trained neural network can be represented as a function $f : R_+^m \rightarrow [0,1]$ mapping the input to the probability of the output class. The input x is a set of gene expression values $x = \{x_g\}$ where g denotes a particular gene. The function $f(x)$ computes the probability that a certain pattern of gene expression values is present w.r.t to the output class. LRP methods apply propagation rules from the output of the neural network to the input in order to quantify the relevance score $R_g(x)$ for each gene g. These relevances show how much gene g influences the prediction $f(x)$:

$$\forall x : f(x) = \sum_g R_g(x). \tag{1}$$

Equation (1) [11] demonstrates that the relevance scores are calculated w.r.t every input data point x.

Graph Convolutional Neural Network and Layer-wise Relevance propagation

Usual CNNs learn data representations on grid-like structures. The Graph-CNN [8] as a deep learning technique is designed to learn features on weighted graphs. The convolution on graphs is used to capture localized patterns of a graph signal. This operation is based on spectral graph theory. The main operator to investigate the spectrum of a graph is the graph Laplacian $L = D - A$, where D is a weighted degree matrix, and A is a weighted adjacency matrix. L is a real symmetric positive semidefinite matrix that can be diagonalized such that $L = U \Lambda U^T$, where $\Lambda = diag([\lambda_1, \ldots, \lambda_m])$ is a diagonal non-negative real valued matrix of eigenvalues, matrix U is composed of eigenvectors. Matrices U and U^T define the Fourier and the inverse Fourier transform respectively. According to the convolution theorem, the operation of graph convolution can be viewed as a filtering operation:

$$y = h_\theta(L)x = h_\theta\left(U \Lambda U^T\right)x = U h_\theta(\Lambda) U^T x, \tag{2}$$

where $x, y \in R^m$, and the filter $h_\theta(\Lambda)$ is a function of eigenvalues (graph frequencies). To localize filters in space, the authors in [8] decided to use a polynomial parametrization

$$h_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k \Lambda^k, \tag{3}$$

where $\theta \in R^k$ is a vector of parameters. The order of the polynomial, which is equal to $K-1$, specifies the local $K-1$ hop neighborhood. The neighborhood is determined by the shortest path distance. The polynomial filter can be computed recursively, as a Chebyshev expansion, which is commonly used in graph signal processing to approximate kernels [33]. The Chebyshev polynomial $T_k(x)$ of order k is calculated as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0 = 1$ and $T_1 = x$. The Chebyshev expansion applies for values that lie in $[-1, 1]$; therefore, the diagonal matrix of eigenvalues Λ has to be derived from a rescaled Laplacian $L = (D - A)/\lambda_{max} - I_n$. Thus, the filtering operation can be rewritten as

$$y = h_\theta(\Lambda)x = \sum_{k=0}^{K-1} \theta_k T_k(L)x = [\bar{x}_0, \ldots, \bar{x}_{K-1}]\,\theta, \tag{4}$$

where $\bar{x}_k = 2L\bar{x}_{k-1} - \bar{x}_{k-2}$ with $\bar{x}_0 = x$ and $x_1 = Lx$. The transition in Eq. 4 is done according to the observation $\left(U \Lambda U^T\right)^k = U \Lambda^k U^T$. The filtering at the convolutional layer boils down to an efficient sequence of $K-1$ sparse matrix-vector multiplications and one dense matrix-vector multiplication [8].

LRP is based on the theoretical framework of deep Taylor decomposition. The function $f(x)$ from Eq. (1) can be decomposed in terms of the Taylor expansion at some chosen root point x^* so that $f(x^*) = 0$. The first order Taylor expansion of f(x) is:

$$\begin{aligned} f(x) &= f(x^*) + \sum_{g=1}^{m} \frac{\partial f}{\partial x}\Big|_{x=x^*} \cdot \left(x_g - x_g^*\right) + \epsilon \\ &= 0 + \sum_{g=1}^{m} R_g(x) + \epsilon \end{aligned} \tag{5}$$

where the relevances $R_g(x)$ are the partial differentials of the function $f(x)$. The details of how to choose a good root point are described in [11]. The $f(x)$ represents an output neuron of a neural network which consists of multiple layers and each layer consists of several neurons. A

neuron receives a weighted sum of its inputs and applies a nonlinear activation function. The idea of the deep Taylor decomposition is to perform a first order Taylor expansion at each neuron of the neural network. These expansions allow to produce relevance propagation rules that compute relevances at each layer in a backward pass. The rules redistribute the relevance from layer to layer starting from output until the input is reached. The value of the output represents the model's decision which is equal to the total relevance detected by the model.

LRP is commonly applied to deep neural networks consisting of layers with rectified linear units (ReLU) nonlinearities. In our experiments, we use only this activation function. Let i and j be single neurons at two consecutive layers at which the relevance should be propagated from j to i. The activation function has this form:

$$a_j = max\left(0, \sum_i a_i w_{ij} + b_j\right) \tag{6}$$

where a_i, a_j are neurons' values, w_{ij} are weights, and b_j is bias. Noticeably, the layers of this type always have non-negative activations. The relevance propagation rule is the following:

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+ + \epsilon} R_j, \tag{7}$$

where w_{ij}^+ corresponds to the positive weights w_{ij} and ϵ stabilizes numerical computations [9]. We set ϵ to 1^{-10}. Equation (7) depicts the z^+ rule coming from deep Taylor decomposition [11]. The z^+ rule is commonly applied to the convolutional and fully connected layers. It favors the effect of only positive contributions to the model decisions. The first input layer can have other propagation rules that are specific to the domain [34]. In our work, we used the rule (7) for the input layer as well since the gene expression data has positive values.

In order to propagate relevance through the filtering (4), we rewrite it as follows:

$$y = \sum_{k=0}^{K-1} \theta_k T_k(L)x = \left[\bar{L}_0, \ldots, \bar{L}_{K-1}\right]\theta x = Wx, \tag{8}$$

where matrix $W \in R^{m\times m}$ connects nodes y and x. The computation of matrix W is done as: $W = \left[\bar{L}_0, \ldots, \bar{L}_{K-1}\right]\theta$, where $\bar{L}_k = 2L\bar{L}_{k-1} - \bar{L}_{k-2}$ with $\bar{L}_0 = I$ and $\bar{L}_1 = L$ are the Chebyshev polynomials of the Laplacian matrix.

Each convolutional layer has F_{in} channels

$$\left[x_1, \ldots, x_{F_{in}}\right] \in R_+^{m\times F_{in}} \tag{9}$$

in the input feature map and F_{out} channels

$$\left[y_1, \ldots, y_{F_{out}}\right] \in R^{m\times F_{out}} \tag{10}$$

of the output feature map. We consider the values of output feature maps before applying ReLU non-linearities on them. The $F_{in} \times F_{out}$ vectors of the Chebyshev coefficients $\theta_{i,j} \in R^k$ are the layer's trainable parameters. The input feature map can be transformed into a vector $\hat{x} = \left[x_1^T, \ldots, x_{F_{in}}^T\right]^T \in R_+^{m\cdot F_{in}}$. We adapt Eq. (8) to compute the j^{th} channel of the output feature map based on the input feature map:

$$\begin{aligned} y_j &= \left[\bar{L}_0, \ldots, \bar{L}_{K-1}\right] \cdot \left[\theta_{1,j}, \ldots, \theta_{F_{in}j}\right] \cdot \left[x_1^T, \ldots, x_{F_{in}}^T\right]^T \\ &= \left[\hat{L}_{1,j}, \ldots, \hat{L}_{F_{in},j}\right] \cdot \left[x_1^T, \ldots, x_{F_{in}}^T\right]^T \\ &= \hat{W}_j \times \hat{x} \in R^m \end{aligned} \tag{11}$$

where $\hat{L}_{i,j} = \left[\bar{L}_0, \ldots, \bar{L}_{K-1}\right]\theta_{i,j} \in R^{m\times m}$, $\hat{W}_j = \left[\hat{L}_{1,j}, \ldots, \hat{L}_{F_{in},j}\right] \in R^{m\times m\cdot F_{in}}$

Since the j^{th} channel of the output feature map is connected through the matrix-vector multiplication with the input feature map, $\hat{W}_j$ can be treated as a matrix of weights joining two fully connected layers. Therefore, the relevance $R_y^j \in R_+^m$ from the j^{th} output channel can be propagated to the input feature map relevance $R_{\hat{x}}^j \in R_+^{m\cdot F_{in}}$ according to the rule (7). Overall, the relevance propagated from the output feature map to the input feature map is:

$$R_{\hat{x}} = \sum_{j=1}^{F_{out}} R_{\hat{x}}^j \in R_+^{m\cdot F_{in}}. \tag{12}$$

For running LRP on graph convolutional layers, one needs to compute huge and dense matrices $\hat{W}_j$. It requires $K-2$ sparse matrix-matrix multiplications and one sparse to dense matrix-matrix multiplication. The computations for relevance propagation are heavier and much more memory demanding compared to the filtering (4). The code implementing our GLRP approach is available in [35].

GLRP on gene expression data

To demonstrate the utility of GLRP, the Graph-CNNs were trained on two gene expression datasets described in the "Gene expression data and molecular network" section. In our previous study [23], the gene expression data were standardized for the training. But in this paper, we did not standardize the data. The argument for it is the following. For the non-image data, to standardize the input features is the usual practice. However, in case of standardization, the input features are treated independently. For an image, the neighboring pixels are highly correlated. If the pixels as features are standardized across the dataset, then this can distort the pattern of the image quite

significantly and lead to misinterpretation. Analogically, feature wise standardization of microarray data changes expression patterns of genes located in the same neighborhood of a molecular network (HPRD PPI in our case). This might affect the explainability of the Graph-CNN that we aim at. Therefore, we trained the Graph-CNN directly on the quantile normalized data avoiding the additional standardization step. Instead, we subtracted the minimal value (5.84847) of the data from each cell of the gene expression matrix to keep the gene expression values non-negative. If initially, GE data was lying in [5.84847, 14.2014], now it is in the interval [0.0, 8.3529]. This transformation allows Graph-CNN to converge faster, to apply the LRP propagation rule (7) suitable for non-negative input values, and to preserve original gene expression patterns in local neighborhoods of the PPI network.

For each of the two gene expression datasets structured by the same prior knowledge (HPRD PPI), we used a 10-fold cross validation over a whole dataset to estimate the predictive performance of Graph-CNN. The hyperparameters such as the number of filters, the presence of pooling, the learning rate, and decay were tweaked manually on this 10-fold cross validation.

The architecture of the Graph-CNN trained on the HUVECs dataset and its performance are given in the "Comparison of subnetworks derived by GLRP to gene-coexpression networks identified by WGCNA" section.

For the breast cancer dataset, the Graph-CNN architecture consisted of two graph convolutional layers following maximum pooling of size 2, and two hidden fully connected layers with 512 and 128 units respectively. Each graph convolutional layer contained 32 filters covering the vertex' neighborhood of size 7. For the performance comparison, we trained a "glmgraph" method [36] implementing network-constrained sparse regression model using HPRD PPI network, and Random Forest without any prior knowledge as baselines. The results on 10-fold cross validations are presented in the "GLRP to deliver patient-specific subnetworks" section.

Further we generated the patient-specific (data point specific) subnetworks via GLRP. For that, each of the gene expression datasets was randomly split again: 90% training and 10% test. We retrained the Graph-CNN on 90% of data using manually selected hyperparameters from 10-fold cross validation, and propagated relevances on test data which was not "seen" by the model during training to make it more challenging. Since the LRP rule (7) propagates only positive contributions, our Graph-CNN had two output neurons for binary classification tasks that showed the probability of these two classes. For each patient in the test set, relevance was propagated by GLRP from the predicted output neuron to the input neurons representing genes (vertices) of the underlying molecular network. The workflow to deliver the patient-specific subnetworks is depicted on Fig. 1. A patient-specific subnetwork explaining the prediction was constructed from the 140 most relevant genes. Selecting more than 140 top relevant vertices entailed visualization issues. The single-

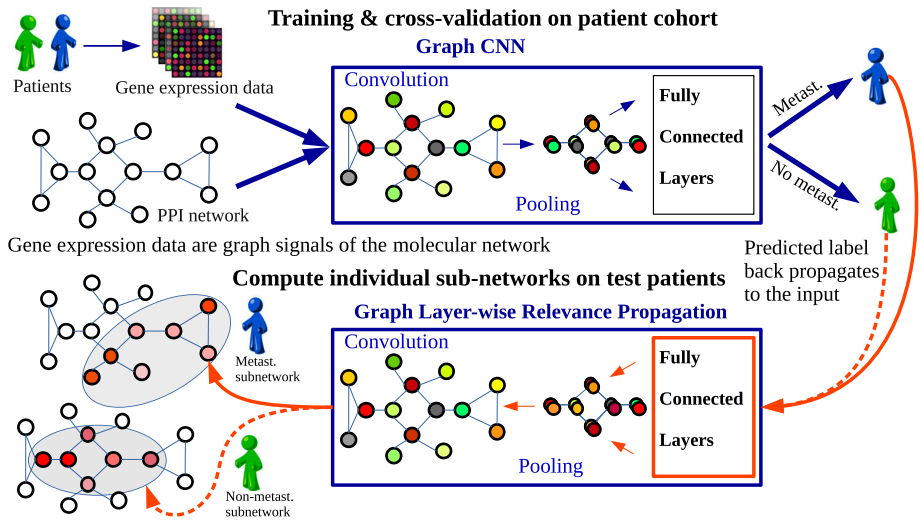

Fig. 1 The workflow to obtain a data point-specific subnetwork. For clarity, a data point represented by a gene expression profile of a patient from the breast cancer dataset. The molecular network (HPRD PPI) structures the genes and is the same for every patient. Patient's gene-expression values are assigned to every vertex of the HPRD PPI so that the patient is represented as a graph signal. Trained Graph-CNN performs graph convolutions and as output classifies the patient as metastatic or non-metastatic. GLRP is applied as a post hoc processing, propagating the relevance from the predicted label up to the input features (vertices of the molecular network). Top 140 highly relevant vertices constitute a molecular subnetwork. Molecular subnetworks differ from one patient to another

tons were deleted so that the subnetwork consisted mainly of around 130 vertices. The same workflow was applied to generate data-point-specific subnetworks for the data described in the "Expression data of HUVECs before and after TNFα stimulation" section.

Pathway analysis

Enrichment of signal transduction pathways annotated in the TRANSPATH® database version 2020.1 [37] in genes prioritized by GLRP were analyzed using the geneXplain platform version 6.1 [38]. The analysis based on the Fisher's exact test [39] was carried out for gene sets obtained for individual patients from the breast cancer dataset as well as for their combination into subtype gene sets.

The following calculations were applied to investigate differences in pathway hits. Let P denote a set of pathway genes and S_i and S_k two subnetwork gene sets, so that $P_i = P \cap S_i$ and $P_k = P \cap S_k$ are the sets of pathway genes matched by the two subnetworks. The difference $\triangle P_{i,k}$ in matched pathway genes was then calculated as $|(P_i \cup P_k) \setminus (P_i \cap P_k)|/|P_i \cup P_k|$ with $|P_i \cup P_k| > 0$. For each selected pathway, we calculated $\triangle P_i, k$ for each pair of subnetworks and reported the median of examined pairs.

Comparison of subnetworks derived by GLRP to gene-coexpression networks identified by WGCNA

To further examine the biological relevance of subnetwork genes prioritized by GLRP and for the purpose of comparison to an already available method that uses expression and network information to prioritize gene sets, we analyzed the gene expression data described in "Expression data of HUVECs before and after TNFα stimulation" section. We compared gene sets identified in our subnetworks to gene modules and differentially expressed genes in response to TNFα identified by Rhead et al. [32]. Rhead et al. [32] reported gene modules obtained by weighted gene co-expression network analysis (WGCNA). The method has been applied in many studies and constructs a gene network based on expression measurements from which it can derive modules of co-expressed genes [40]. We trained a Graph-CNN on the gene expression data to classify the TNFα treatment status of HUVECs. The Graph-CNN architecture consisted of 2 convolutional layers with 4 and 8 filters respectively followed by one hidden fully connected layer with 128 nodes. The vertex's neighborhood covered by graph convolutions was of size 7. No pooling was used. The performance of the Graph-CNN in 10-fold cross validation: mean 100*AUC, accuracy, and F1-weighted were 99.49, 96.25% and 96.06%, respectively. A random forest achieved the same performance. We generated the subnetworks according to the "GLRP on gene expression data" section, retrained the Graph-CNN on 70 randomly selected samples, and applied GLRP on 8 test samples (4 treated and 4 not treated). The test samples were predicted correctly. For each of the 8 test samples, we constructed a subnetwork. Associations between subnetwork genes sets and 16 gene modules defined by Rhead et al. [32] as well as 589 upregulated genes (log-fold change > 0.5, FDR < 0.01), 425 downregulated genes (log-fold change < -0.5, FDR < 0.01), and the combined set of 1014 DE genes were analyzed using the *Functional classification* tool of the geneXplain platform [41]. Fisher test calculations were carried out with a total contingency table count corresponding to the number of genes in [32, file S1 of] after mapping to Ensembl [42] gene ids (10022 genes). Rhead et al. [32] assigned a color code to the 16 gene co-expression modules and denoted them as *black*, *blue*, *brown*, *cyan*, *green*, *greenyellow*, *grey*, *magenta*, *midnightblue*, *pink*, *purple*, *red*, *salmon*, *tan*, *turquoise*, and *yellow* which is maintained in results reported here.

