Multi-omics Classification on Kidney Samples Exploiting Uncertainty-Aware Models

Marta Lovino[1(✉)], Gianpaolo Bontempo[1], Giansalvo Cirrincione[2,3], and Elisa Ficarra[1]

[1] Department of Control and Computer Engineering, Politecnico di Torino, 10129, Turin, Italy
{marta.lovino,elisa.ficarra}@polito.it,
gianpaolo.bontempo@studenti.polito.it
[2] LTI Lab, University of Picardie Jules Verne, 80025, Amiens, France
exin@u-picardie.fr
[3] University of South Pacific, Suva, Fiji

Abstract. Due to the huge amount of available omic data, classifying samples according to various omics is a complex process. One of the most common approaches consists of creating a classifier for each omic and subsequently making a consensus among the classifiers that assigns to each sample the most voted class among the outputs on the individual omics.

However, this approach does not consider the confidence in the prediction ignoring that a biological information coming from a certain omic may be more reliable than others. Therefore, it is here proposed a method consisting of a tree-based multi-layer perceptron (MLP), which estimates the class-membership probabilities for classification. In this way, it is not only possible to give relevance to all the omics, but also to label as *Unknown* those samples for which the classifier is uncertain in its prediction. The method was applied to a dataset composed of 909 kidney cancer samples for which these three omics were available: gene expression (mRNA), microRNA expression (miRNA) and methylation profiles (meth) data. The method is valid also for other tissues and on other omics (e.g. proteomics, copy number alterations data, single nucleotide polymorphism data). The accuracy and weighted average f1-score of the model are both higher than 95%. This tool can therefore be particularly useful in clinical practice, allowing physicians to focus on the most interesting and challenging samples.

Data availability: the code is freely accessible at https://github.com/Bontempogianpaolo1/Consunsus-on-multi-omics, while mRNA, miRNA and meth data can be obtained from the GDC database [2] or upon request to the authors.

Keywords: Bayesian neural networks · Gene expression · mRNA · miRNA · Methylation · Multi-Layer Perceptron (MLP) · Multi-omics · Multi-omics classification

D.-S. Huang and K.-H. Jo (Eds.): ICIC 2020, LNCS 12464, pp. 32–42, 2020.
https://doi.org/10.1007/978-3-030-60802-6_4

1 Introduction

In recent years, the reduction of costs for the sequencing of biological molecules including DNA, RNA and proteins has allowed the widespread of huge amounts of data both in the form of large structured databases and in the form of repositories specially created for the study of particular pathologies [1–3].

In this context, various omic data can be taken into account for the study and analysis of samples, either tumor or healthy: gene expression data (mRNA), microRNA expression data (miRNA), methylation data (meth), copy number alterations data (CNA), single nucleotide polymorphism data (SNV), proteomics and phosphoproteomics data.

Two large strands are typically available in multi-omics analysis: first the subdivision of the samples into its own classes [8–11] and second, the identification of specific pathways and gene patterns in the dataset [12, 13].

This work is focus exclusively on the first strand; in particular, some methods are presented for the classification of kidney cancer samples by simultaneously exploiting the information from the mRNA, miRNA, and methylation (meth) data. Although the work is focused on mRNA, miRNA and meth omics, it must be noticed that the same algorithms can be applied to a greater number of omics or other omics in place of them.

In the multi-omics classification approach, a crucial step is represented by the algorithm by which to integrate the classification results from each omic. One of the standard approaches is to make a consensus among the various omics, such that the multi-omic class is the most voted class among the outputs on the individual omics [14, 15]. However, this approach has two main limitations. At first, it is difficult to attribute to the multi-omic class in the case in which all the outputs of the individual omics are completely disjoint or more than one class is equally voted among all the omics. Secondly, each omic carries characteristics that may not be present in the other omics. For classification purposes, therefore, the contribution of a single omic should be considered according to the certainty in its classification.

