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:

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

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:

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

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:

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

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.

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

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

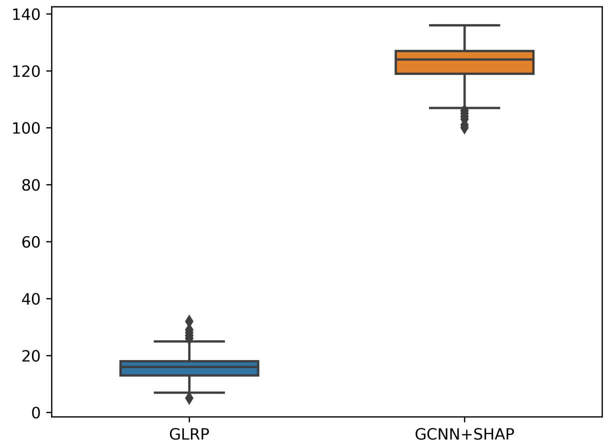

Fig. 1. The distribution of the number of connected components in patients' subnetworks. The left boxplot corresponds to the subnetworks obtained by GLRP while the right one corresponds to the subnetworks obtained by GCNN+SHAP.

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.