Results

Sanity check of the implemented graph LRP

To initially validate our implemented LRP, we applied Graph-CNN on the MNIST dataset [43] in the same way as described in the paper [8]. The MNIST dataset contains 70,000 images of hand-written digits each having a size of 28 by 28 pixels. To apply Graph-CNN on the image data, we constructed an 8 nearest-neighbors graph similarly to the schema proposed in [8], with the exception that all the weights are equal to 1. The weight 1 is more natural for the graph connecting neighboring image pixels. Thus, each image is a graph signal represented by node attributes—pixel values. We achieved high classification accuracy (99.02%) on the test set for the Graph-CNN, which is comparable to the performance of classical CNN (99.33%) reported in [8]. The number of parameters was the same for both methods.

Usually, to manage box-constrained pixel values, the special pixel-specific LRP rule is applied for the input layer. This pixel-specific rule highlights not only the digits itself, but also the contours of the digits [34, Fig. 13 of]. In contrast, the rule (7) highlights only those positively relevant parts of the image where the signal of the digit is present. We kept the propagation rule (7) for the input and all other layers in all our experiments. Further, we visually compared on the same digits how the heatmaps generated by implemented GLRP correspond to the heatmaps generated by usual LRP procedure applied on classical CNN (Fig. 2).

The heatmaps were rendered only for the classes predicted by classical CNN and Graph-CNN. In this case, the classes are "6" and "3". For the Graph-CNN, a bigger part of the digit is relevant for the classification since the covered neighborhood can be expanded up to 24 hops. Graph-CNN's filters are isotropic; thus, they tend to cover

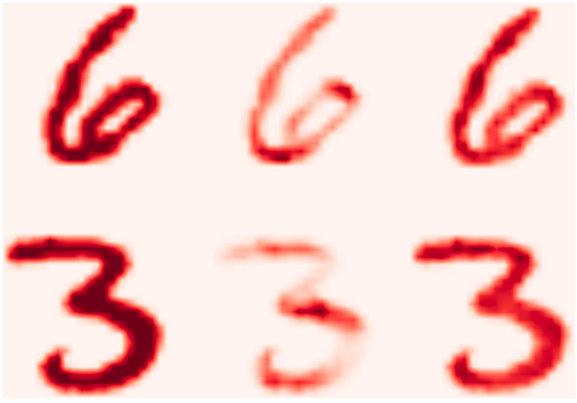

Fig. 2 From left to right: initial image, LRP on classical CNN and GLRP on Graph-CNN

roundish areas that concern rounded patterns (curves) of the digit (Additional file 1: Fig. S1).

Genes selected by GLRP correlate with modules identified by gene co-expression network analysis

In the analysis of TNF-induced gene expression changes in HUVECs, our procedure prioritized in total 168 genes of which 105 genes were found in subnetworks of all eight test samples (Additional file 2). Remarkably, the *green* gene module, which was the most strongly correlated one with TNFα upregulation [32], showed significant association (adjusted p value < 0.05) with the combined set of subnetwork genes, with genes found in the majority of subnetworks and also with 5 of the 8 subnetworks (Additional file 2). At the same significance level, the *turquoise* gene module described in [32] was strongly associated with 2 of 8 subnetworks and with genes found in all 8 subnetworks. In addition, both the *green* and the *turquoise* modules showed moderate association (adjusted p value < 0.1) with the majority of gene sets defined on the basis of the test subnetworks. Furthermore, we found strong (adjusted p value < 0.05) or moderately (adjusted p value < 0.1) significant overlap between upregulated genes and some subnetwork gene sets. The gene modules *cyan*, *greenyellow*, and *midnightblue* did not overlap with GLRP-derived subnetworks. These results demonstrate partial agreement between gene sets suggested by GLRP, another gene network analysis and classical differential expression analysis. Hence, the GLRP-based subnetworks gathered biologically meaningful genes and may even complement other approaches in revealing important properties of the underlying biological systems. Additionally, another two gene sets were compared with WGCNA modules: the intersection of subnetworks genes and genes that occurred in more than in 4 test samples subnetworks. Notably, the individual subnetworks shared more genes with the *green* and *turquoise* WGCNA modules than those described gene sets, pointing out the ability of GLRP to identify sample-specific genes.

GLRP to deliver patient-specific subnetworks

We applied the GLRP to the Graph-CNN trained on gene expression data from the "Breast cancer data" section. The gene expression data was structured by a protein-protein interaction network. The standardization of features was not performed as described in the "GLRP on gene expression data" section. The prediction task performed by the Graph-CNN was to classify patients into 2 groups, metastatic and non-metastatic. The results of a 10-fold cross validation are depicted in Table 1. While Graph-CNN and glmgraph utilized the HPRD PPI network topology, a random forest did not use any prior knowledge. glmgraph was not evaluated on non-standardized data, since it had convergence issues in this case. The metrics were averaged over folds and the standard errors of their means were calculated.

The GLRP was applied as described in the"GLRP on gene expression data" section. We retrained the Graph-CNN on 872 patients and generated relevances for 97 test patients. The relevances were propagated from the Graph-CNN's output node corresponding to the correctly predicted class. The most frequently selected features are summarized in Additional file 1: Table S1. The eukaryotic translation elongation factor EEF1A1, which is overexpressed in the majority of breast cancers and protects tumor cells from proteotoxic stress [44], was the sole factor that was selected in all of the 97 test set patients. Other frequently selected features in both non-metastatic as well as metastatic patients included genes such as the epithelial-to-mesenchymal-transition (EMT)-related gene VIM (46/58 non-metastatic, 30/39 metastatic patients), the extracellular matrix protein FN1 (43/58 non-metastatic, 22/39 metastatic patients), the actin cytoskeleton regulator CFL1 (7/58 non-metastatic, 7/39 metastatic

Table 1 Performance of Graph-CNN on metastatic event prediction, depending on normalization

Method	Std	100*AUC	Accuracy, %	F1-weighted, %
Graph-CNN	-	82.57±1.25	76.07±1.30	75.82±1.33
Random Forest	-	81.27±1.66	74.23±1.73	73.47±1.84
Graph-CNN	+	82.16±1.25	76.18±1.36	75.86±1.35
Random Forest	+	81.40±1.76	74.74±1.67	74.00±1.82
glmgraph	+	80.88±1.37	75.14±1.30	74.73±1.39

Std stands for standardization of features (genes)

patients), and the estrogen receptor ESR1 28/58 non-metastatic, 10/39 metastatic patients) that are all known to be linked with breast cancer development and progression [45–48]. This indicates that our method successfully identified relevant key players with a general role in breast tumorigenesis.

Additionally, we show individualized PPI subnetworks delivered for four correctly predicted breast cancer patients (Table 2) from the microarray data set. Two of them had been assigned with the most common subtype luminal A (LumA), while the other two suffered from the highly aggressive basal-like subtype. In each group, one patient with early metastasis was picked and one who did not develop any within at least 5 years of follow-up.

The generated PPI subnetworks are displayed in Fig. 3. The sequence of pictures in order ABCD is the same as in the table.

Interestingly, the networks of both LumA patients contained ESR1 which fits well since this subtype is considered as estrogen receptor positive [49]. In contrast, genes often associated with the basal-like subtype and a poor prognosis such as MCL1, CTNNB1, EGFR, or SOX4 were found in the basal-like patient GSM519217 suggesting that the generated networks are capable of extracting breast cancer subtype-specific features. The comparison of the subnetworks of the non-metastatic and the metastatic patients furthermore revealed some patient-specific genes which might give valuable information about specific mechanisms of tumorigenesis and therapeutic vulnerabilities in the respective patient. In general, it seemed that the subnetworks of the non-metastatic patients contained more genes that have been linked to better prognostic outcomes such as JUP, PCBP1, and HMGN2 in GSM615695 [50–52] or RASA1, IL6ST, KRT19, and RPS14 in GSM150990 [53–56], while the networks of both metastatic patients harbored genes that are known to be involved in aggressive tumor growth or therapy resistance which might explain the early metastatic spread in these patients. Some examples are CDK1, SFN, and XPO1 in GSM519217 [57–59] or CAV1, PTPN11, and FTL in GSM615233 [60–62].

Table 2 Patients that the PPI subnetworks are generated for

Patient's ID	Subtype	Metastatic event	Time of metastases, years	Last follow-up, years
GSM519217	Basal	1	0.9	-
GSM615233	LumA	1	0.79	-
GSM615695	Basal	0	-	5.38
GSM150990	LumA	0	-	9.93

However, not only the presence of specific genes might be important, but also their overall expression level. Our analyses identified, e.g., the EMT-related gene VIM as one of the most relevant nodes in the subnetworks of both metastatic patients in which the gene was highly expressed (> 75% quantile based on the gene expression throughout the whole patient cohort). In contrast, VIM was also present in the subnetworks of the two non-metastatic patients, however, with a lower relevance and a particularly low expression (< 25% quantile). VIM is an important marker for EMT and high expression levels correlate with a motile, mesenchymal-like cancer cell state, thus making VIM an essential effector of metastasis [45].

A comparison of subnetwork genes of 79 correctly predicted test set patients to a database of signal transduction pathways confirmed significant enrichment of pathways that have previously been associated with cancer disease mechanisms such as the EGF, ER-alpha, p53, and TGFbeta pathways as well as Caspase and beta-catenin networks. Comparisons were performed for each patient as well as for subtype gene sets formed by combining subnetwork genes of patients associated with a breast cancer subtype. Results for the 238 signaling pathways from the TRANSPATH® database that were significantly enriched with subtype genes are visualized in Fig. 4. Differences in enrichment significance may suggest that the importance of some signaling pathways detected this way is subtype-specific, e.g., for YAP ubiquitination or the VE-cadherin network (orange heatmap, Fig. 4, see also Additional file 1: Table S2 for details). The pattern of enrichment found on the level of cancer subtypes coincided well with the findings for subnetwork genes of individual patients revealing several molecular networks with elevated significance in both subtype and patient gene sets such as the EGF pathway, although the patient-level visualization did not suggest subtype-specific enrichment (green heatmap, Fig. 4). One source of these observations can be that patient subnetworks tend to be associated with certain pathways but cover different pathway components (genes). We therefore compared pathway genes in pairs of patient subnetworks for the 33 largest pathways. In 18 pathways, the median pair of patient subnetworks differed in 33% or more of the genes matched within a pathway (see also Additional file 1: Table S3 for details). These results demonstrate that the subnetworks obtained by

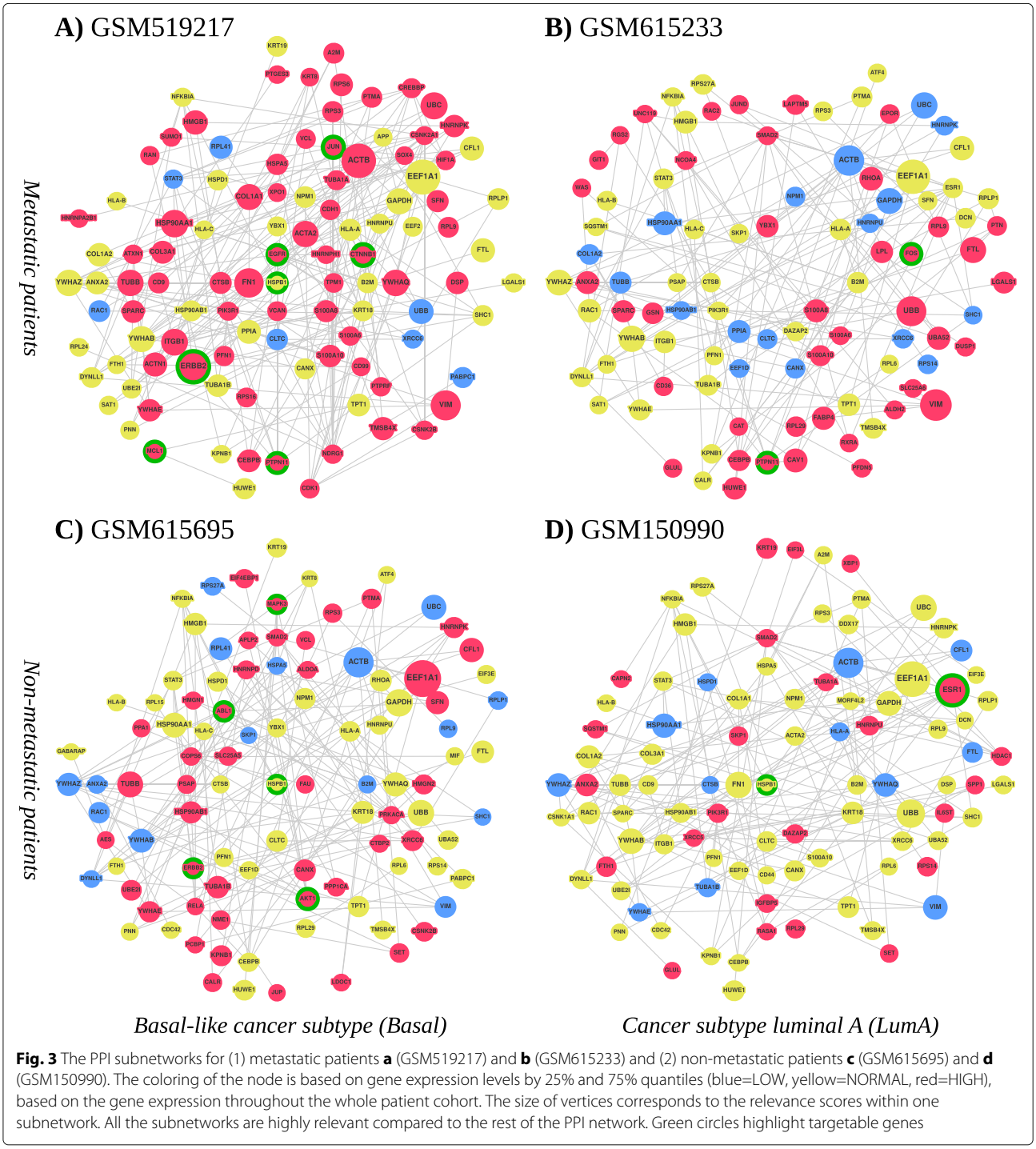

Fig. 3 The PPI subnetworks for (1) metastatic patients **a** (GSM519217) and **b** (GSM615233) and (2) non-metastatic patients **c** (GSM615695) and **d** (GSM150990). The coloring of the node is based on gene expression levels by 25% and 75% quantiles (blue=LOW, yellow=NORMAL, red=HIGH), based on the gene expression throughout the whole patient cohort. The size of vertices corresponds to the relevance scores within one subnetwork. All the subnetworks are highly relevant compared to the rest of the PPI network. Green circles highlight targetable genes

Graph-CNN were enriched with common signaling pathways relevant for the respective disease and can assign patient-specific priorities to pathway components.

Finally, we tested whether the subnetworks can also be used for finding potentially targetable genetic vulnerabilities that could open new options for personalized treatment decisions. We applied the "MTB report" methodology described in [63] to identify actionable genes present in the subnetworks. For that, we extended the algorithm to match high expression with gain of function alterations, and low expression with loss of function alterations. The results are summarized in Table 3.

Although information about the presence of actionable genetic variants is missing from our patient microarray

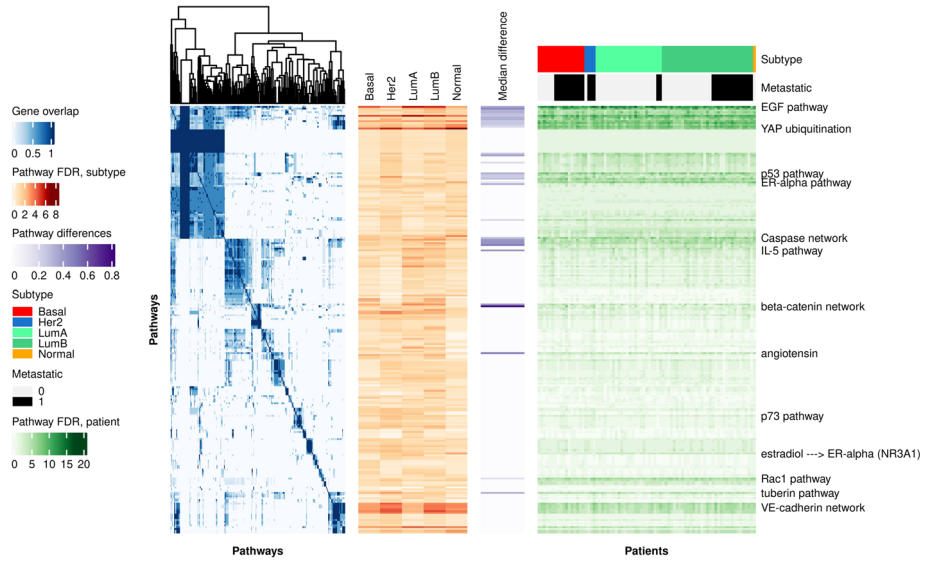

Fig. 4 Signal transduction pathway analysis of subnetwork genes reported for 79 patients in 5 subtypes. (From left to right) Blue heatmap: 238 signaling pathways clustered according to proportion of shared subnetwork genes; Orange heatmap: Enrichment significance of pathways in subnetwork genes combined from patients of given subtype. Darker orange indicates higher significance; Purple heatmap: Median difference in matched pathway genes observed in pairwise comparisons of subnetwork gene sets from patients mapped to 33 pathways. Darker purple indicates higher tendency of pairs of subnetwork gene sets to coincide with different pathway genes; Green heatmap: Enrichment significance of pathways in subnetwork genes of 79 patients. Darker green indicates higher significance. Corresponding subtypes and metastatic status are shown by the annotation above the heatmap. A detailed version of this figure capturing pathway and sample names is provided in Additional file 1: Fig. S2

data, the information generated by the PPI subnetworks could be used to define specific panels for subsequent sequencing. Indeed, the MTB reports highlighted specific genes that could be targeted therapeutically in each of the four patients: In the non-metastatic LumA patient GSM150990 ESR1 was proposed as therapeutic target which is in line with current treatment regimens that use hormone therapy as the main first-line treatment of choice for this patient subgroup. In contrast, in the metastatic LumA patient GSM615233 FOS and PTPN11 were identified as novel actionable alterations. In the often rapidly relapsing basal-like patients HSPB1 and ERBB2 were