This work proposes the use of a learning method that for each omic returns not only the corresponding class, but also its membership probability to that class, overcoming the main problem of standard consensus when the same sample is assigned to different classes across the omics or when there is no clear class prevalence.

In addition, the use of the class-membership probability allows to filter samples according to the class probability and consequently postpone for further analyses those samples on which there is not enough certainty in the classification across all the omics. This approach is particularly useful in creating automatic tools that, integrating different omic information, may favor clinical practice, by proposing a classification label when all the omics are enough certain in their classification and, an *Unknown* label when discrepancies are found across the omics. In this way, physicians can have a quick look at well-defined samples and focus more on the most interesting and challenging cases where human control is crucial.

2 Biological Data

Although the proposed method can be applied to any tissue and pathology, this work is deal with the study of kidney tumor samples freely available in the Genomic Data Commons (GDC) database [2]. The samples used in this study belong to three main kindney tumor subtypes: kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRCH) and kidney chromophobe (KICH). In addition, a reduced number of healthy samples is available both for KIRP and KICH subtypes (usually these tissues are healthy areas surrounding a KIRP or KICH tumor).

For KIRP, KIRCH and KICH subtypes, only samples samples available are selected for mRNA, miRNA and meth data, obtaining a final dataset of 909 samples.

The mRNA, miRNA and meth data are tabular data commonly represented as matrices, where the value in position (i, j) represents the amount of a specific biological product or the intensity of a phenomenon (mRNA, miRNA and meth respectively) in a specific sample. The mRNA, miRNA and meth matrices carry different biological information.

The mRNA expression value is strictly related to the amount of its protein (higher is the number, higher the amount of the protein) which regulates a specific pathway in the cell life cycle.

The miRNA expression value indicates the amount of a specific miRNA, a small noncoding RNA molecule which intervenes in the post-transcriptional process, regulating the amount of produced final protein.

Methylation value refers to the methylation beta value, an estimate of the methylation level computed as the ratio of intensities between methylated and unmethylated alleles. The biological effect of the methylation consists of the change of the activity of a DNA segment without changing its sequence (when methylation occurs, it reduces the DNA transcription, thus consequently reducing the amount of protein).

It must be noticed that many biological molecules act together in order to regulate the cell activity and that changes in the values of one or more omics can be correlated to a specific pathology or a tumor subtype.

2.1 Data Preprocessing

After downloading and selecting samples for which both mRNA, miRNA and meth data are available, the following preprocessing is performed:

- **mRNA**: 5000 features × 909 samples. Raw count data have originally about 60000 mRNA genes and have been normalized using the Variance Stabilizing Trasformation (VST) [16]. Then all not protein coding genes have been discarded reaching about 20000 mRNA genes and z-score transformation has been performed. In the end, the top 5000 mRNA genes with the highest standard deviation are selected.
- **miRNA**: 1200 features × 909 samples. The miRNA data have about 2000 miRNAs and have been normalized using deseq [17]. Then pseudo-counts have been computed as $log_2(count\_value + 1)$. In the end, z-score transformation has been performed and the top 1200 miRNAs with the highest standard deviation have been selected.

- **meth:** 5000 features × 909 samples. Among the 27000 features in methylation array data obtained with Illumina Human Methylation 27 platform, the top 5000 with the highest standard deviation are selected. Since original data are intrinsically normalized, no further normalization is required.

The 909 samples belong to 5 classes: **tumor KIRCH**: 509, **tumor KIRP**: 288, **tumor KICH**: 65, **healthy KIRCH**: 24, **healthy KIRP**: 23.

They have been further divided into training set (75% of the samples) and test set (25% of the samples) such that the latter includes the same proportion of samples belonging to the different classes.

In order to test the model on samples that do not belong to the kidney classes, 37 stomach samples have been obtained from GDC [2], by applying the same preprocessing steps described in 2.1. This dataset is used only as test test, without re-training the kidney model to evaluate the ability of the probabilistic approaches, such as the tree MLP classifier, to recognize unseen classes.

3 Method

An extension of the multi-layer perceptron (MLP) combining several MLPs in a tree architecture (tree MLP) is here proposed. Such an architecture has been designed to face with the classification of samples where no clear class prevalence was obtained through the consensus of the various omic-based classifiers. Moreover, it aims at identifying and filtering out samples uncertainly classified.

Since a MLP equipped with a cross-entropy loss function, with associated either logistic sigmoid (two class problem) or softmax (multiclass problem), outputs the class-membership posterior probabilities of the inputs [21], the proposed tree MLP classifier is therefore able to return the class label and the associated probability of the sample belonging to a class.

As it can be seen in Fig. 1, a tree-like architecture was created with MLP models as nodes and trained separately on subsets of the training set. For this specific problem, there are a root node (trained to recognize healthy from tumor samples) and two leaf nodes. The former is trained on healthy samples and classifies them into KIRP and KIRCH healthy tissues. The latter is trained on tumor samples and classifies them into KIRP, KIRCH, and KICH tumors.

Therefore, given a new sample S, it will be classified by the root MLP as healthy or tumor (y_{root}) with a class-membership probability P_r. After selecting the leaf node corresponding to y_{root}, it returns the subclass label y_{leaf} (tumor_KIRP, tumor_KIRCH, and tumor_KICH for tumor leaf MLP; healthy_KIRP and healthy_KIRCH for normal leaf MLP) with its class-membership probability P_{leaf}. The final class y_{pred} is equal to y_{leaf}.

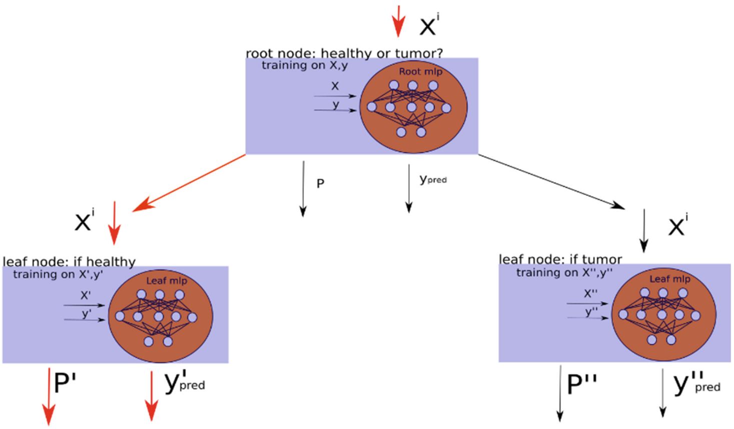


Fig. 1. Proposed tree MLP model: i) each node is trained on three different subsets of the original dataset. (X, y) aims to distinguish between healthy and tumor samples, (X′, y′) between subtypes of healthy samples and (X″, y″) between subtypes of tumor samples; ii) the output of each node consists of the predicted label y_{pred} and the class-membership probability P.

Once the classification on each individual omic is performed, the final consensus is built taking into account the final probabilities on each omic. Given:

- n: the number of the omics,
- m: the number of the classes,
- th: threshold on the omics, in order to filter predictions with low probabilities across all the omics,
- tr: threshold on the classes, in order to select only samples with a not uniform distribution of the class-membership probabilities across the m classes,
- P_{ij}: the class membership probability for class i and omic j,
- $S_i = sum_{j=1}^{n} P_{ij}$: the sum of the probabilities on all the omics for a single class,
- $S_a = sum_{i=1}^{m} S_i$: the sum of the probabilities on all the omics and all the samples,
- $S_m = S_i/n$: the mean of the probabilities on all the omics for a single class.

The consensus for a sample is built according to the next formula:

$$\begin{cases} Unknown,\ if\ \max_{\mathrm{i}}(S_m) < th or \max_{\mathrm{i}}(S_i/S_a) < tr \\ arg \max_{\mathrm{i}}(S_i) otherwise \end{cases}$$

In that way, a sample with a low mean probability across all the omics is labelled as *Unknown*.