Table 3 Actionable genes identified by the MTB report workflow

Patient	Gene	Expression	Known Var	Predicts
615695	HSPB1	Normal	expression	Response to gemcitabine
	ABL1	High	GoF	Response to ABL TK inhibitors (imatinib, desatininb, ponatinib, regorafenib. . .)
	AKT1	High	GoF	Response to PI3K, AKT, MTOR inhibitors; resistance to BRAF inhibitors
	ERBB2	High	GoF	Response to ERBB2, EGFR, MTOR, AKT inhibitors
	MAPK3	High	GoF	Resistance to EGFR inhibition
519217	HSPB1	Normal	expression	Response to gemcitabine
	CTNNB1	High	GoF	Response to everolimus + letrozole; resistance to Tankyrase inhibitors
	EGFR	High	GoF	Response to EGFR, ERBB2, HSP90 and MEK inhibitors
	ERBB2	High	GoF	Response to ERBB2, EGFR, MTOR, AKT inhibitors
	JUN	High	overexpr	Response to irbesartan (angiotensin II antagonist)
	MCL1	High	GoF	Resistance to anti-tubulin agents
	PTPN11	High	GoF	Response to MEK inhibitors
615233	FOS	High	overexpr	Response to irbesartan (angiotensin II antagonist)
	PTPN11	High	GoF	Response to MEK inhibitors
150990	HSPB1	Normal	expression	Response to gemcitabine
	ESR1	High	GoF	Response to novel ER degraders, fulvestrant, tamoxifen

Genes from the PPI subnetworks were matched to known genomic alterations (Known Var) that predict either response or resistance to drugs (Predicts). High and low gene expression were matched to gain of function (GoF) and loss of function (LoF) genomic variants, respectively

identified as common targets as well as MAPK3, AKT1, and ABL1 for the non-metastatic patient GSM615695 or EGFR, MCL1, CTNNB1, PTPN11, and JUN for the metastatic patient GSM519217, thereby suggesting novel possibilities for combinatory or alternative treatments. Taken together, GLRP provides subnetworks centered around known oncogenic drivers that seem reasonable in the context of cancer biology and can help to identify patient-specific cancer dependencies and therapeutic vulnerabilities in the context of precision oncology.

Discussion

In our work, we focused on the interpretability of a deep learning method utilizing molecular networks as prior knowledge. We implemented LRP for Graph-CNN and provided the sanity check of the developed approach on the MNIST dataset. Essentially, the main aim of the paper was to explain the prediction of metastasis for breast cancer patients by providing an individual molecular subnetwork specific for each patient. The patient-specific subnetworks provided interpretability of the deep learning method and demonstrated clinically relevant results on the breast cancer dataset.

Supposedly, the performance of Graph-CNN can be improved. The batch normalization technique [64] that is used to accelerate the training of deep neural networks is not seen to be available for the Graph-CNN, so this can be the way to enhance its performance. The LRP rule for batch normalization layers is yet another procedure to be adapted for Graph-CNN.

Another possibility to identify genes (and construct subnetworks out of them) influencing classifier decisions is to apply model-agnostic SHAP and LIME explanation methods. LIME method provides explanations of a data point based on feature perturbations. The method samples perturbations from a Gaussian distribution, ignoring correlations between features. It leads to the instability of explanations that is not favorable for personalized medicine. SHAP provides Shapley values for each feature of a data point as well but does not have such an issue, so we attempted to derive patients-specific subnetworks applying TreeExplainer and KernelExplainer from SHAP python module on Random Forest and Graph-CNN respectively. The subnetworks were build on the basis of HPRD PPI utilizing positive Shapley values, which were pushing prediction to a higher probability of corresponding class (metastatic or non-metastatic). The subnetworks obtained were mostly consisting from single vertices. In contrast, the subnetworks from GLRP and Graph-CNN were mostly connected. The SHAP's DeepExplainer approach suitable for convenient deep learning models is not applicable for Graph-CNN. The model-agnostic KernelExplainer computes SHAP values out of a debiased lasso regression. Reevaluating the model happens several thousands numbers of times specified by a user as well as a small background dataset is needed for integrating out features. Hence, the KernelExplainer is not scalable and application of it on Graph-CNN resulted in not connected subnetworks as well.

Furthermore, the sensitivity of Graph-CNN to the changes of prior knowledge is still to be investigated. Authors in [8] showed that for the MNIST images a random graph connecting pixels significantly decreases the performance destroying local connectivity. In our case, the permutation of the vertices of the PPI network does not influence the classifier performance on standardized gene expression data. Yet, PPI network is a small world network and its degree distribution fits to the power law with the exponent $\alpha = 2.70$. It implies great connectivity between proteins and means that any two nodes are separated by less than six hops. The filters of convolutional layers cover a 7-hop neighborhood of each vertex, so we assume it still might be enough to capture the gene expression patterns. In our future work, we will investigate how the properties of the prior knowledge influence the performance and explainability of Graph-CNN.

The subnetworks generated by GLRP contained common potential oncogenic drivers which indicates that they can extract the essential cancer pathways. Indeed, our analyses identified genes associated with hormone receptor-positive breast cancer (e.g. ESR1, IL6ST, CD36, GLUL, RASA1) in the networks from the patients with estrogen receptor positive, LumA breast cancer and genes associated with the basal-like subtype (e.g., EGFR, SOX4, AKT1 as well as high levels of HNRNPK) in the basal-like patients, underlining the biological relevance of the networks. Next to subtype-specific genes, the networks contained several oncogenes that were found in all four patients and could thus represent common drivers of breast cancer initiation and progression. One example is the actin-binding protein cofilin (CFL1) that regulates cancer cell motility and invasiveness [46]. Another interesting candidate is STAT3 which is activated in more than 40% of breast cancers and can cause deregulated cell proliferation and epithelial-to-mesenchymal transition (EMT) [65]. Our graphs not only displayed patient-specific PPI subnetworks, but also concisely visualized the relevance of each node and its expression levels. This information is potentially relevant to judge the biological significance of the gene in a patient-specific context.

Next to the common genes found in all four networks, each network was characterized by several special, cancer-associated genes which are of high interest because they might represent patient-specific central signaling nodes and therapeutic vulnerabilities. Some examples are PTPN11 that is known to activate a transcriptional program associated with cancer stem cells or the EMT-related genes SOX4 or VIM that might be responsible for the high

invasive capacity of the tumors and their early metastasis formation [45, 61, 66, 67]. Interestingly, the network of the metastatic patient GSM615233 harbored the genes FABP4 and LPL which both have been shown to interact with CD36, another highly expressed node in the network, to support cell proliferation and counteract apoptosis [68–70]. In contrast, in the non-metastatic patient GSM150990 especially the interleukin receptor IL6ST and the Ras GTPase-activating protein 1 (RASA1) seem to be interesting because for both high expression levels have been linked with a favorable prognosis [53, 54]. In the other non-metastatic patient GSM615695 high levels of HMGN2 and PCBP1 were identified which both have been shown to be able to inhibit cell proliferation [51, 52]. Although the experimental validation for the networks is still missing, it is tempting to speculate that these genes might contribute to the benign phenotype of the tumor in these patients.

All patient-specific subnetworks contained relevant drug targets that have been largely studied in breast cancer (e.g., ERBB2, ESR1, EGFR, AKT1). Yet, resistance mechanisms in breast cancer targeted therapies represent a big challenge; many of the identified therapeutic approaches have failed [71] due to the highly interconnected nature of signaling pathways and potential circumvents. A promising way forward could involve the molecular characterization of the tumor with transcriptomics and a parallel culture of patient-derived organoids. PPI networks could elucidate the right combination strategy by identifying central signaling nodes. Different therapeutic strategies could be tested on organoids and confirm the best strategy that synergistically blocks cancer cell escape routes and minimizes the emergence of survival mechanisms. Only the identification of relevant mechanisms of action for cell survival as well as of the factors involved in resistance for each patient, together with a more precise and personalized characterization of each cancer phenotype, may provide useful improvements in current therapeutic approaches.

Conclusions

We present a novel Graph-CNN-based feature selection method that benefits from prior knowledge and provides patient-specific subnetworks. We adapted the existing Layer-wise Relevance Propagation technique to the Graph-CNN, demonstrated it on MNIST data, and showed its applicability on a large breast cancer dataset. Our new approach generated individual patient-specific molecular subnetworks that influenced the model's decision in the given context of a classification problem. The subnetworks selected by the developed method utilizing general prior knowledge are relevant for prediction of metastasis in breast cancer. They contain common as well as subtype-specific cancer genes that match the clinical subtype of the patients, together with patient-specific genes that could potentially be linked to aggressive/benign phenotypes. In the context of a breast cancer dataset GLRP provides patient-specific explanations for the Graph-CNN that largely agree with clinical knowledge, include oncogenic drivers of tumor progression, and can help to identify therapeutic vulnerabilities. We therefore conclude that our method GLRP in combination with Graph-CNN is a new, useful, and interpretable ML approach for high-dimensional genomic data-sets. Generated classifiers rely on prior knowledge of molecular networks and can be interpreted by patient-specific subnetworks driving the individual classification result. These subnetworks can be visualized and interpreted in a biomedical context on the individual patient level. This approach could thus be useful for precision medicine approaches such as for example the molecular tumorboard.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00845-7.

Additional file 1: Contains Supplementary Tables S1-S3 and Supplementary Figures S1, S2.

Additional file 2: Subnetwork genes obtained for 8 test samples and analysis of their association with gene modules reported by [32] as well as differentially expressed (DE) genes.
Worksheet *Subnetwork genes 8 samples* provides identifiers and gene symbols of 167 subnetwork genes, in how many and in which samples they were selected. Worksheet *Gene module enrichment* presents results of Fisher test calculations comparing subnetwork gene sets to gene modules and DE gene sets. Each row contains data for a DE gene set or a gene module consisting of the total group size and column tripletts with *p*-value, adjusted *p*-value as well as the number of hits, respectively, observed in comparisons to the union of genes from 8 subnetworks, the set of genes occurring in the majority, the set of genes found in all of the subnetworks and each of the 8 samples. Highlighted are rows corresponding to *green* and *turquoise* gene modules, which were most often significantly associated with subnetwork gene sets (grey), adjusted *p*-values below 0.05 (red) and between 0.05 and 0.1 (yellow).

Abbreviations

ML: Machine learning; LRP: Layer-wise Relevance Propagation; GNN: Graph Neural Network; CNN: Convolutional Neural Network; GLRP: Graph Layer-wise Relevance Propagation; WGCNA: Weighted gene co-expression network analysis; GDPR: General Data Protection Regulation; GEO: Gene Expression Omnibus; HPRD: Human Protein Reference Database; PPI: Protein-protein interaction; ReLU: Rectified linear unit; HUVEC: Human umbilical vein endothelial cells; EMT: Epithelial-to-mesenchymal transition

Acknowledgements

We would like to acknowledge Michaela Bayerlová, Mark Gluzman, and Vladyslav Yushchenko for fruitful discussions. HC is a member of the International Max Planck Research School for Genome Science, part of the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciences. TB is a member of the Göttingen Campus Institute Data Science.

Authors' contributions

HC and TB designed the study. HC developed and implemented the approach and performed the computational experiments. AB, FK, and PS provided major contributions to the study design. AB and KM provided clinical insights as well as JPB and PS provided biological insights, performing analyses of patient-specific subnetworks. FA developed the web-site to visualize the subnetworks.

TB and AL provided machine learning insights. HC, TB, AL, KM, and PS wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by the German Ministry of Education and Research (BMBF) e:Med project *MyPathSem* (031L0024) and the project *MTB-Report* by the big data initiative of the Volkswagenstiftung. KM was supported by German Research Foundation (DFG) project 424252458. We acknowledge support by the Open Access Publication Funds of the Göttingen University. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
The utilized breast cancer datasets are accessible from Gene Expression Omnibus (GEO) [28] data repository (accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, GSE6532). The HUVECs gene expression data [32] is available in GEO database (GSE144803). The HPRD PPI network can be found in [26]. The preprocessed breast cancer data, the adjacency matrix of the HPRD PPI network, and the code of the GLRP method are provided in http://mypathsem.bioinf.med.uni-goettingen.de/resources/glrp [31] and https://gitlab.gwdg.de/UKEBpublic/graph-lrp [35]. The web-site to explore patient-specific subnetworks is in http://mypathsem.bioinf.med.unigoettingen.de/MetaRelSubNetVis/ [72].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
PS is an employee of geneXplain GmbH, Germany. The remaining authors declare that they have no competing interests.

Author details
[1]Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany. [2]Dept. of Medicine A (Hematology, Oncology, Hemostaseology and Pulmonology), University Hospital Münster, Münster, Germany. [3]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. [4]geneXplain GmbH, Wolfenbüttel, Germany. [5]IT Infrastructure for Translational Medical Research, University of Augsburg, Augsburg, Germany. [6]Medical Statistics, University Medical Center Göttingen, Göttingen, Germany. [7]Campus-Institute Data Science (CIDAS), University of Göttingen, Göttingen, Germany.

Received: 26 August 2020 Accepted: 5 February 2021
Published online: 11 March 2021

References

1. Perera-Bel J, Leha A, Beißbarth T. In: Badve S, Kumar GL, editors. Bioinformatic methods and resources for biomarker discovery, validation, development, and integration. Cham: Springer; 2019, pp. 149–64. https://doi.org/10.1007/978-3-319-95228-4_11.
2. Rivenbark AG, O'Connor SM, Coleman WB. Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine. Am J Pathol. 2013;183(4):1113–24. https://doi.org/10.1016/j.ajpath.2013.08.002.
3. Sørlie T. Molecular classification of breast tumors: toward improved diagnostics and treatments. In: Target Discovery and Validation Reviews and Protocols. Totowa: Humana Press; 2007. p. 91–114. https://doi.org/10.1385/1-59745-165-7:91.
4. Fragomeni SM, Sciallis A, Jeruss JS. Molecular subtypes and local-regional control of breast cancer. Surg Oncol Clin N Am. 2018;27(1):95–120. https://doi.org/10.1016/j.soc.2017.08.005.
5. Porzelius C, Johannes M, Binder H, Beißbarth T. Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients. Biom J. 2011;53(2):190–201. https://doi.org/10.1002/bimj.201000155, Accessed 01 Dec 2020.
6. Johannes M, Brase JC, Fröhlich H, Gade S, Gehrmann M, Fälth M, Sültmann H, Beissbarth T. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics. 2010;26(17):2136–44. https://doi.org/10.1093/bioinformatics/btq345.
7. Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 5115–24.
8. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS); 2016. p. 3844–52.
9. Yang Y, Tresp V, Wunderle M, Fasching PA. Explaining therapy predictions with layer-wise relevance propagation in neural networks. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018. p. 152–62. https://doi.org/10.1109/ICHI.2018.00025.
10. Parliament and C. of the European Union. General data protection regulation. 2016. https://gdpr-info.eu/.
11. Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. 2017;65:211–22. https://doi.org/10.1016/j.patcog.2016.11.008.
12. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2016. p. 1135–44.
13. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS); 2017. p. 4768–77.
14. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV. Cham: Springer; 2014. p. 818–33. https://doi.org/10.1007/978-3-319-10590-1_53.
15. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. 2015;10(7):0130140. https://doi.org/10.1371/journal.pone.0130140.
16. Klauschen F, Müller K-R, Binder A, Bockmayr M, Hägele M, Seegerer P, Wienert S, Pruneri G, de Maria S, Badve S, Michiels S, Nielsen TO, Adams S, Savas P, Symmans F, Willis S, Gruosso T, Park M, Haibe-Kains B, Gallas B, Thompson AM, Cree I, Sotiriou C, Solinas C, Preusser M, Hewitt SM, Rimm D, Viale G, Loi S, Loibl S, Salgado R, Denkert C. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. Semin Cancer Biol. 2018;52:151–7. https://doi.org/10.1016/j.semcancer.2018.07.001. Immuno-oncological biomarkers.
17. Binder A, Bockmayr M, Hägele M, Wienert S, Heim D, Hellweg K, Stenzinger A, Parlow L, Budczies J, Goeppert B, Treue D, Kotani M, Ishii M, Dietel M, Hocke A, Denkert C, Müller K-R, Klauschen F. Towards computational fluorescence microscopy: machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv:1805.11178 [cs]. 2018.
18. Xie S, Lu M. Interpreting and understanding graph convolutional neural network using gradient-based attribution method. arXiv:1903.03768 [cs]. 2019. Accessed 12 July 2020.
19. Schwarzenberg R, Hübner M, Harbecke D, Alt C, Hennig L. Layerwise relevance visualization in convolutional text graph classifiers. arXiv:1909.10911 [cs]. 2019. Accessed 06 Nov 2020.
20. Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. Explainability methods for graph convolutional neural networks. In: 2019 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 10764–73. https://doi.org/10.1109/CVPR.2019.01103. ISSN: 2575-7075.
21. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 [cs, stat]. 2016. Accessed 09-01-2017.
22. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. Adv Neural Inf Process Syst. 2019;32:9240–51.
23. Chereda H, Bleckmann A, Kramer F, Leha A, Beissbarth T. Utilizing molecular network information via graph convolutional neural networks to predict metastatic event in breast cancer. Stud Health Technol Inform. 2019;267:181–6. https://doi.org/10.3233/SHTI190824.
24. Rhee S, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization; 2018. p. 3527–34. https://doi.org/10.24963/ijcai.2018/490. https://www.ijcai.org/proceedings/2018/490.