In addition, when a sample receives similar S_i values for more than one class, the model is uncertain in its prediction. Therefore, a tr threshold is set in order to select only samples with a not uniform distribution of the class-membership probabilities across the m classes.

This final consensus can be applied using any number of omics as long as each omic represents different points of view of the same sample. Obviously, the larger the number of the omics, the more reliable the consensus prediction can be.

Very small architectures in the implemented neural networks were used (e.g. MLP with a single hidden layer with 20 neurons and a single activation layer) since the chosen number of main PCA components is low. This structure is the same for all the nodes of the tree MLP. Many hyper-parameters configurations have been considered. In the end, gradient descent with back propagation and the cross entropy as loss function were used. The optimizer was Adam.

In order to have a baseline for the results, a support vector machine (SVM) and a random forest (RF) classifier have been applied to the training set (with hyper-parameters optimization). Unless these models do not output a class-membership probability, they can provide valuable insights onto the data. Since they are unable to estimate the certainty of their prediction, the implementation of the consensus has been slightly modified. The final consensus for SVM and RF classifiers is given by the majority voting between the different omics.

On the other hand, to compare the tree MLP architecture with other methods that return a class-membership probability, a standard MLP classifier and a Bayesian neural network (BNN) were built.

The BNN model has the same structure as the MLP; however, it works in a completely different way. Indeed, as the loss is modified with a Bayesian regularization term, its weights are no longer deterministic like a standard MLP, but probabilistic, and each neuron learns to follow probabilistic distributions.

Therefore, it is possible to infer the level of uncertainty of the class-membership probability estimation of the input, which represents how much a sample belongs to a given class. The model is applied to the sample *n* times and the median value among all the output probabilities is selected as the final probability. For instance, if the median value is 0.95, it means that the output is highly stable and its classification uncertainty is very low.

All models have been tested both on the test set, consisting of kidney samples belonging to the 5 classes of the training set, and on the 37 stomach cancer samples.

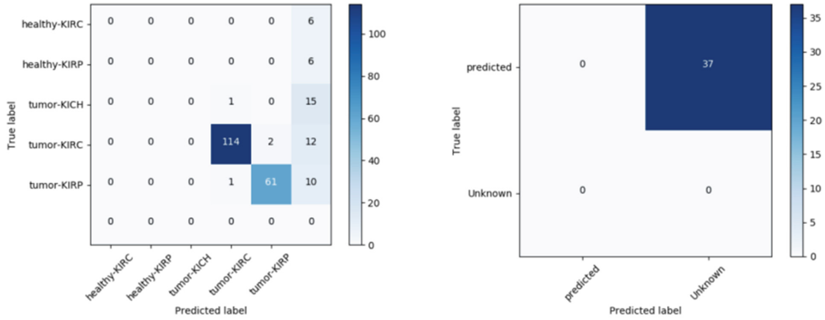
All models were implemented in Pytorch framework [22]. In addition, the Pyro library [23] was used for the BNN to transform the parameters into random variables and to run stochastic variational inference.

4 Results

In this section, the results related to the proposed method are presented, as well as those of SVM, RF, MLP and BNN models. All performance metrics are obtained setting $th = 0.9$ and $tr = 0.25$. For tree MPL, MLP and BNN classifiers, all metrics are computed discarding *Unknown* samples.

In detail, concerning the tree MLP classifier, it is reported the confusion matrix with the classification results as well as the final consensus both on kidney test set (Fig. 2 (a)) and on 37 stomach samples (Fig. 2 (b)). Globally, the tree MLP method reached the 98% of accuracy and 97% of weighted average f1-score (Table 1). The metrics were computed disregarding *Unknown* samples as they had not been assigned to any classes. The tree MLP classifier selected as *Unknown* the 21, 49% of kidney test set samples and misclassified the 2% of the not *Unknown* samples.

Concerning the 37 stomach samples, all of them were correctly labelled as *Unknown* (Table 1).



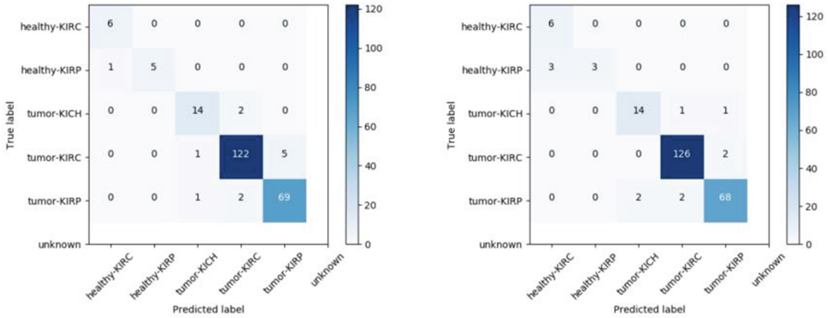
(a) consensus on kidney test set

(b) consensus on stomach samples

Fig. 2. Confusion matrices for tree MLP classifier on (a) kidney test set and (b) stomach set.

Consensus confusion matrices for SVM and RF classifiers on kidney samples are reported in Fig. 3 (a-b). Both SVM and RF reached the 95% of accuracy and weighted average f1-score.

It should be noticed that the consensus creation for SVM and RF is different from that used in tree MLP, MLP and BNN models, since SVM and RF does not output the class-membership probabilities. Therefore, for SVM and RF classifiers the consensus is based on the majority voting on the three omics without considering class probabilities. As a consequence, the results for SVM and RF on stomach samples are not reported, since all the 37 stomach samples will be forced to one of the five kidney classes.

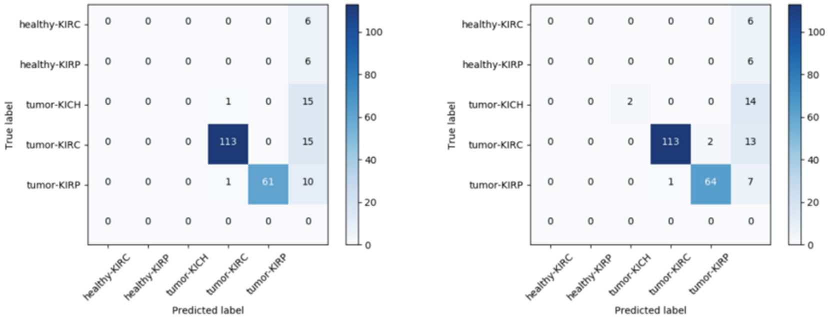


(a) consensus SVM on kidney test set

(b) consensus RF on kidney test set

Fig. 3. Consensus confusion matrices on kidney test set on (a) SVM, (b) RF classifiers.

In addition, the performances with standard MLP model on the kidney test set were evaluated. After proper hyper-parameter tuning, the results obtained for the consensus are reported in Fig. 4 (a). Globally, it reached the 99% of accuracy and 99% of weighted average f1-score. All metrics were computed disregarding *Unknown* samples. Standard MLP model classified as *Unknown* the 22, 80% of kidney samples and misclassified the 1, 10% of the not *Unknown* samples. Concerning the 37 stomach samples, all of them were correctly labelled as *Unknown* (see in Fig. 5 (a)).

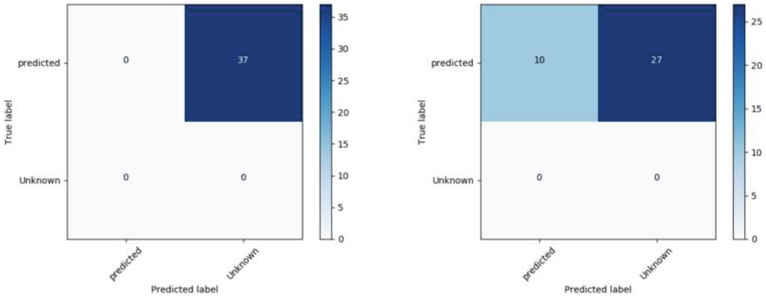


(a) consensus. MLP on kidney test set

(b) consensus. BNN on kidney test set

Fig. 4. Consensus confusion matrices on kidney test set on (a) MLP, (b) BNN classifiers.