25. Schnake T, Eberle O, Lederer J, Nakajima S, Schütt KT, Müller K-R, Montavon G. XAI for graphs: explaining graph neural network predictions by identifying relevant walks. arXiv:2006.03589 [cs, stat]. 2020. Accessed 29 Oct 2020.
26. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database?2009 update. Nucleic Acids Res. 2009;37:767–72. https://doi.org/10.1093/nar/gkn892.
27. Bayerlová M, Menck K, Klemm F, Wolff A, Pukrop T, Binder C, Beißbarth T, Bleckmann A. Ror2 signaling and its relevance in breast cancer progression. Front Oncol. 2017;7:135. https://doi.org/10.3389/fonc.2017.00135.
28. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41(Database issue):991–5. https://doi.org/10.1093/nar/gks1193.
29. Irizarry RA, Hobbs B, Collin F, Beazer–Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64. https://doi.org/10.1093/biostatistics/4.2.249.
30. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, Haibe-Kains B. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32(7):1097–9. https://doi.org/10.1093/bioinformatics/btv693.
31. Bayerlová M, Chereda H. Preprocessed breast cancer data. 2020. http://mypathsem.bioinf.med.uni-goettingen.de/resources/glrp.
32. Rhead B, Shao X, Quach H, Ghai P, Barcellos LF, Bowcock AM. Global expression and CpG methylation analysis of primary endothelial cells before and after TNFa stimulation reveals gene modules enriched in inflammatory and infectious diseases and associated DMRs. PLoS ONE. 2020;15(3):0230884. https://doi.org/10.1371/journal.pone.0230884.
33. Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. Appl Comput Harmon Anal. 2011;30(2):129–50. https://doi.org/10.1016/j.acha.2010.04.005.
34. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
35. Chereda H. Graph layer-wise relevance propagation (GLRP). Gitlab. 2020. https://gitlab.gwdg.de/UKEBpublic/graph-lrp.
36. Chen L, Liu H, Kocher J-PA, Li H, Chen J. glmgraph: an R package for variable selection and predictive modeling of structured genomic data. Bioinformatics. 2015;31(24):3991–3. https://doi.org/10.1093/bioinformatics/btv497.
37. Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E. TRANSPATH ® : an integrated database on signal transduction and a tool for array analysis. Nucleic Acids Res. 2003;31(1):97–100. http://dx.doi.org/10.1093/nar/gkg089. https://academic.oup.com/nar/article-pdf/31/1/97/7127458/gkg089.pdf.
38. Koschmann J, Bhar A, Stegmaier P, Kel A, Wingender E. "Upstream analysis": an integrated promoter-pathway analysis approach to causal interpretation of microarray data. Microarrays. 2015;4(2):270–86. https://doi.org/10.3390/microarrays4020270.
39. Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Stat Soc. 1922;85(1):87–94. https://doi.org/10.2307/2340521.
40. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:17. https://doi.org/10.2202/1544-6115.1128.
41. Kolpakov F, Poroikov V, Selivanova G, Kel A. GeneXplain–identification of causal biomarkers and drug targets in personalized cancer pathways. J Biomol Tech. 2011;22(Suppl):16.
42. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. Ensembl 2020. Nucleic Acids Res. 2020;48(D1):682–8. https://doi.org/10.1093/nar/gkz966.
43. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324. https://doi.org/10.1109/5.726791.
44. Lin C-Y, Beattie A, Baradaran B, Dray E, Duijf PHG. Contradictory mRNA and protein misexpression of EEF1A1 in ductal breast carcinoma due to cell cycle regulation and cellular stress. Sci Rep. 2018;8(1):13904. https://doi.org/10.1038/s41598-018-32272-x.
45. Sharma P, Alsharif S, Fallatah A, Chung BM. Intermediate filaments as effectors of cancer development and metastasis: a focus on keratins, vimentin, and nestin. Cells. 2019;8(5):497. https://doi.org/10.3390/cells8050497.
46. Wang W, Eddy R, Condeelis J. The cofilin pathway in breast cancer invasion and metastasis. Nat Rev Cancer. 2007;7(6):429–40. https://doi.org/10.1038/nrc2148.
47. Lin T-C, Yang C-H, Cheng L-H, Chang W-T, Lin Y-R, Cheng H-C. Fibronectin in cancer: Friend or foe. Cells. 2019;9(1):27. https://doi.org/10.3390/cells9010027.
48. Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, Ji X, Liu W, Huang B, Luo W, Liu B, Lei Y, Du S, Vuppalapati A, Luu HH, Haydon RC, He T-C, Ren G. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes Dis. 2018;5(2):77–106. https://doi.org/10.1016/j.gendis.2018.05.001.
49. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø., Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale A-L, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52. https://doi.org/10.1038/35021093.
50. Bailey CK, Mittal MK, Misra S, Chaudhuri G. High motility of triple-negative breast cancer cells is due to repression of plakoglobin gene by metastasis modulator protein SLUG. J Biol Chem. 2012;287(23):19472–86. https://doi.org/10.1074/jbc.m112.345728.
51. Shi H, Li H, Yuan R, Guan W, Zhang X, Zhang S, Zhang W, Tong F, Li L, Song Z, Wang C, Yang S, Wang H. PCBP1 depletion promotes tumorigenesis through attenuation of p27 Kip1 mRNA stability and translation. J Exp Clin Cancer Res. 2018;37(1):187. https://doi.org/10.1186/s13046-018-0840-1.
52. Fan B, Shi S, Shen X, Yang X, Liu N, Wu G, Guo X, Huang N. Effect of HMGN2 on proliferation and apoptosis of MCF-7 breast cancer cells. Oncol Lett. 2018;17(1):1160–6. https://doi.org/10.3892/ol.2018.9668.
53. Liu Y, Liu T, Sun Q, Niu M, Jiang Y, Pang D. Downregulation of Ras GTPase-activating protein 1 is associated with poor survival of breast invasive ductal carcinoma patients. Oncol Rep. 2014;33(1):119–24. https://doi.org/10.3892/or.2014.3604.
54. Mathe A, Wong-Brown M, Morten B, Forbes JF, Braye SG, Avery-Kiejda KA, Scott RJ. Novel genes associated with lymph node metastasis in triple negative breast cancer. Sci Rep. 2015;5(1):15832. https://doi.org/10.1038/srep15832.
55. Saha S, Kim K, Yang G-M, Choi H, Cho S-G. Cytokeratin 19 (KRT19) has a role in the reprogramming of cancer stem cell-like cells to less aggressive and more drug-sensitive cells. Int J Mol Sci. 2018;19(5):1423. https://doi.org/10.3390/ijms19051423.
56. Zhou X, Hao Q, Liao J-M, Liao P, Lu H. Ribosomal protein S14 negatively regulates c-Myc activity. J Biol Chem. 2013;288(30):21793–801. https://doi.org/10.1074/jbc.m112.445122.
57. Alexandrou S, George S, Ormandy C, Lim E, Oakes S, Caldon C. The proliferative and apoptotic landscape of basal-like breast cancer. Int J Mol Sci. 2019;20(3):667. https://doi.org/10.3390/ijms20030667.
58. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe J-P, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo W-L, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006;10(6):515–27. https://doi.org/10.1016/j.ccr.2006.10.008.

59. Taylor J, Sendino M, Gorelick AN, Pastore A, Chang MT, Penson AV, Gavrila EI, Stewart C, Melnik EM, Chavez FH, Bitner L, Yoshimi A, Lee SC-W, Inoue D, Liu B, Zhang XJ, Mato AR, Dogan A, Kharas MG, Chen Y, Wang D, Soni RK, Hendrickson RC, Prieto G, Rodriguez JA, Taylor BS, Abdel-Wahab O. Altered nuclear export signal recognition as a driver of oncogenesis. Cancer Discov. 2019;9(10):1452–67. https://doi.org/10.1158/2159-8290.cd-19-0298.
60. Qian X-L, Pan Y-H, Huang Q-Y, Shi Y-B, Huang Q-Y, Hu Z-Z, Xiong L-X. Caveolin-1: a multifaceted driver of breast cancer progression and its application in clinical treatment. OncoTargets Ther. 2019;12:1539–52. https://doi.org/10.2147/ott.s191317.
61. Aceto N, Sausgruber N, Brinkhaus H, Gaidatzis D, Martiny-Baron G, Mazzarol G, Confalonieri S, Quarto M, Hu G, Balwierz PJ, Pachkov M, Elledge SJ, van Nimwegen E, Stadler MB, Bentires-Alj M. Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop. Nat Med. 2012;18(4):529–37. https://doi.org/10.1038/nm.2645.
62. Chekhun VF, Lukyanova NY, Burlaka AP, Bezdenezhnykh NA, Shpyleva SI, Tryndyak VP, Beland FA, Pogribny IP. Iron metabolism disturbances in the MCF-7 human breast cancer cells with acquired resistance to doxorubicin and cisplatin. Int J Oncol. 2013;43(5):1481–6. https://doi.org/10.3892/ijo.2013.2063.
63. Perera-Bel J, Hutter B, Heining C, Bleckmann A, Fröhlich M, Fröhling S, Glimm H, Brors B, Beißbarth T. From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. Genome Med. 2018;10(1):18. https://doi.org/10.1186/s13073-018-0529-2.
64. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, PMLR. 2015;37:448–456. http://proceedings.mlr.press/v37/ioffe15.html.
65. Banerjee K, Resat H. Constitutive activation of STAT 3 in breast cancer cells: a review. Int J Cancer. 2015;138(11):2570–8. https://doi.org/10.1002/ijc.29923.
66. Bentires-Alj M, Paez JG, David FS, Keilhack H, Halmos B, Naoki K, Maris JM, Richardson A, Bardelli A, Sugarbaker DJ, Richards WG, Du J, Girard L, Minna JD, Loh ML, Fisher DE, Velculescu VE, Vogelstein B, Meyerson M, Sellers WR, Neel BG. Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. Cancer Res. 2004;64(24):8816–20. https://doi.org/10.1158/0008-5472.can-04-1923.
67. Zhang J, Liang Q, Lei Y, Yao M, Li L, Gao X, Feng J, Zhang Y, Gao H, Liu D-X, Lu J, Huang B. SOX4 induces epithelial?mesenchymal transition and contributes to breast cancer progression. Cancer Res. 2012;72(17):4597–608. https://doi.org/10.1158/0008-5472.can-12-1045.
68. Guaita-Esteruelas S, Bosquet A, Saavedra P, Gumà J, Girona J, Lam EW-F, Amillano K, Borràs J, Masana L. Exogenous FABP4 increases breast cancer cell proliferation and activates the expression of fatty acid transport proteins. Mol Carcinog. 2016;56(1):208–17. https://doi.org/10.1002/mc.22485.
69. Liang Y, Han H, Liu L, Duan Y, Yang X, Ma C, Zhu Y, Han J, Li X, Chen Y. CD36 plays a critical role in proliferation, migration and tamoxifen-inhibited growth of ER-positive breast cancer cells. Oncogenesis. 2018;7(12):98. https://doi.org/10.1038/s41389-018-0107-x.
70. Kuemmerle NB, Rysman E, Lombardo PS, Flanagan AJ, Lipe BC, Wells WA, Pettus JR, Froehlich HM, Memoli VA, Morganelli PM, Swinnen JV, Timmerman LA, Chaychi L, Fricano CJ, Eisenberg BL, Coleman WB, Kinlaw WB. Lipoprotein lipase links dietary fat to solid tumor cell proliferation. Mol Cancer Ther. 2011;10(3):427–36. https://doi.org/10.1158/1535-7163.mct-10-0802.
71. Nakai K, Hung MC, Yamaguchi H. A perspective on anti-EGFR therapies targeting triple-negative breast cancer. Am J Cancer Res. 2016;6(8):1609–23.
72. Auer F. Patient specific molecular sub-networks responsible for metastasis in breast cancer. 2020. http://mypathsem.bioinf.med.uni-goettingen.de/MetaRelSubNetVis.

Publisher's Note

4 Stability of feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation

Reference

Chereda H, Leha A, Beissbarth T. Stability of feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation. bioRxiv. 2021 Dec 27. doi: 10.1101/2021.12.26.474194

Original Contribution

HC and TB conceived and designed the study. HC developed and implemented the software necessary to run the computational experiments. AL provided machine learning insights. HC wrote the paper, TB and AL contributed to editing the manuscript.

Stability of feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation

Hryhorii Chereda [1], Andreas Leha [2], Tim Beißbarth [1,3,*]

[1]Medical Bioinformatics, University Medical Center Göttingen, Germany
[2]Medical Statistics, University Medical Center Göttingen, Germany
[3]Campus-Institute Data Science (CIDAS), University of Göttingen, Göttingen, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: High-throughput technologies play a more and more significant role in discovering prognostic molecular signatures and identifying novel drug targets. It is common to apply Machine Learning (ML) methods to classify high-dimensional gene expression data and to determine a subset of features (genes) that is important for decisions of a ML model. One feature subset of important genes corresponds to one dataset and it is essential to sustain the stability of feature sets across different datasets with the same clinical endpoint since the selected genes are candidates for prognostic biomarkers. The stability of feature selection can be improved by including information of molecular networks into ML methods. Gene expression data can be assigned to the vertices of a molecular network's graph and then classified by a Graph Convolutional Neural Network (GCNN). GCNN is a contemporary deep learning approach that can be applied to graph-structured data. Layer-wise Relevance Propagation (LRP) is a technique to explain decisions of deep learning methods. In our recent work we developed Graph Layer-wise Relevance Propagation (GLRP) — a method that adapts LRP to a graph convolution and explains patient-specific decisions of GCNN. GLRP delivers individual molecular signatures as patient-specific subnetworks that are parts of a molecular network representing background knowledge about biological mechanisms. GLRP gives a possibility to deliver the subset of features corresponding to a dataset as well, so that the stability of feature selection performed by GLRP can be measured and compared to that of other methods.
Results: Utilizing two large breast cancer datasets, we analysed properties of feature sets selected by GLRP (GCNN+LRP) such as stability and permutation importance. We have implemented a graph convolutional layer of GCNN as a Keras layer so that the SHAP (SHapley Additive exPlanation) explanation method could be also applied to a Keras version of a GCNN model. We compare the stability of feature selection performed by GCNN+LRP to the stability of GCNN+SHAP and to other ML based feature selection methods. We conclude, that GCNN+LRP shows the highest stability among other feature selection methods including GCNN+SHAP. It was established that the permutation importance of features among GLRP subnetworks is lower than among GCNN+SHAP subnetworks, but in the context of the utilized molecular network, a GLRP subnetwork of an individual patient is on average substantially more connected (and interpretable) than a GCNN+SHAP subnetwork, which consists mainly of single vertices.
Keywords: gene expression data, explainable AI, personalized medicine, precision medicine, classification of cancer, deep learning, prior knowledge, molecular networks.
Availability: https://gitlab.gwdg.de/UKEBpublic/graph-lrp
Contact: tim.beissbarth@bioinf.med.uni-goettingen.de

1 Introduction

Microarray and especially high-throughput technologies have become commonly used tools for genome-wide gene-expression profiling. Gene expression patterns elucidate the molecular mechanisms of such heterogeneous disease as breast cancer (Sørlie, 2007) As a result, large amounts of data produced by high-throughput sequencing are utilized to identify predictive gene signatures and discover individual biomarkers in cancer prognosis (Perera, Leha, and Beissbarth, 2019)

One of the tasks of clinical cancer research is to identify prognostic gene signatures that are able to predict the clinical outcome (Johannes et al., 2010) From a machine learning perspective, the clinical endpoint is usually presented as a classification task, and the challenge is to find a subset of important features containing the most information about the clinical outcome. Prediction is performed by a ML model, which is trained on a

high-dimensional gene expression dataset. A predictive gene signature is a feature subset driving the classification result of the ML model. However, when the number of genes is much higher than the number of patients, the feature selection for the ML model has to deal with the "curse of dimensionality" (Porzelius et al., 2011) It leads to instability in the selected feature subsets across different datasets with the same clinical endpoint.

The stability of a feature selection algorithm is essentially the robustness of the algorithm's feature preferences. The feature selection is unstable when small changes in training data lead to large changes in the chosen feature subsets. The quantification of stability can be performed by providing different samples from the same training data and measuring the changes among chosen feature subsets. According to (Nogueira, Sechidis, and Brown, 2018) the measurement of stability addresses the question — *how much we can trust the algorithm?* From biomedical standpoint, it is crucial to guarantee the reproducibility of the given feature selection methods when finding proper sets of biomarkers (Lee et al., 2013)

Incorporation of prior knowledge of molecular networks (e.g. pathways) into a ML algorithm improves stability (Johannes et al., 2010) since genes connected in close proximity should have similar expression profiles and should not be treated independently. Molecular networks represent molecular processes in a given biological system and are widely used by biologists to interpret the results of a statistical analysis (Porzelius et al., 2011) The nodes of a molecular network depict molecules: genes, RNA, proteins and metabolites. The interactions between molecules are represented by edges. Different molecular networks can be used to approximate the interactions between features (genes). ML-based feature selection methods benefit from molecular network information in terms of interpretability of selected gene signatures (Johannes et al., 2010; Porzelius et al., 2011)

In our recent work (Chereda et al., 2021) we presented the Graph Layer-wise Relevance Propagation (GLRP), an adaptation to GCNN (Defferrard, Bresson, and Vandergheynst, 2016) of the Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) method explaining deep neural networks. The GCNN method utilizes prior knowledge of a molecular network structuring gene expression data. The GLRP approach delivers patient-specific predictive subnetworks, which are parts of a molecular network representing background knowledge about molecular mechanisms. In our previous work (Chereda et al., 2021) we used a protein-protein interaction network as a molecular network. The vertices of a predictive subnetwork are selected genes that are highly relevant for a classifier's individual decision. Additionally, the GLRP approach allows for selecting not only a feature subset relevant for an individual patient, but also a subset of features important for the classifier decisions made over a whole dataset. Here we aim to estimate the stability of feature subsets selected by GLRP w.r.t. different training samples provided from the same data.