Consensus performances obtained with the BNN model are reported in Fig. 4 (b). The BNN model classified as *Unknown* the 20, 17% of kidney samples and misclassifies the 2% of the not *Unknown* samples. In addition, it achieved the 98% of accuracy and the 98% of weighted average f1-score. However, concerning the 37 stomach samples, only the 73% of them were correctly labelled as *Unknown* (see in Fig. 5 (b)).



(a) MLP consensus on stomach samples

(b) BNN consensus on stomach samples

Fig. 5. Confusion matrices for (a) MLP classifier and (b) BNN on 37 stomach samples.

In the end, the main results achieved for all the classifiers are reported in Table 1.

Table 1. Comparison between all the methods on the kidney test set. All the reported metrics are computed with weighted average only on not *Unknown* samples. The support metric, or the number of not *Unknown* samples, is the value on which the other metrics are based.

	Precision	Recall	F1-score	Accuracy	Support	*Unknown*
RF	95%	95%	95%	95%	228	–
SVM	95%	95%	95%	95%	228	–
tree MLP	97%	98%	97%	98%	179	21.4%
MLP	98%	99%	99%	99%	176	22.8%
BNN	98%	98%	98%	98%	182	20.17%

5 Discussion

As reported above, all the classifiers perform generally well. In fact, the accuracy and weighted average f1-score is always higher or equal to 95% (see Table 1).

In detail, SVM and RF models reached high classification rates on the kidney test set with accuracy and weighted average f1-score equal to 95%. Since these two models always force a prediction, they prevent labelling samples as *Unknown*. Although it could seem a minor issue, in real clinical practice, it is suitable to receive an *Unknown* label when the classifier is uncertain in its prediction.

Compared to the majority voting consensus used for SVM and RF classifiers, the proposed method analyses the probability values obtained on each omic and provides an integrated assessment of all the probability values. Considering tree MLP, standard MLP and BNN classifiers, they labelled as *Unknown* a similar percentage of kidney test set samples (21.49%, 22.80%, 20.17%, respectively) and had a similar weighted average f1-score (97%, 99%, 98%, respectively).

It should be noticed that, considering a tissue which the classifiers were not trained on (stomach samples), tree MLP and MLP classifiers labelled all the 37 stomach samples as *Unknown*, against the 73% of the BNN classifier.

Unlike the standard MLP, in a tree MLP model it is possible to retrain one of its nodes separately. This aspect is crucial in the biological domain. In fact, new molecular subtypes of the same tumor are continually redefined. In this case, the tree MLP model can be updated on the new classes retraining only the involved nodes and not the entire classifier, avoiding spare of time. In the MLP architectures, the threshold represents a cut with respect to the class-membership point-wise posterior probabilities of the inputs. On the other hand, in the BNN architecture, all the output probabilities estimated for each sampling are summarized by a median value. This scalar can be used for recognition thresholding. However, even if both techniques look identical, the probability value on which they act is completely different in nature. Therefore, a direct comparison between the two MLP-based methods and the BNN architecture is not completely possible. In the presented results, same *th* and *tr* values were applied to the three probabilistic approaches for the sake of comparison. This choice probably leaded the BNN to be less selective in the classification of stomach samples.

In addition, it can be noticed that a key role in the results is played by the criteria that has been used to obtain the final consensus across all the omics. In fact, the proposed consensus algorithm labels as *Unknown* samples with a low mean probability across all the omics or with similar sum probabilities on the classes (S_i). In that way, it prevents an unsafe labelling.

6 Conclusions

In the multi-omics classification task, the main limitation of the standard consensus is given by the absence of a measure to check the relevance of each individual omic in the classification.