Besides, we applied the SHAP method (Lundberg and Lee, 2017) to GCNN to interpret its individual decisions, and to deliver patient-specific subnetworks that can be compared with the subnetworks delivered by GLRP (GCNN+LRP). As well as GLRP, SHAP allows for the selection of a general subset of features by quantifying feature importance scores over a whole dataset. We analyze the stability estimates for GCNN+LRP and GCNN+SHAP and the properties of subnetworks delivered by these two approaches.

The contributions of this work are the following:

- Present the Keras (Chollet, 2015) compatible graph convolutional layer of the GCNN method (Defferrard, Bresson, and Vandergheynst, 2016) allowing for creating a Keras *Sequential* GCNN model, so that the SHAP method could explain it.
- Estimate and compare the stability of feature selection performed by GCNN+LRP, GCNN+SHAP and other machine learning based approaches.
- Compare and analyze the subnetworks delivered by GCNN+LRP and GCNN+SHAP: quantify the permutation importance of the features among patient-specific subnetworks as well as their connectivity.

2 Materials and Methods

2.1 Protein-Protein Interaction Network

The gene expression data was structured with the Human Protein Reference Database (HPRD) protein-protein interaction (PPI) network (Keshava Prasad et al., 2009) It contains protein-protein interaction information based on yeast two-hybrid analysis, in vitro and in vivo methods. The set of binary interactions between pairs of proteins in the HPRD PPI network represented as an undirected graph. The graph is not connected.

2.2 Breast Cancer Data

2.2.1 Metastases Dataset

We applied our methods to a large breast cancer patient dataset that we previously studied and preprocessed (Bayerlová et al., 2017) That data is compiled out of 10 public microarray datasets measured on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A arrays. The datasets are available from the Gene Expression Omnibus (GEO) (Barrett et al., 2013) data repository and have the accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, GSE6532. The data preprocessing is the same as in our previous work (Chereda et al., 2021, "Breast cancer data" section of) After pre-processing, the dataset consisted of 12179 genes and 969 patients. The patients were divided into two classes: 393 patients with distant metastasis occurred within the first 5 years, and 576 patients without metastasis having the last follow-up between 5 and 10 years.

After genes were mapped to the vertices of the HPRD PPI network, the main connected component of the resulting graph consisted of 6888 vertices. GCNN'S input dimensionality is equal to 6888 as well.

2.2.2 Subtype Dataset

We have also applied our approaches on another RNA-seq based gene expression dataset of human breast cancer patient samples. A label of each patient corresponds to a breast cancer molecular subtype. The expression (batch normalized from Illumina HiSeq_RNASeqV2) and clinical data are provided by The Cancer Genome Atlas (TCGA), were downloaded from (*cBioPortal TCGA-BRCA PanCancer data* 2018) The expression data comprise the collection of 20531 genes and 1082 samples. After mapping sample's IDs to clinical data (containing subtype labels) we ended up with 981 samples of breast cancer, corresponding to five subtypes: luminal A (499 samples), luminal B (197 samples), basal-like (171 samples), HER2-enriched (78 samples) and normal-like (36 samples).

Neighboring genes within a molecular network should have similar expression profiles. To promote gene expression similarities, the gene expression data was normalized utilizing the gene length corrected trimmed mean of M-values (GeTMM) method (Smid et al., 2018) It allows for inter- and intrasample analyses with the same normalized data set. After that we applied $log_2(x+1)$ transform to reduce the scale. The expression data were mapped to vertices of PPI resulting in 8469 genes in the main connected component.

2.3 ML methods for feature selection

2.3.1 GCNN+LRP

In our recent work (Chereda et al., 2021) we developed the Graph Layer-wise Relevance Propagation (GLRP) — a method that adapts LRP (Bach et al., 2015) to graph convolution layers of GCNN (Defferrard, Bresson, and Vandergheynst, 2016) and explains GCNN's patient-specific decisions. GCNN was applied to two breast cancer datasets ("2.2 Breast Cancer Data"). The HRPD PPI network ("2.1 Protein-Protein Interaction Network") was used to structure gene expression data. The GLRP method (can be also referred as GCNN+LRP) computes a relevance value for each feature of an individual data point representing a cancer patient. A single relevance value shows how much a particular feature influences a classifier's decision.

As in our previous work (Chereda et al., 2021, "GLRP on gene expression data" section of) GCNN is trained on training data and the subnetworks are generated by GLRP on test data. The number of GCNN's output neurons corresponds to the number of classes in a classification task. Also for binary classification, GCNN had two output neurons that showed the probability of the two classes. For each patient in a test set, relevance was propagated by GLRP from the output neuron (corresponding to the ground truth label even if a data-point was misclassified) to the input neurons representing genes (vertices) of the underlying molecular network. In our setup, GLRP propagates only positive contributions to a predicted class.

Let g_p be the set of 140 most relevant genes for a single patient where p corresponds to a patient's index. The genes of the set g_p are mapped to the vertices of an underlying molecular network, creating a patient-specific subnetwork. This subnetwork, that explain the prediction of a single patient, consists from 140 genes in the set g_p and corresponding to g_p edges from the underlying molecular network. The description of how to construct a feature subset using GCNN+LRP is given in "2.5 Selecting a feature subset via LRP and SHAP" section.

2.3.2 GCNN+SHAP

Additionally, we generated patient-specific subnetworks applying SHAP method (Lundberg and Lee, 2017) to GCNN trained on breast cancer subtype data ("2.2.2 Subtype Dataset"). The SHAP method explains single decisions of a classifier in a similar to LRP manner, but instead of relevances it estimates Shapley values. The Shapley value is a term established in cooperative game theory. According to Molnar, 2019, the game theory setup behind Shapley values is the following: The "game" is the prediction task for a single data point. The "payout" is the difference between the actual prediction for this data point and the average prediction for all instances. The "players" are the feature values of the data point that collaborate to receive the "payout" (predict a certain value). Shapley values indicate how to fairly distribute the "payout" among the features. A single Shapley value represents an importance measure of a particular feature value of a data point that was fed into the classifier.

The SHAP's DeepExplainer approach suitable for convenient deep learning models was not applicable for GCNN and in our previous work (Chereda et al., 2021, "Discussion" section of) the KernelExplainer was utilized to explain GCNN, although the estimation of Shapley values took very long. To make explanations delivered faster within 10-fold cross validation, we have implemented graph convolution as a separate Keras layer and built a GCNN model as a Keras sequential model. The SHAP's DeepExplainer approach was applied to our Keras implementation of GCNN. Similarly to GLRP, for each patient we create a set g_p of top 140 genes with the highest positive Shapley values, which were pushing prediction to a higher probability of the ground truth label. As background data for integrating out the features we used training dataset, and the Shapley values were estimated for the test test. The positive Shapley values are referred as feature relevance values in "2.5 Selecting a feature subset via LRP and SHAP" section that describes how to construct a feature subset using GCNN+SHAP.

2.3.3 MLP+LRP and MLP+SHAP

Multi-Layer Perceptron (MLP) is a feed-forward neural network. In this work MLP was trained on breast cancer subtype data ("2.2.2 Subtype Dataset"). MLP consisted of three hidden fully-connected layers with 1024 units each. Rectified linear unit was used as activation function. Five output neurons correspond to five subtypes of breast cancer. For the performance results on ("2.2.1 Metastases Dataset") we refer the reader to (Chereda et al., 2019)

The set g_p of 140 most relevant genes can be generated as a data point specific explanation of a single MLP's decision. For comparison, we applied both LRP and SHAP to MLP to deliver patient-specific explanations. The MLP approach does not use prior knowledge. Thus, in the context of MLP, we refer to patient subnetworks only as a set g_p for the sake of simplicity. A feature subset, corresponding to a dataset, is built with MLP in the same way as with GCNN and described in "2.5 Selecting a feature subset via LRP and SHAP" section.

2.3.4 GLMGRAPH and Random Forest

Chen et al., 2015 developed a 'glmgraph' method that implements network-constrained sparse regression model. HPRD PPI was used as an underlying network. The idea of the network constraint is to shrink the difference between the estimated coefficients of the connected predictors. The selection of tuning parameters for the sparsity and network constraints was performed within a separate run of 5-fold cross-validation. For 'glmgraph', important features were selected according to the ranking of their absolute coefficients in the linear model.

Random Forest is a tree-based ensemble machine learning technique that combines bagging and random subspace method. It does not incorporate any prior knowledge, but is widely used as a baseline tool for high-dimensional data analysis. We trained Random Forest with 10000 trees. Important features were selected on the basis of mean decrease in Gini impurity.

2.4 Measuring the stability of a feature selection algorithm

The input of a feature selection procedure is the data set $\{x_i, y_i\}_{i=1}^n$ where each x_i is a m-dimensional feature vector and y_i is the associated label. Feature selection identifies a feature subset S of the dimensionality $k < m$ (Nogueira, Sechidis, and Brown, 2018) The subset S conveys the most relevant information about the label y. The output of a feature selection approach is either a scoring on the features, a ranking of the features, or a subset of the features. Thus, the output of any feature selection method can be treated as a subset selection. Further in this paper, we do not consider the scoring information about features selected and treat them as a set. The input dataset of a feature selection technique is a finite sample that is created by a generating distribution. In the case of varying samples, the selected feature subset may vary as well. The variation of the feature subset is the stability that we aim to measure.

A typical approach to measure stability is to produce M subsamples of the dataset at hand, to apply a feature selection approach to each one of them, and then to measure the variability in the M feature sets obtained (Nogueira, Sechidis, and Brown, 2018) Let $Z = \{S_1, ..., S_M\}$ be a collection of feature sets. Let $\phi(S_i, S_j)$ be a symmetric function taking two feature sets as input and returning their similarity value and let $\hat{\Phi}$ be a function taking Z as input and returning a stability value. Nogueira, Sechidis, and Brown, 2018 provide a good overview over stability measuring techniques. We utilize similarity based approach, so that $\hat{\Phi}$ can be defined as the average pairwise similarity between the

$M(M-1)/2$ possible pairs of feature sets in Z:

$$\hat{\Phi} = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j>i}^{M} \phi(S_i, S_j). \quad (1)$$

One of the techniques to generate subsamples is bootstrap. Another approach is random subsampling (Wald, Khoshgoftaar, and Dittman, 2012) In this work we use subsamples within 10-fold cross validation, therefore $M = 10$. As an easily interpretable pairwise similarity function, we use Jaccard distance:

$$\phi(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \quad (2)$$

2.5 Selecting a feature subset via LRP and SHAP

Since the stability measure in equation (1) requires correspondence of a single feature subset to a single dataset, we used two generic ways to construct a feature subset S for the approaches GCNN+LRP, GCNN+SHAP, MLP+LRP, MLP+SHAP. Further in this section, which is a follow-up of 2.3.1, 2.3.1 and 2.3.3, we refer as feature relevances to both the values delivered by LRP and the values computed by SHAP. Given the set g_p of top 140 genes with the highest feature relevances of a single patient, we denote a set $\hat{S} = \cup_p g_p$ as a union of subnetworks' genes of all the patients in test data. The two ways to construct a feature subset, which can be used to measure the stability of feature selection, are the following:

1. We rank genes among patient subnetworks genes $\hat{S}$ according to their frequency in subnetworks. There, we select the set $\hat{S}^{140}$ of 140 top frequent genes among subnetworks.
2. We compute average feature relevances of genes across patients in the test set and select top 140 genes with the highest average feature relevances into the set $\bar{S}^{140}$.

The feature subset $\hat{S}$ was not used to estimate the stability of feature selection. This subset represents rather differences across patients, while subsets $\hat{S}^{140}$ and $\bar{S}^{140}$ contain features that are common or averaged across patients.

We also compare the stability measures based on $\bar{S}^{140}$ to the stability measures of top 140 important features from Random Forest using no prior knowledge and from 'glmgraph' method (Chen et al., 2015) implementing network-constrained sparse regression model using HPRD PPI network.

Two types of feature subsets, that can be delivered by GCNN+LRP, GCNN+SHAP, MLP+SHAP, and MLP+LRP ($\hat{S_i}^{140}$ and $\bar{S_i}^{140}$, $i \in \{1, 2, ..10\}$) are generated in scopes of 10-fold cross validation. The stability measures on the subsets above are presented in "3 Results" section.

2.6 Measuring the permutation importance of patient-specific subnetworks prioritized by LRP and SHAP

Apart from the feature selection stability, one can estimate another valuable property — the permutation importance of features that are relevant for individual decisions made by a particular ML model. The permutation importance of a particular feature is calculated as a drop in classification score when the values of this feature are permuted. We measure the permutation importance of all the genes that are included in patients' subnetworks. Following the notations from the previous section, we define the set of important genes as the union of the subnetworks' genes of all the patients in the dataset:

$$G = \bigcup_{i=1}^{M} \hat{S}_i = \bigcup_{p=1}^{n} g_p, \quad (3)$$

Table 1. Stability of gene selection, metastases prediction. In the last column, for Random Forest top important 140 features are selected according to the decrease of Gini impurity, while for 'glmgraph' according to the absolute value of their coefficients.

Method	Top 140 most frequent genes within subnetworks per fold, subsets $\hat{S_i}^{140}$, %	Top important 140 genes per fold, subsets $\bar{S_i}^{140}$, %
GLRP	92.13	92.10
Random Forest	-	63.61
glmgraph	-	56.22

where $M = 10$ since subnetworks are generated using 10-fold cross-validation, and n is a number of patients in the dataset. The subnetworks can be generated either by LRP or SHAP methods.

The permutation importance of the genes G was calculated in another additional run within 10-fold cross validation. Inside of each iteration, we provide three test sets instead of one: $T_i^1, T_i^2, T_i^3, i \in \{1, 2, ..10\}$. The first T_i^1 is a usual one as it was during the initial run of 10-fold cross validation generating subnetworks. The second one T_i^2 is based on T_i^1, but the gene expression values of genes G are randomly and independently permuted across patients. The third one T_i^3 is created by shuffling expression values of $|G|$ randomly selected genes. The performance difference between T_i^1 and T_i^3-like test sets is used as a baseline to compare with the performance difference between T_i^1 and T_i^2-like test sets.

3 Results

3.1 Stability of feature selection

3.1.1 GLRP on the metastases dataset

The stability of feature selection performed by GLRP on the dataset described in "2.2.1 Metastases Dataset" section was measured as it is written in "2.5 Selecting a feature subset via LRP and SHAP" section. The GCNN architecture consisted of two graph convolutional layers following maximum pooling of size 2, and two hidden fully connected layers with 512 and 128 units respectively. Each graph convolutional layer contained 32 filters covering a vertex' neighborhood with seven hops. We utilized two other baselines as we did in our previous research (Chereda et al., 2021) a 'glmgraph' method (Chen et al., 2015) implementing network-constrained sparse regression model (HPRD PPI as prior knowledge), and Random Forest (no prior knowledge). 'glmgraph' was evaluated on standardized data, since it had convergence issues otherwise. The performance results of 10-fold cross validation of these methods are available in (Chereda et al., 2021, Table 1 of) The stability estimates shown in Table 1 are based on feature subsets $\hat{S_i}^{140}$ and $\bar{S_i}^{140}$ described in "2.5 Selecting a feature subset via LRP and SHAP" section. The stability metrics demonstrate that the feature selection using GLRP is substantially more stable than using 'glmgraph' or Random Forest. For 'glmgraph', the top 140 important features were selected according to their absolute coefficients in the linear model. For Random Forest, top 140 important features were selected on the basis of mean decrease in Gini impurity.

3.1.2 GLRP on the subtype dataset

On the RNA-seq dataset described in "2.2.2 Subtype Dataset" section a slightly different GCNN architecture was applied and our analyses additionally included multilayer perceptron (MLP) method. The GCNN architecture consisted of two graph convolutional layers following average pooling of size 2, and two hidden fully connected layers with 512 units each. Each graph convolutional layer contained 32 filters covering a vertex' neighborhood with seven hops. MLP consisted of three hidden fully-connected layers with 1024 units each.

Table 2. Performance of GCNN predicting the breast cancer subtype. 'glmgraph' performs binary classification, LumA vs rest.

Method	Multiclass	Accuracy, %	F1-weighted, %
GCNN	+	91.33±0.77	91.29±0.71
MLP	+	91.54±0.68	91.30±0.73
Random Forest	+	87.06±0.83	85.82±1.00
glmgraph	-	88.99±1.55	88.99±1.54

Table 3. Stability of gene selection, breast cancer subtype prediction. In the last column, for Random Forest top important 140 features are selected according to the decrease of Gini impurity, while for 'glmgraph' according to the absolute value of their coefficients.

Method	Top 140 most frequent genes within subnetworks per fold, subsets $\hat{S_i}^{140}$, %	Top important 140 genes per fold, subsets $\bar{S_i}^{140}$, %
GLRP	92.29	92.68
Random Forest	-	83.96
glmgraph	-	58.21
MLP+LRP	34.93	34.84
MLP+SHAP	62.07	39.84
GCNN+SHAP	55.88	25.63

While the RNA-seq dataset has 5 different classes, the 'glmgraph' method is only suitable for binary classification. Thus, 'glmgraph' performed luminal A (499 data points) vs other subtypes (482 data points) binary classification. The data was standardized only for 'glmgraph'. The performance of the methods was measured using 10-fold cross validation and the results are depicted in Table 2. As we can see in Table 2, the MLP and GCNN demonstrate similar performances, while Random Forest and 'glmgraph' show worse classification scores.

To have more holistic picture on how LRP and SHAP influence the stability of feature selection, we applied LRP and SHAP to MLP and compared GCNN+LRP (GLRP) with GCNN+SHAP. The stability estimates are presented in Table 3 and were obtained according to the procedure detailed in "2.5 Selecting a feature subset via LRP and SHAP" section.

Compared to the metastases dataset, the stability of GLRP, Random Forest, and 'glmgraph' applied to the subtype dataset were higher. GLRP demonstrated a slight increase in stability w.r.t. $\bar{S_i}^{140}$ subsets (92.10 % vs 92.68 %). Random Forest showed higher stability estimates (63.61 % vs 83.96 %) as well as 'glmgraph' (56.22 % vs 58.21 %). The rise of the stability estimates indicates that the subtype dataset has higher quality than the metastases dataset. While the stability estimates are lower for LRP than for SHAP when both are applied to MLP, the situation is the opposite when both applied to GCNN utilizing the prior knowledge. Furthermore, GLRP provides the highest stability compared to other methods shown in Table 3.

3.2 Comparing properties of subnetworks prioritized by LRP and SHAP

The results showed in the previous section highlight the differences between stability estimates computed for the SHAP and LRP methods explaining MLP or GCNN models. We examine these differences further on the same breast cancer subtype dataset ("2.2.2 Subtype Dataset") by computing the permutation importance for the set of important genes G, which is the union of the subnetworks' genes of all the patients in the dataset. The permutation importance was calculated within 10-fold cross validation. Inside of each iteration, we provide three test sets instead of one. The first test set is a usual one. The second test set has shuffled expression values across patients for the genes from the set G. The third one has shuffled expression values across patients for $|G|$ randomly selected genes. Comparing classification performances on those three test sets, one can evaluate the permutation feature importance as a performance drop caused by shuffling the expression values of the subnetworks' genes G. The results are presented in Table 4. The set G as well as the procedure to measure the permutation importance are described in "2.6 Measuring the permutation importance of patient-specific subnetworks prioritized by LRP and SHAP" section.

One notices, that the performance drop between T_i^2, T_i^3 when GLRP prioritizes 140 top genes per patient, is quite moderate - a bit more than 3 % (Table 4). Also, the set G contains quite small amount of genes - 836 out of 8469. In the second row of Table 4, the increase of the size of a patient's subnetwork to 600 genes ($|G| = 2712$) lead to the increase of the performance drop between T_i^2, T_i^3 up to around 10 %. The stability estimates (when a patient subnetwork consists of 600 genes) for the subsets $\hat{S_i}^{600}$ and $\bar{S_i}^{140}$ are the following: 92.66% and 92.68%. It indicates that increase of subnetworks' size does not influence the stability estimates.

The permutation importance of the features selected by GCNN+SHAP is demonstrated in the third row of Table 4. The feature set G contains higher number of genes (4172 for GCNN+SHAP vs 836 for GLRP), which indicates that the individual patient subnetworks differ across the patients much more than in the case of GLRP. The performance drop between T_i^2, T_i^3 is around 40 % that shows that genes selected by SHAP carry higher importance for classification decisions than genes selected by LRP. In other words, from the perspective of feature selection, the fraction of false positive genes among patient subnetworks prioritized by LRP is higher than the fraction of false positive genes among patient subnetworks prioritized by SHAP. Another cornerstone of the patient's subnetworks is interpretability in the context of underlying prior knowledge (HPRD PPI network). We compared the connectivity of individual subnetworks delivered by GCNN+SHAP and GLRP by counting the number of connected components in them. The distributions of the number of connected components in subnetworks are displayed as boxplots in Figure 1. While the subnetworks generated by GLRP have on average 16 connected components, the subnetworks generated by GCNN+SHAP have 126 of them. In contrary to the GLRP subnetworks, the genes prioritized by GCNN+SHAP can hardly be interpreted in the context of the HPRD PPI network since a subnetwork generated by GCNN+SHAP consists mainly of singletons.

The last two rows in Table 4 compare the behavior of SHAP and LRP applied to MLP that does not use any prior knowledge. As in the case of GCNN, SHAP features has lower amount of false positives than LRP features. Comparing the fourth and the second row, one can notice that the performance drop on MLP+LRP is higher than that on GLRP even though the number of genes with permuted vertices was lower for MLP+LRP. Perhaps the reason for GLRP to demonstrate such a behavior is that if a gene which is not that important for classification is adjacent to an important one, it can be assigned abundant relevance if the expression values of these genes are similar and the corresponding weights of graph convolutional filters have similar values.

4 Discussion

The focus of our paper is to investigate the stability of feature selection performed by the GCNN+LRP approach (GLRP) and to compare it to the stability of feature selection performed by GCNN+SHAP. Moreover, the stability of GLRP was compared to that of more commonly used algorithms, such as Random Forest and network-constrained sparse regression model. The stability estimates for GLRP are the highest among all the feature selection approaches used in this paper. Surprisingly, for GCNN+SHAP the stability estimates are among the lowest and the

Table 4. Performance drop by permuting subnetworks' genes values across patients in test sets. The measure for the peformance drop is F1-weighted score.

Method	Performance, usual test sets T_i^1, %	Performance, permuting values of subnetworks' genes G, test sets T_i^2, %	Performance, permuting values of $\|G\|$ randomly selected genes, test sets T_i^3, %	$\|g_p\|$, number of selected top relevant genes per patient,	$\|G\|$, number of genes with permuted values
GLRP	91.29±0.71	87.55±0.71	90.79±0.95	140	836
GLRP	90.17±0.95	76.18±1.40	86.54±0.97	600	2712
GCNN+SHAP	91.60±0.84	43.81±1.18	82.73±1.33	140	4172
MLP+LRP	91.30±0.73	69.62±1.84	87.79±0.66	140	2372
MLP+SHAP	91.17±0.71	40.78±1.19	86.54±1.35	140	2952

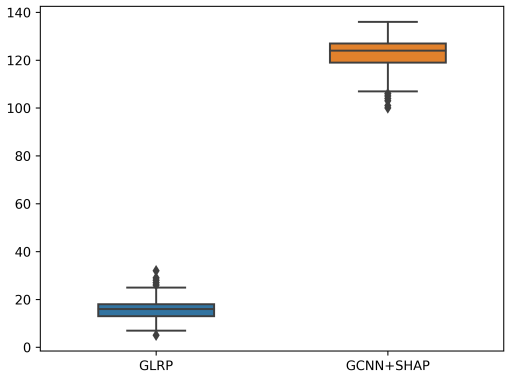

Fig. 1. The distribution of the number of connected components in patients' subnetworks. The left boxplot corresponds to the subnetworks obtained by GLRP while the right one corresponds to the subnetworks obtained by GCNN+SHAP.

GCNN+SHAP subnetworks are much less similar between patients than the GLRP subnetworks. As for the permutation importance, the situation is completely opposite: the subnetworks' genes prioritized by GCNN+SHAP are more important for GCNN's decisions than the the subnetworks' genes prioritized by GCNN+LRP. Although one should take into account that the number of all subnetwork genes is more than four times higher for GCNN+SHAP than for GLRP.

One one hand it is expected to have very different patient-specific subnetworks because cancer is a heterogeneous disease. On the other hand, the connectivity properties of GCNN+SHAP subnetworks are poor since they mainly consist of single vertices that are disconnected within the HPRD PPI network. On contrary, GLRP produces connected subnetworks. We hypothesise, that the GLRP method smoothes the relevances across layer's nodes of a neural network while propagating them from output to input layers.

In the case of MLP models, the permutation importance is also substantially higher for SHAP features than for LRP features that perhaps supports our previous claim. Comparing GLRP and MLP+LRP, one can notice that the permutation importance of the genes prioritized by MLP+LRP is higher than that of the genes prioritized by GCNN+LRP. Investigating properties of the distribution of relevance, gene expression values, and weights among input features of GCNN and MLP, one could potentially check the hypothesis mentioned in the previous paragraph but we leave it for our future research.

Additionally, we noticed that the frequency, with which a gene is prioritized by LRP (for both GLRP and MLP+LRP), correlates with the expression value of a gene - this correlation is around 0.47. For the SHAP method the same correlation is less then 0.10. We assume that the LRP has a slight bias towards genes with higher expression values, and this property also needs to be investigated further.

The performances of MLP and GCNN on the breast cancer subtype data are basically the same. This fact questions the superiority of GCNN over other ML methods in classification tasks. In our recent research (Alachram et al., 2021) we utilized three additional microarray cancer datasets. We have checked how the GCNN's performance depends on prior knowledge and also compared it to the performance of Random Forest. We found out that the performances of GCNN and Random Forest were comparable. Moreover, permutation of nodes of an underlying molecular network did not substantially alter the classification performance of GCNN (Alachram et al., 2021) It can be explained by our assumption that the expression correlations between genes did not coincide well with provided network topologies (Alachram et al., 2021) This property is worth to be studied further as well.

5 Conclusion

We have investigated the stability of feature selection procedure based on the GLRP (GCNN+LRP) approach delivering patient-specific subnetworks. Its stability was also compared to the stability of feature selection of more classical methods such as Random Forest and generalized linear model with graph constraints. Additionally, we have studied the prioritization of features performed by the SHAP and LRP explanation methods that were applied to GCNN and MLP. We conclude that GLRP provides the highest stability in feature selection compared to other approaches. Patient-specific features prioritized by SHAP had consistently higher permutation importance than patient-specific LRP features when LRP and SHAP were applied to GCNN as well as to MLP. It was also established, that highly unstable approach MLP+LRP (no prior knowledge) prioritizes features with permutation importance higher than that of features prioritized by GCNN+LRP. Our further investigation of subnetworks that were prioritized by GCNN+LRP and GCNN+SHAP showed that while the subnetworks generated by GCNN+SHAP had higher permutation importance for GCNN's decisions, the subnetworks generated by GLRP were much more connected in contrast to the subnetworks delivered by GCNN+SHAP that consisted mainly of single vertices. Therefore, the subnetworks generated by GLRP are more interpretable in the context of prior knowledge compared to the subnetworks obtained from GCNN+SHAP.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was funded by the German Ministry of Education and Research (BMBF) e:Med project *MyPathSem* (031L0024). We would like to acknowledge Johannes Söding for fruitful discussions. H.C. is a member of the International Max Planck Research School for Genome Science, part of the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciences. T.B. is a member of the Göttingen Campus Institute Data Science.

References

Alachram, Halima et al. (2021) "Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks". en. In: *PLOS ONE* 16.10. Publisher: Public Library of Science, e0258623.

Bach, Sebastian et al. (2015) "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". en. In: *PLOS ONE* 10.7, e0130140.

Barrett, Tanya et al. (2013) "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic Acids Res* 41.Database issue, pp. D991–D995.

Bayerlová, Michaela et al. (2017) "Ror2 Signaling and Its Relevance in Breast Cancer Progression". English. In: *Front. Oncol.* 7. Publisher: Frontiers.

cBioPortal TCGA-BRCA PanCancer data (2018) `https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018`.

Chen, Li et al. (2015) "glmgraph: an R package for variable selection and predictive modeling of structured genomic data". In: *Bioinformatics* 31.24. Publisher: Oxford Academic, pp. 3991–3993.

Chereda, Hryhorii et al. (2019) "Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer". eng. In: *Stud Health Technol Inform* 267, pp. 181–186.

Chereda, Hryhorii et al. (2021) "Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer". In: *Genome Medicine* 13.1, p. 42.

Chollet, François (2015) *Keras*. `https://github.com/fchollet/keras`.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016) "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *arXiv:1606.09375 [cs, stat]*. arXiv: 1606.09375.

Johannes, Marc et al. (2010) "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients". eng. In: *Bioinformatics* 26.17, pp. 2136–2144.

Keshava Prasad, T. S. et al. (2009) "Human Protein Reference Database—2009 update". In: *Nucleic Acids Res* 37.Database issue, pp. D767–D772.

Lee, Hae Woo et al. (2013) "Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery". In: *Statistical Applications in Genetics and Molecular Biology* 12.2, pp. 207–223.

Lundberg, Scott and Su-In Lee (2017) "A Unified Approach to Interpreting Model Predictions". In: *arXiv:1705.07874 [cs, stat]*.

Molnar, Christoph (2019) *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

Nogueira, Sarah, Konstantinos Sechidis, and Gavin Brown (2018) "On the Stability of Feature Selection Algorithms". In: *Journal of Machine Learning Research* 18.174, pp. 1–54.

Perera, Julia, Andreas Leha, and Tim Beissbarth (2019) "Bioinformatic Methods and Resources for Biomarker Discovery, Validation, Development, and Integration". In: *Predictive Biomarkers in Oncology*. Ed. by Sunil Badve and George Kumar. Springer International Publishing. Chap. 11, pp. 149–164.

Porzelius, Christine et al. (2011) "Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients". en. In: *Biometrical Journal* 53.2, pp. 190–201.

Smid, Marcel et al. (2018) "Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons". In: *BMC Bioinformatics* 19.1, p. 236.

Sørlie, Therese (2007) "Molecular Classification of Breast Tumors: Toward Improved Diagnostics and Treatments". In: *Target Discovery and Validation Reviews and Protocols*.: Humana Press, pp. 91–114.

Wald, Randall, Taghi Khoshgoftaar, and David Dittman (2012) "A New Fixed-Overlap Partitioning Algorithm for Determining Stability of Bioinformatics Gene Rankers". In: *2012 11th International Conference on Machine Learning and Applications*. Vol. 2, pp. 170–177.

5 Discussion

In this thesis, the main aim was to develop the methodology that allows for delivering patient-specific molecular subnetworks. The methodology is based on a data-driven ML approach that is guided by prior knowledge of molecular networks and predicts a clinical outcome. The ML approach consists of an ML model and an explanation method, applied to the ML model. The explanation method generates patient-specific subnetworks that drive individual decisions of the ML model. These molecular subnetworks are parts of a large molecular network used as prior knowledge, and contain genes that are potentially druggable drivers of tumor progression [16]. Since the subnetworks are delivered on the individual patient level, the developed methodology can be not only useful in precision medicine approaches, promoting *individualized medicine*, but also applicable as a general feature selection approach [17].

This chapter is structured into three main parts (see section 1.4.3 *Organization of the thesis*) according to Chapters 2, 3, 4, and reflecting the aims of the thesis described in section 1.4:

1. **Predicting clinical endpoints with GCNN utilizing prior knowledge**. The prediction task is formulated as a classification task and is performed by a ML method utilizing prior knowledge of a molecular network. In this thesis, a PPI network was utilized by the GCNN method [21] that performs classification of patients' gene expression profiles. The gene expression profiles are structured by a PPI network, forming graph signals (see Figure 1.1), and GCNN learns their patterns to classify patients. Two different classification tasks, which correspond to two different clinical endpoints, were considered in the scopes of this thesis. First, to predict an occurrence of a metastatic event during the first 5 years after treatment (see Chapters 2, 3); Second, to classify patients according to breast cancer subtypes (see Chapter 4).

2. **Explaining GCNN to deliver patient-specific subnetworks responsible for the prediction of the clinical outcome**. Individual decisions of a GCNN model performing predictions can be explained by the methods of XAI. In scopes of this thesis, the GLRP method [16] was developed as an adaptation of the LRP [4, 7] explanation method to graph convolutional layers of GCNN [21]. GLRP can explain a patient-specific decision of a GCNN model by assigning non-negative relevance values to the input features (genes). A patient-specific subnetwork is constructed out of the most relevant for an individual decision genes (see Figure 1.2). The details

on GLRP method as well potential utility of subnetworks in precision medicine approaches are given in Chapter 3 *Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer*.

3. **Stability of feature selection performed by GLRP**. Another caveat point is the stability of feature selection that can be performed by GLRP for the whole dataset. The stability of feature selection is intrinsically linked to the biological intepretability of the selected features [52, 24, 22]. Furthermore, it is crucial to compare the stability of feature selection performed by GLRP (GCNN+LRP) also to other feature selection approaches, including SHAP as a XAI method [43]. The graph convolutional layer of GCNN initially implemented in TensorFlow [1] was reimplemented as a Keras [18] layer (see Chapter 4) such that SHAP could be applied to a Keras GCNN model. Furthermore, both SHAP and LRP can derive patient-specific features when applied to a MLP model solving the same classification task. To enhance holism in the view on feature selection approaches, I compared the permutation importance of patient-specific features prioritized by LRP and SHAP (patient-specific subnetworks when a GCNN model is explained). Additionally, the connectivity of the subnetworks delivered by GCNN+LRP or GCNN+SHAP was analyzed. Chapter 4 *Stability of feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation* provides a systematic overview on the properties of features prioritized by GLRP.

5.1 Predicting clinical endpoint with GCNN utilizing prior knowledge

5.1.1 Questioning the superiority of GCNN's performance

In the study [15], the GCNN method [21] was applied to gene expression data structured by *Human Protein Reference Database* (HPRD) PPI network [30] to predict occurence of metastasis. The GCNN method slightly outperformed Random Forest, MLP and Lasso logistic regression, although the difference in performance metrics was not substantial. In Rhee et al. [61] GCNN utilizing prior knowledge of the PPI network from STRING database [74] showed better classification result than, for example, Random Forest and Support Vector Machine to predict breast cancer subtype. Surprisingly, in the task of breast cancer subtype prediction, the MLP approach without any prior knowledge had the same performance as the GCNN method utilizing HPRD PPI network [17, Table 2 of].