Here, to overcome this problem, a tree MLP architecture is proposed to take into account the reliability of the classification on the individual omics exploiting uncertainty-aware models. Compared to the standard MLP and BNN architectures to classify kidney test set, the tree MLP represents a good compromise in terms of percentage of samples labelled as *Unknown*, and misclassification rate on the remaining samples (21, 49% and 2% respectively). In addition, the tree MLP model significantly outperforms the BNN model when predicting samples coming from a tissue on which the model has not been trained. This aspect is particularly relevant in clinical practice, since usually it is preferable to receive an *Unknown* label instead of a wrong prediction. Moreover, compared to a standard MLP, the tree structure is particular effective in applications where there is an ever-evolving knowledge, such as genetic complex diseases studies, preventing the classifier to be trained from scratch.

References

1. Weinstein, J.N., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10), 1113 (2013)
2. Grossman, R.L., et al.: Toward a shared vision for cancer genomic data. N. Engl. J. Med. **375**(12), 1109–1112 (2016)
3. Leinonen, R., Sugawara, H., Shumway, M.: International nucleotide sequence database collaboration: the sequence read archive. Nucleic Acids Res. **39**((suppl_1)), D19–D21 (2010)
4. Pochet, N., De Smet, F., Suykens, J.A., De Moor, B.L.: Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics **20**(17), 3185–3195 (2004)
5. Lee, G., Rodriguez, C., Madabhushi, A.: An empirical comparison of dimensionality reduction methods for classifying gene and protein expression datasets. In: Măndoiu, I., Zelikovsky, A. (eds.) ISBRA 2007. LNCS, vol. 4463, pp. 170–181. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72031-7_16
6. Kim, P.M., Tidor, B.: Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. **13**(7), 1706–1718 (2003)

7. Lu, M., Zhan, X.: The crucial role of multiomic approach in cancer research and clinically relevant outcomes. EPMA J. **9**(1), 77–102 (2018)
8. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al.: Similarity network fusion for aggregating data types on a genomic scale. Nat. Meth. **11**(3), 333 (2014)
9. Argelaguet, R., et al.: Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. **14**(6), e8124 (2018)
10. Robles, A.I., Arai, E., Mathé, E.A., Okayama, H., Schetter, A.J., Brown, D., et al.: An integrated prognostic classifier for stage I lung adenocarcinoma based on mRNA, microRNA, and DNA methylation biomarkers. J. Thorac. Oncol. **10**(7), 1037–1048 (2015)
11. Tang, W., Wan, S., Yang, Z., Teschendorff, A.E., Zou, Q.: Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics **34**(3), 398–406 (2018)
12. Cantini, L., Medico, E., Fortunato, S., Caselle, M.: Detection of gene communities in multi-networks reveals cancer drivers. Sci. Rep. **5**, 17386 (2015)
13. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nat. Rev. Genet. **12**(1), 56–68 (2011)
14. Fuchs, M., Beißbarth, T., Wingender, E., Jung, K.: Connecting high-dimensional mRNA and miRNA expression data for binary medical classification problems. Comput. Meth. Programs Biomed. **111**(3), 592–601 (2013)
15. Mallik, S., Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: Integrated analysis of gene expression and genome-wide DNA methylation for tumor prediction: an association rule mining-based approach. In: 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), (pp. 120–127). IEEE April 2013
16. Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M.: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics **18**((suppl_1)), S96–S104 (2002)
17. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014)
18. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1-3), 37–52 (1987)
19. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
20. Ramchoun, H., Idrissi, M.A.J., Ghanou, Y., Ettaouil, M.: Multilayer perceptron: architecture optimization and training. IJIMAI **4**(1), 26–30 (2016)
21. Christopher, M.: Bishop.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
22. Paszke, A., et al. Automatic differentiation in pytorch (2017)
23. Bingham, E., et al.: Pyro: deep universal probabilistic programming. J. Mach. Learn. Res. **20** (1), 973–978 (2019)