Since GCNN [21] is an extension of usual CNNs to graph domain it is worth mentioning previous works [58, 44, 51]. These three works utilized the same dataset comprising gene expression data from roughly 11000 samples of 33 cancer types of *The Cancer Genome Atlas* (TCGA) project [13]. While Ramirez et al. [58] applied GCNN [21] incorporating

co-expression or PPI network, authors in [44] and [51] applied CNNs to a gene expression profile of a patient. CNNs utilize 1D, 2D, or 3D data, thus gene expression had to be structured. Lyu et al. [44] defines a single gene expression profile as a 1D vector, where elements (genes) are ordered according to their chromosomal positions. In contrary, Mostavi et al. [51] arbitrarily order genes in a 1D vector. Then, this 1D vector is converted into a 2D grayscale image [44, 51], e.g. a vector of length 10000 can be reshaped as 100x100 matrix, which is fed as a single input data point to CNN. Classifying 2D gene expression profiles, Lyu et al. [44] achieved 95.6 % accuracy, while Mostavi et al. [51] reached 95.7 % accuracy on the same TCGA dataset. Comparing the performances stated in the works of Mostavi et al. [51] and Lyu et al. [44], one should notice that the order of genes did not affect the performance of CNNs. The accuracy of GCNN in Ramirez et al. [58] is 94.6 %, which is lower than the accuracy of CNNs. It can be explained by the fact that the GCNN method utilized only one convolutional layer with only one filter, while in [51] and [44] the convolutional layers had dozens of filters.

Therefore, one cannot conclude that the GCNN method [21] benefits (in terms of performance) from molecular network information while classifying gene expression profiles of patients. The absence of this benefit could be connected to the barely noticeable sensitivity of the GCNN method to an underlying network structuring gene expression data.

5.1.2 Sensitivity of GCNN's performance to an underlying molecular network

Initially, the sensitivity of the GCNN method to an underlying graph was studied in [21]. There, GCNN was applied on the MNIST dataset [39] containing 70,000 images of handwritten digits (ten classes from "0" to "9") each having a size of 28 by 28 pixels. The regular grid underlying each of the images was converted into a graph with $28 \cdot 28 = 784$ nodes. Each of the nodes is connected to its eight nearest neighbors [21]. The constructed graph is shared across images and each image is represented as a graph signal, where a single pixel value is a node attribute. It is shown in [21, Table 5 of] that a random graph connecting pixels substantially deteriorates GCNN's performance. It proves the sensitivity of GCNN to a graph structuring the data.

The dependency of GCNN's classification performance on underlying networks across several different cancer datasets was studied in Alachram et al. [3]. The classification performance was measured by training and testing GCNNs on data structured by each one out of four types of prior knowledge: HPRD PPI network [30], text-mining based embedding network [3], the same network but with permuted gene labels over nodes, and completely random network. The performance of Random Forest that did not use any prior knowledge was also comparable to that of GCNN. The performance of GCNN did not substantially differ w.r.t. four types of prior knowledge on liver, lung and breast cancer datasets. Completely random network with random weights deteriorated convergence (and classification scores) of GCNN on prostate and colorectal cancer datasets. On these datasets, GCNN with text-mining based embedding network had performance comparable to the performance of

GCNN with the network of the same topology but permuted nodes. The random network did not substantially alter GCNN's performance on breast and liver cancer datasets, but deteriorated GCNN's convergence on lung, prostate and colorectal cancer datasets [3]. As possible explanation for the low sensitivity of GCNN to prior knowledge, one could speculate, that the complex correlation structure within patients' gene expression profiles might not be that well reflected in the topologies of provided molecular networks [3].

Change in the prior knowledge alters the patterns of graph signals formed by patients' gene expression profiles. Thus, GCNN learns altered patterns that can involve different sets of genes driving the classification result. The fact that the performance of GCNN does not substantially changes w.r.t. prior knowledge is in line with the previous discoveries [23, 77] based on the analysis of breast cancer gene signatures. In the latter study, 47 published breast cancer outcome signatures were compared to the random gene signatures. Twenty-eight of the published signatures (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worse predictors than the median random signature [77]. The authors of this study provide the following explanation: more than 50% of the breast cancer transcriptome is correlated with cell proliferation, which integrates most prognostic information in this disease [77].

5.2 Explaining GCNN to deliver patient-specific subnetworks responsible for the prediction of the clinical outcome

The general workflow of the methodology developed within this thesis is depicted in Figure 1.2. GCNN utilizes a molecular network that structures gene expression data and is beneficial for the biological interpretability of the patient-specific subnetworks. The patient-specific subnetworks are delivered by GLRP (GCNN+LRP) method [16], but other explanation methods can be applied to GCNN as well. Model agnostic LIME [62] and SHAP [43] explanation approaches provide importance scores for each feature value of an input data point. The LIME method computes explanations based on feature perturbations from a Gaussian distribution, ignoring correlations between features. It leads to the instability of importance scores for a fixed individual data-point. This instability is not favourable for personalized medicine approaches [16]. The applicability of the GCNN+SHAP approach is described and compared with GCNN+LRP in Chapter 4 and is further discussed in section 5.3, and LRP as an explanation approach is discussed in the next section.

5.2.1 Peculiarities of applying LRP to neural networks

The LRP method was adapted to the graph convolutional layers of the GCNN method [21], since LRP is broadly applicable [49], and has great benchmark performance [64]. The LRP method is based on the set of different propagation rules [47] that redistribute relevance from an output node (corresponding to a predicted class) through the hidden layers up to

the input layer of a neural network. Some of the propagation rules are embedded into the theoretical framework of deep Taylor decomposition [48] performing a first order Taylor expansion at each neuron of the neural network (see Chapter 4 for details).

The PatternAttribution method [31] is derived from the context of a purely linear model and data stemming from a linear generative model. PatternAttribution provides explanations in a manner similar to LRP, advancing deep Taylor decomposition by learning a point from data, at which the expansion occurs. Further in this section, the applicability of this method is considered in the context of LRP.

The presence of different relevance propagation rules as well as their applicability to different layers of a CNN model poses a question — *what is the best practice of applying LRP?* According to [49] the default choice of LRP rules should be on the ones derived from the framework of deep Taylor decomposition [48]. There are three such rules and their application is dependent on the restrictions on the input space. The rules correspond to real-valued input spaces that are unconstrained, non-negative or box-constrained (like pixel values). In this thesis, the input data for GCNN are gene expression values which are non-negative. The rectified linear units, producing the non-negative outputs, were used as activations of the GCNN models trained in this thesis. Thus, the same LRP rule, dealing with non-negative values ([16, Equation (7) of]), was used for every layer of the GCNN models. Besides the deep Taylor decomposition derived rules, there are at least five of others [47, Appendix A of]. According to Kohlbrenner et al. [35], and Montavon et al. [47] the composite LRP, where different parts of CNN are decomposed using purposed rules, provides robustness of explanations against local artifacts while sustaining class sensitivity. In some highly specific cases (see [35, Figure 1 of]) a uniformly applied single LRP rule as well as the PatternAttribution approach are not capable of highlighting class-specific features, while composite LRP is class sensitive. One has to take into account that advantages of composite LRP [35] were shown on deep CNN models with at least a dozen of layers that are applied in visual object detection setting. In the scopes of this thesis the GCNN models were much shallower with four weighted hidden layers. While for shallower NN the same LRP rule applied network-wide works well [35], application of different LRP rules for different layers of GCNN models is worth exploring in future work.

While quantitative approaches are available for evaluating explanations [49], in image classification the examination of LRP derived explanations is commonly done qualitatively. In image domain, such explanations can be easily evaluated by a human. In graph domain, visualization of explanations is more challenging, which is discussed in the next section.

5.2.2 Perception of explanations differs when domain is switched from images to graphs

In image classification, the LRP method explains a CNN model by assigning a relevance value to each pixel of an input image. These relevance values can be visualized as a heatmap showing the regions that were important to classify a particular image. A heatmap contains

patterns that are familiar for human perception. Examples of such heatmaps can be found in [37, 4]. A researcher, perceiving a pattern on a heatmap, can not only estimate if the learned pattern highlights a novel yet unnoticed finding, but also judge if a model relies on the correct strategy for its predictions [38], and potentially identify data selection biases and artifacts.

The developed GLRP method [16] explains decisions of a GCNN model in the same manner as LRP — it assigns relevance values to the input features (genes) that are nodes of the underlying molecular network. This relevance values are non-negative and distributed over a graph of a molecular network that can have more than ten thousand vertices. For instance, a processed PPI network structuring breast cancer data in [16] contains 6888 vertices (genes). Unfortunately, visualizing such a large network with a pattern of distributed relevance values is an extremely challenging task. In this thesis I present patient-specific subnetworks that are highly relevant parts of an underlying molecular network (see Chapter 3). For a single patient, the top 140 most relevant for an individual GCNN's decision genes are selected. Next, a patient-specific subnetwork with 140 vertices is constructed as subgraph of the prior knowledge molecular network (see [16, Figure 3 of]). The number of 140 was selected arbitrarily to preserve visual interpretability of a subnetwork by preventing the "hairy ball" effect.

An explanation in a form of a relevant subnetwork is conceptually different from an explanation in a form of a heatmap, which is common in image classification. This difference is highlighted by the fact, that for a subnetwork, 140 top relevant genes among 6888 network's genes embrace only around 8.5 % of full relevance for a fixed patient. This implies the questions of i) how appropriate is 140 as a threshold for selecting most relevant genes that constitute a subnetwork, ii) how much of importance for classifier's decisions is contained in patient-specific subnetworks. The first question could potentially be resolved by studying advanced graph visualization techniques. Additionally, one could analyze relevance distributions across patients and find an appropriate way to build a patient-specific heatmap which would be ideologically similar to an explanation in image domain. The second question can be studied by measuring the permutation importance of genes that are in patient-specific subnetworks (see Chapter 4 for details).

5.2.3 Sensitivity of GLRP to prior knowledge and its potential applicability in clinical setting

The findings pointed out in section 5.1.2 highlight specific aspects of high-dimensional gene expression data. If gene expression data is structured by a prior knowledge molecular network, then a change in an underlying molecular network alters the formation of patterns within patients' gene expression profiles. For one molecular network, GCNN learns patterns located over some neighborhoods of vertices (genes), while for another, where the latter neighborhoods might be drastically changed, GCNN learn patterns located (vertex-wise) over other neighborhoods, involving altered patient-specific subnetworks important

for classifying the same single patient. Thus, the biological information in a molecular network influences the learning of graph signal representations that implies bias in explanations delivered by GLRP in form of patient-specific subnetworks.

The study [16] showed that PPI based patient-specific subnetworks delivered by the GLRP approach could allow a researcher to learn new insights about biological mechanisms involved in a patient's tumor development. The authors in [16] established that the subnetworks prioritized by GLRP contain biologically meaningful genes. The biological validation was performed by analysing subnetworks' genes delivered by GLRP that was applied to gene expression data from HUVEC treated or not treated with tumor necrosis factor alpha [60]. For the details, the reader can be referred to [16]. Furthermore, when applied to a large breast cancer dataset, subnetworks delivered by GLRP were enriched with common signaling pathways associated with the respective disease and could assign patient-specific priorities to pathway components [16]. The "MTB report methodology" described in [54] was applied to subnetworks to identify actionable genes individually for four selected patients. The *Molecular Tumor Board* (MTB) reports highlighted known and possibly novel genes that could be targeted therapeutically [16]. Although the experimental validation of the patient-specific subnetworks is still missing, it would be promising to involve a parallel culture of patient-derived organoids. Different therapeutic strategies, that could be built on the basis of studying central signalling nodes within patient-specific subnetwork, could be tested on patients organoids [16]. The testing of therapeutic strategies could potentially identify relevant mechanisms and factors involved in tumor resistance on individual patient level and select the best strategy that may provide improvements in current treatment approaches.

5.3 Stability of feature selection performed by GLRP

The text in this section is inspired by section 4 *Discussion* of [17]. The GLRP (GCNN+LRP) approach [16] explains single decisions of a GCNN model via patient-specific subnetworks. GLRP allows for general feature selection, where a feature subset corresponds to a whole dataset. The stability of feature selection performed by GCNN+LRP was estimated by utilizing two approaches of selecting a feature subset that are described in [17]. Additionally, the feature selection performed by GLRP was compared to the feature selection approaches based on other ML methods: "glmgraph", Random Forest and MLP. The "glmgraph" method utilizes prior knowledge molecular network to impose graph constraints on regression coefficients. Random Forest and MLP do not use any prior knowledge. The LRP and SHAP explanation methods were compared in the context of feature selection by applying them to MLP and GCNN. The SHAP explanation method applied to GCNN can deliver patient-specific subnetworks. The properties of subnetworks delivered by GCNN+LRP and GCNN+SHAP are further discussed in this seciton.

Initially, the SHAP's *DeepExplainer* approach suitable for commonly used deep learning

architectures was not applicable [16] to the GCNN method used in this thesis. To fix it, the graph convolution approach of Defferrard et al. [21] was implemented as a separate Keras layer [17]. As a result, a GCNN model could be built as a Keras *Sequential* model. Then, the SHAP's *DeepExplainer* was applied to GCNN models implemented in Keras.

The GCNN+LRP approach exhibits the highest stability in feature selection among the "glmgraph", Random Forest, GCNN+SHAP, MLP+SHAP, MLP+LRP methods applied to two large breast cancer datasets (see Chapter 4). While the stability of GCNN+SHAP is substantially lower than the stability of GCNN+LRP, the permutation importance of genes in the patient-specific subnetworks delivered by GCNN+SHAP is substantially higher. Also, when LRP and SHAP explain a MLP model, then the permutation importance is higher for SHAP features than for LRP features. Probably LRP smoothes the relevance values while propagating them from output to input layers of a NN model [17]. To check this assumption, one could study the properties of the distribution of relevance values across input nodes of NN and GCNN, but it is left for future research.

The connectivity properties of GCNN+SHAP subnetworks are poor. A single patient-specific subnetwork consists mainly of single vertices. In contrary, the GCNN+LRP approach produces connected subnetworks that are interpretable in the context of a prior knowledge molecular network. Presumably the connectivity of GCNN+LRP subnetworks is linked to their lower permutation importance. This link could be explained by the following: if an non-important for the classification gene is adjacent (or in close proximity) to an important one, the LRP method by smoothing out relevance values can assign an abundant relevance to the non-important gene. This phenomena is worth considering in future work by checking the intersection of subnetwork' genes delivered by GCNN+SHAP with the those delivered by GCNN+LRP and checking the proximity of how close GCNN+SHAP genes are located to GCNN+LRP genes within the underlying molecular network.

6 Conclusion

Molecular biomarkers based on data generated by NGS technologies play an increasing role in the prediction of tumor progression or therapy response. Utilizing molecular biomarkers, precision medicine aims to get a broaden view for individualized treatment decisions. Individualize treatment decisions warrant the need to combine molecular biomarkers with the vast amount of knowledge on biological networks, allowing for more holistic view of the patient status.

The main contribution of this thesis towards precision medicine has been the development of the methodology that presents meaningful and interpretable patient-specific molecular subnetworks to clinicians and researchers in order to enable further medical and pharmaceutical insights. The methodology is data-driven and based on transcriptomics data, multinomial clinical endpoint, and a prior knowledge molecular network. By incorporating the molecular network information, researcher introduces bias in patient-specific molecular subnetworks. The methodology consists of two methods: the GCNN method utilizing a prior knowledge molecular network, and an explanation method applicable to the trained GCNN model. The GLRP explanation method was developed in the scopes of this thesis. GLRP is an adaption of the LRP expalantion method to graph convolutional layers of GCNN. GLRP generates patient-specific molecular subnetworks that are relevant for individual classification decisions of GCNN.

The experiments on a large breast cancer dataset has shown, that the subnetworks prioritized by GLRP utilizing general prior knowledge are relevant for the prediction of clinical endpoint (prediciton of metastasis). GLRP provides patient-specific explanations for GCNN that largely agree with clinical knowledge, include oncogenic drivers of tumor progression, and can help to identify therapeutic vulnerabilities. The subnetworks contain not only subtype-specific cancer genes that match the clinical subtype of the patients, but also patient-specific genes that could potentially be linked to aggressive/benign phenotypes.

I have investigated the stability of feature selection procedure based on the GLRP (GCNN+LRP) approach. In the scopes of this investigation I implemented the GCNN method as a Keras compatible *Sequential* model so that SHAP could be applied to GCNN (GCNN+SHAP) and compared to GCNN+LRP. The GCNN+LRP approach has shown the highest stability in feature selection among the methods that use prior knowledge molecular network ("glmgraph", GCNN+SHAP) and methods that do not use it (Random Forest, MLP+SHAP, MLP+LRP) methods. The comparison of the properties of patient-specific features revealed that the features delivered by SHAP had higher permutation importance than the features delivered by LRP, when SHAP and LRP applied to both GCNN and

LRP. While the subnetworks prioritized by GCNN+SHAP had a higher permutation importance for GCNN's decisions, the subnetworks delivered by GLRP were much more connected. The subnetworks prioritized by GCNN+SHAP consisted mainly of single vertices. Thus, the subnetworks prioritized by GLRP are more interpretable in the context of prior knowledge than the subnetworks generated by GCNN+SHAP.

Therefore, the GLRP is a novel and interpretable ML approach for high-dimensional genomic datasets. The subnetworks prioritized by GLRP can be visualized and interpreted in a biomedical context on the individual patient level. GLRP could thus be useful for precision medicine approaches such as for example the molecular tumorboard.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] TaeJin Ahn, Taewan Goo, Chan-hee Lee, SungMin Kim, Kyullhee Han, Sangick Park, and Taesung Park. Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1748–1752, December 2018.

[3] Halima Alachram, Hryhorii Chereda, Tim Beißbarth, Edgar Wingender, and Philip Stegmaier. Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks. *PLOS ONE*, 16(10):e0258623, 2021. Publisher: Public Library of Science.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[6] Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Albrecht Stenzinger, Laura Parlow, Jan Budczies, Benjamin Goeppert, Denise Treue, Manato Kotani, Masaru Ishii, Manfred Dietel, Andreas Hocke, Carsten Denkert, Klaus-Robert Müller, and Frederick Klauschen. Towards

computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. *arXiv:1805.11178 [cs]*, May 2018.

[7] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus Muller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning - 25th International Conference on Artificial Neural Networks, ICANN 2016, Proceedings*, pages 63–71. Springer Verlag, 2016.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[9] Ann M. Bode and Zigang Dong. Recent advances in precision oncology research. *npj Precision Onc*, 2(1):1–6, April 2018. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Editorial Publisher: Nature Publishing Group Subject_term: Cancer Subject_term_id: cancer.

[10] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs, stat]*, May 2021. arXiv: 2104.13478.

[11] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. Conference Name: IEEE Signal Processing Magazine.

[12] Ikram A. Burney and Ritu Lakhtakia. Precision Medicine. *Sultan Qaboos Univ Med J*, 17(3):e255–e258, August 2017.

[13] Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113–1120, October 2013.

[14] Li Chen, Han Liu, Jean-Pierre A. Kocher, Hongzhe Li, and Jun Chen. glmgraph: an R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics*, 31(24):3991–3993, December 2015. Publisher: Oxford Academic.

[15] Hryhorii Chereda, Annalen Bleckmann, Frank Kramer, Andreas Leha, and Tim Beissbarth. Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer. *Stud Health Technol Inform*, 267:181–186, September 2019.

[16] Hryhorii Chereda, Annalen Bleckmann, Kerstin Menck, Júlia Perera-Bel, Philip Stegmaier, Florian Auer, Frank Kramer, Andreas Leha, and Tim Beißbarth. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Medicine*, 13(1):42, 2021.

[17] Hryhorii Chereda, Andreas Leha, and Tim Beissbarth. Stability of feature selection utilizing graph convolutional neural network and layer-wise relevance propagation. *bioRxiv*, 2021.

[18] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[19] European Commission. Europe's Beating Cancer Plan — communication from the Commission to the European Parliament and the Council. `https://ec.europa.eu/health/sites/default/files/non_communicable_diseases/docs/eu_cancer-plan_en.pdf`, 2021. Accessed: 2021-11-21.

[20] Joseph M. de Guia, Madhavi Devaraj, and Carson K. Leung. DeepGx: deep learning using gene expression for cancer classification. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pages 913–920, New York, NY, USA, August 2019. Association for Computing Machinery.

[21] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *arXiv:1606.09375 [cs, stat]*, June 2016. NIPS 2016.

[22] Eytan Domany. Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer Res*, 74(17):4612–4621, September 2014.

[23] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, January 2005.

[24] Liat Ein-Dor, Or Zuk, and Eytan Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *PNAS*, 103(15):5923–5928, April 2006.

[25] Eric Faulkner, Anke-Peggy Holtorf, Surrey Walton, Christine Y. Liu, Hwee Lin, Eman Biltaj, Diana Brixner, Charles Barr, Jennifer Oberg, Gurmit Shandhu, Uwe Siebert, Susan R. Snyder, Simran Tiwana, John Watkins, Maarten J. IJzerman, and Katherine Payne. Being Precise About Precision Medicine: What Should Value Frameworks Incorporate to Address Precision Medicine? A Report of the Personalized Precision Medicine Special Interest Group. *Value in Health*, 23(5):529–539, May 2020.

[26] Simona Maria Fragomeni, Andrew Sciallis, and Jacqueline S Jeruss. Molecular subtypes and local-regional control of breast cancer. *Surgical oncology clinics of North America*, 27(1):95—120, January 2018.

[27] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

[28] Thomas Hofmarcher, Peter Lindgren, Nils Wilking, and Bengt Jönsson. The cost of cancer in Europe 2018. *European Journal of Cancer*, 129:41–49, April 2020.

[29] Marc Johannes, Jan C. Brase, Holger Fröhlich, Stephan Gade, Mathias Gehrmann, Maria Fälth, Holger Sültmann, and Tim Beissbarth. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, 26(17):2136–2144, September 2010.

[30] T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–772, January 2009.

[31] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, and Sven Dähne. Patternnet and patternlrp–improving the interpretability of neural networks. *arXiv preprint arXiv:1705.05598*, 3, 2017.

[32] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, September 2016.

[33] F. Klauschen, K.-R. Müller, A. Binder, M. Bockmayr, M. Hägele, P. Seegerer, S. Wienert, G. Pruneri, S. de Maria, S. Badve, S. Michiels, T.O. Nielsen, S. Adams, P. Savas, F. Symmans, S. Willis, T. Gruosso, M. Park, B. Haibe-Kains, B. Gallas, A.M. Thompson, I. Cree, C. Sotiriou, C. Solinas, M. Preusser, S.M. Hewitt, D. Rimm, G. Viale, S. Loi, S. Loibl, R. Salgado, and C. Denkert. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. *Seminars in Cancer Biology*, 52:151 – 157, 2018. Immuno-oncological biomarkers.

[34] Lev Klebanov and Andrei Yakovlev. Diverse correlation structures in gene expression data and their utility in improving statistical inference. *The Annals of Applied Statistics*, 1(2):538–559, December 2007. Publisher: Institute of Mathematical Statistics.

[35] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards Best Practice in Explaining Neural Network Decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, July 2020. ISSN: 2161-4407.

[36] Paulina Krzyszczyk, Alison Acevedo, Erika J. Davidoff, Lauren M. Timmins, Ileana Marrero-Berrios, Misaal Patel, Corina White, Christopher Lowe, Joseph J. Sherba, Clara Hartmanshenn, Kate M. O'Neill, Max L. Balter, Zachary R. Fritz, Ioannis P. Androulakis, Rene S. Schloss, and Martin L. Yarmush. The growing role of precision and personalized medicine for cancer treatment. *Technology (Singap World Sci)*, 6(3-4):79–100, 2018.

[37] Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat Commun*, 10(1):1096, December 2019. arXiv: 1902.10178.

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. Number: 7553 Publisher: Nature Publishing Group.

[41] Hae Woo Lee, Carl Lawton, Young Jeong Na, and Seongkyu Yoon. Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery. *Statistical Applications in Genetics and Molecular Biology*, 12(2):207–223, 2013.

[42] Ron Levie, Federico Monti, Xavier Bresson, and Michael M. Bronstein. CayleyNets: Graph Convolutional Neural Networks With Complex Rational Spectral Filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, January 2019. Conference Name: IEEE Transactions on Signal Processing.

[43] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, November 2017.

[44] Boyu Lyu and Anamul Haque. Deep Learning Based Tumor Type Classification Using Gene Expression Data. *bioRxiv*, page 364323, July 2018.

[45] Kalifa Manjang, Shailesh Tripathi, Olli Yli-Harja, Matthias Dehmer, Galina Glazko, and Frank Emmert-Streib. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Sci Rep*, 11(1):156, January 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Breast cancer;Cancer;Computational biology and bioinformatics;Mathematics and computing;Microarrays;Statistical methods;Systems biology Subject_term_id: breast-cancer;cancer;computational-biology-and-bioinformatics;mathematics-and-computing;microarrays;statistical-methods;systems-biology.

[46] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Brief Bioinform*, 18(5):851–869, September 2017.

[47] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise Relevance Propagation: An Overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 193–209. Springer International Publishing, Cham, 2019.

[48] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017.

[49] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018.

[50] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[51] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13(5):44, April 2020.

[52] Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

[53] Parliament and C. of the European Union. General data protection regulation. `https://gdpr-info.eu/`, 2016.

[54] Júlia Perera-Bel, Barbara Hutter, Christoph Heining, Annalen Bleckmann, Martina Fröhlich, Stefan Fröhling, Hanno Glimm, Benedikt Brors, and Tim Beißbarth. From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. *Genome Medicine*, 10(1):18, 2018.

[55] Júlia Perera-Bel, Andreas Leha, and Tim Beißbarth. *Bioinformatic Methods and Resources for Biomarker Discovery, Validation, Development, and Integration*, pages 149–164. Springer International Publishing, Cham, 2019.

[56] Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability Methods for Graph Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10764–10773, June 2019. ISSN: 2575-7075.

[57] Christine Porzelius, Marc Johannes, Harald Binder, and Tim Beißbarth. Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients. *Biometrical Journal*, 53(2):190–201, 2011.

[58] Ricardo Ramirez, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. Classification of Cancer Types Using Graph Convolutional Neural Networks. *Front. Phys.*, 8, 2020. Publisher: Frontiers.

[59] European Commission — Press Release. Europe's Beating Cancer Plan: A new EU approach to prevention, treatment and care. `https://ec.europa.eu/commission/presscorner/detail/en/ip_21_342`, 2021. Accessed: 2021-11-21.

[60] Brooke Rhead, Xiaorong Shao, Hong Quach, Poonam Ghai, Lisa F. Barcellos, and Anne M. Bowcock. Global expression and CpG methylation analysis of primary endothelial cells before and after TNFa stimulation reveals gene modules enriched in inflammatory and infectious diseases and associated DMRs. *PLOS ONE*, 15(3):e0230884, March 2020. Publisher: Public Library of Science.

[61] Sungmin Rhee, Seokjun Seo, and Sun Kim. Hybrid Approach of Relation Network and Localized Graph Convolutional Filtering for Breast Cancer Subtype Classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3527–3534, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization.

[62] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, August 2016.

[63] Ashley G. Rivenbark, Siobhan M. O'Connor, and William B. Coleman. Molecular and Cellular Heterogeneity in Breast Cancer. *Am J Pathol*, 183(4):1113–1124, October 2013.

[64] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, November 2017. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[65] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans Neural Netw*, 20(1):61–80, January 2009.

[66] Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller, and Grégoire Montavon. XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv:2006.03589 [cs, stat]*, June 2020.

[67] Robert Schwarzenberg, Marc Hübner, David Harbecke, Christoph Alt, and Leonhard Hennig. Layerwise Relevance Visualization in Convolutional Text Graph Classifiers. *arXiv:1909.10911 [cs]*, September 2019. arXiv: 1909.10911.

[68] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

[69] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013.

[70] Therese Sørlie. Molecular classification of breast tumors: Toward improved diagnostics and treatments. In *Target Discovery and Validation Reviews and Protocols*, pages 91–114. Humana Press, 2007.

[71] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2015.

[72] Kun Sun, Jiguang Wang, Huating Wang, and Hao Sun. GeneCT: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies. *Bioinformatics*, 34(23):4129–4130, December 2018.

[73] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017.

[74] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos,

Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 43(Database issue):D447–452, January 2015.

[75] Erico Tjoa and Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, November 2021. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[76] Khoa A. Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D. Williams, John V. Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1):152, September 2021.

[77] David Venet, Jacques E. Dumont, and Vincent Detours. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLOS Computational Biology*, 7(10):e1002240, October 2011. Publisher: Public Library of Science.

[78] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, December 2020.

[79] Sarah C. P. Williams. News Feature: Capturing cancer's complexity. *Proc Natl Acad Sci U S A*, 112(15):4509–4511, April 2015.

[80] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, January 2021. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[81] Shangsheng Xie and Mingming Lu. Interpreting and Understanding Graph Convolutional Neural Network using Gradient-based Attribution Method. *arXiv:1903.03768 [cs]*, April 2019.

[82] Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching. Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 152–162, June 2018.

[83] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. *Adv Neural Inf Process Syst*, 32:9240–9251, December 2019.

[84] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham, 2014. Springer International Publishing.

[85] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 543–559, Cham, 2016. Springer International Publishing.

[86] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An End-to-End Deep Learning Architecture for Graph Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. Number: 1.

Hryhorii Chereda

MACHINE LEARNING ON GRAPHS | HIGH-DIMENSIONAL DATA ANALYSIS

Location: Göttingen, Germany

hryhorii.chereda@bioinf.med.uni-goettingen.de | GregoryDS | hryhorii-chereda

Summary

Researcher in interdisciplinary area intersecting machine learning, systems biology, and precision medicine. Experienced in applying and developing machine learning approaches on high-dimensional data. Adapted and implemented an explanation method to interpret classification decisions of a Graph convolutional neural network performing graph signal classification. Interested in geometric deep learning, interpretability and robustness of machine learning approaches.

Experience

Department of Medical Bioinformatics, University Medical Center Göttingen — *Göttingen, Germany*

RESEARCH FELLOW — *Aug. 2017 - present*

- Developed and implemented an explanation method for a Graph convolutional neural network
- Co-developed a machine learning pipeline generating patient-specific subnetworks for personalized treatment decisions
- Communicated and presented machine learning approaches to biologists and clinicians
- Supervision of students: internship, project work, master thesis
- Lecturing and teaching method courses

Cardio 4u, a student startup in the healthcare domain developing software for "smart garments". — *Kyiv, Ukraine*

SIGNAL PROCESSING ENGINEER — *Dec. 2014 - May 2017*

- Digital preprocessing of electrocardiograms (ECGs) and photoplethysmograms (PPGs)
- Adjusted existing ECG delineation tools to be applicable to ECGs with different sampling rates

Education

Georg August University of Göttingen | International Max Planck Research School for Genome Science (IMPRS-GS) — *Göttingen, Germany*

PH.D. STUDENT AT THE DEPARTMENT OF MEDICAL BIOINFORMATICS, UNIVERSITY MEDICAL CENTER GÖTTINGEN — *Oct. 2017 - present*

- **Thesis:** Explaining decisions of graph convolutional neural networks for analyses of molecular subnetworks in cancer
- **Supervisor:** Prof. Dr. Tim Beißbarth.

National Technical University of Ukraine "Kyiv Polytechnic Institute" — *Kyiv, Ukraine*

M.SC. WITH HONOURS IN SYSTEM ANALYSIS AND CONTROL — *Sep. 2015 - Jun. 2017*

- **Thesis:** Adaptive Filtration of a Time Series with Variable Sampling
- **Subjects:** System Analysis, Control Theory, Mathematical Programming, Intelligent Systems for Decision-Making Support

Nicolaus Copernicus University — *Toruń, Poland*

STUDIES WITHIN ERASMUS+ EXCHANGE PROGRAM, COMPUTER SCIENCE — *Feb. 2016 - Jun. 2016*

- **Subjects:** Python, Massive Data Mining, Algorithms for Scalable Data Processing, Signal and Data Analysis

National Technical University of Ukraine "Kyiv Polytechnic Institute" — *Kyiv, Ukraine*

B.SC. IN SYSTEM ANALYSIS — *Sep. 2011 - Jun. 2015*

- **Thesis:** Automatic annotation of digitalised ECG signals with wavelets
- **Subjects:** Probability and Statistics, Optimization Methods, Modelling of Complex Systems, Fundamentals of System Analysis

Skills

Programming languages	*High:* Python; *Intermediate:* R, Matlab
Machine Learning	*High:* Tensorflow (1.x and 2.x), Keras; *Intermediate:* Scikit-learn
Data Processing	*High:* Pandas, Numpy; *Intermediate:* SciPy; *Basic:* OpenCV, PIL
Visualisation	Matplotlib, Seaborn
IDE	PyCharm, Jupyter Notebook
Others	*Intermediate:* Latex, Docker, Linux; *Basic:* Git
Languages	Ukrainian, Russian, English (C1), German (A2), Polish (A2)

Honors & Awards

2021 **Göttingen Very Important Publication Award,** GöVIP-25 - Kategorie Klinische Forschung — *Göttingen, Germany*

2016 **Metro Challenge Prize,** Food{hacks} hackathon — *Berlin, Germany*

2015 **Finalist as a member of Cardio 4u team,** IV Festival of innovative projects "Sikorsky Challenge" — *Kyiv, Ukraine*

Publications

Stability of feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation *Link*

Chereda H, Leha A, Beißbarth T (2021) Submitted. bioRxiv: 2021.12.26.474194.

Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks *Link*

Alachram H, Chereda H, Beißbarth T, Wingender E, Stegmaier P (2021) PLoS ONE 16(10): e0258623.

Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer *Link*

Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, Auer F, Kramer F, Leha A, Beißbarth T (2021) Genome medicine, 13(1), 42.

Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer *Link*

Chereda H, Bleckmann A, Kramer F, Leha A, Beißbarth T (2019) Studies in Health Technology and Informatics 267:181-186.

Sampling Rate Independent Filtration Approach for Automatic ECG Delineation *Link*

Chereda H, Nikolaiev S, Tymoshenko Y (2016) arXiv: 1611.08537.

Selected courses

Machine Learning Summer School (MLSS) 2018 — *Madrid, Spain*

Participant — *Sep. 2018*

Data Science Summer School 2018 — *Göttingen, Germany*

Participant — *Aug. 2018*

Machine Learning Course from Andrew Ng — *Coursera*

Successfully completed at Coursera MOOC platform — *Jul. 2014 - Sep. 2014*

Extracurricular Activity

2nd IMPRS-GS doctoral retreat — *Volpriehausen, Germany*

Organizer — *Jun. 2019*

- Communication with invited speakers
- Organisation of social activities