

A Weakly Supervised Clustering Method for Cancer Subgroup Identification

Duygu Ozcelik and Oznur Tastan

Abstract—Identifying subgroups of cancer patients is important as it opens up possibilities for targeted therapeutics. A widely applied approach is to group patients with unsupervised clustering techniques based on molecular data of tumor samples. The patient clusters are found to be of interest if they can be associated with a clinical outcome variable such as the survival of patients. However, these clinical variables of interest do not participate in the clustering decisions. We propose an approach, WSURFC (Weakly Supervised Random Forest Clustering), where the clustering process is weakly supervised with a clinical variable of interest. The supervision step is handled by learning a similarity metric with features that are selected to predict this clinical variable. More specifically, WSURFC involves a random forest classifier-training step to predict the clinical variable, in this case, the survival class. Subsequently, the internal nodes are used to derive a random forest similarity metric among the pairs of samples. In this way, the clustering step utilizes the nonlinear subspace of the original features learned in the classification step. We first demonstrate WSURFC on handwritten digit datasets, where WSURFC can capture salient structural similarities of digit pairs. Next, we apply WSURFC to find breast cancer subtypes using mRNA, protein, and microRNA expressions as features. Our results on breast cancer show that WSURFC could identify interesting patient subgroups more effectively than the widely adopted methods.

Index Terms—Clustering, cancer subtype identification, patient subgroup identification.

I. INTRODUCTION


A major hurdle in devising more effective cancer therapies is the accurate stratification of patients into subgroups [1]. This stems from the fact that most cancer types are heterogeneous at the molecular level; seemingly similar tumors that are classified into the same cancer type may have distinct molecular profiles resulting in distinct clinical trajectories [2]. The availability of large sets of patient molecular data has opened up opportunities to redefine the subtypes of cancers [3].

The widely adopted approach for grouping cancer patients using patient molecular data is to apply unsupervised clustering techniques such as k-means, hierarchical clustering or non-negative matrix factorization (NMF) on the genomic data

Duygu Ozcelik conducted this research at the Department of Computer Engineering, Bilkent University, Ankara, TURKEY. e-mail: duyguozcelik89@gmail.com

 <https://orcid.org/0000-0001-8980-6200>

Oznur Tastan is with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, TURKEY. Coresponding author e-mail: otastan@sabanciuniv.edu

 <https://orcid.org/0000-0001-7058-5372>

of patients, sometimes combined with consensus clustering [4]–[11]. There are also more advanced methods that use probabilistic modeling of multi-omics data to project the data to a lower dimension (reviewed and benchmarked in [12] and [13]) such as PARADIGM [14], iCluster [15], multi-omics perturbation based approaches such as PINSplus [16] or methods that integrates biological pathways and multi-omics data, such as PAMOGK [17]. All of these approaches take an unsupervised approach, and the patient clusters are deemed interesting if they are found to be associated with a clinical variable of interest, such as the survival rate of the patients. Therefore, the clinical variable of interest does not participate in the clustering decisions to guide the clusters.

One body of method that incorporates clinical variable of interest makes use of the association of the features with the clinical variable of interest in the feature selection step. Bair et al. [18] test the null hypothesis of no association between the feature and the outcome variable and applies feature selection based on the test-statistic. Next, with the remaining feature, they perform clustering using a conventional clustering algorithm such as k-means or hierarchical clustering. Koestler et al. [19] propose a method called “semi-supervised recursively partitioned mixture models” with the same rationale. It also calculates a score for each feature and measures the association between the features and the outcome variable of interest. Next, clustering is carried out using only the features with the largest scores. The difference between Bair et al. and Koestler et al. is that the latter applies a recursively partitioned mixture models algorithm [20] instead of a standard clustering algorithm. Although these approaches are simple, they are limited in the sense that they only take account of the univariate relationship of features with the clinical variable.

We utilize weakly supervised learning and propose a new approach, WSURFC, where the clinical variable of interest guides the clustering process. The weak supervision is achieved by learning a similarity metric by predicting a clinical variable, in our study, the survival class as long and short survivors. With guidance from the clinical label, WSURFC learns a non-linear similarity metric among patients, which is then used to cluster the patients. We first demonstrate the methodology on an unrelated, but easier-to-expect problem, digit classification. Then we apply it to breast cancer subgroup identification.

II. METHODS

This section presents a detailed description of our proposed method and the datasets used. Let \mathcal{D} represent the set of n

Algorithm 1 WSURFC: Weakly Supervised Random Forest Clustering

Input: \mathcal{D} the patient set, n number of patients, p , number of features, $\mathbf{X} \in \mathbb{R}^{n \times p}$ feature matrix, \mathbf{y} class labels associated with the clinical variable of interest for the patient set, \mathcal{F} the random forest classifier, d_l and d_u : the lower bound and upper bounds of the range where the depth will be sampled, respectively, \mathbf{r} Random Forest parameters, \mathbf{c} clustering parameters.

Output: Partitioning of C_i 's where $\mathcal{D} = \cup_{i=1}^k C_i$, k is the number of clusters and $C_i \cap C_j = \emptyset$ for all $i, j \in 1, \dots, k$.

1. $\mathcal{F} \leftarrow \text{RFClassifier}(\mathbf{X}, \mathbf{y}, \mathbf{r})$
2. $\mathbf{S} \leftarrow \text{RFSimilarity}(\mathcal{F}, d_l, d_u)$
2. Convert \mathbf{S} to a distance matrix, \mathbf{D}
3. $C \leftarrow \text{Cluster}(\mathbf{D}, \mathbf{c})$
4. return C

patients. We denote the feature vector derived from patient molecular data with $\mathbf{x}^{(i)} \in \mathbb{R}^p$ and the clinical variable of interest for patient i is denoted with $y^{(i)}$. In this work, we assume that the clinical variable of interest is dichotomized; however, it can be generalized to continuous outputs by using a random forest regression or random forest survival model instead of a classifier. We will represent the $n \times p$ feature matrix with \mathbf{X} and \mathbf{y} is the $n \times 1$ label vector. We want to find a partitioning \mathcal{C} such that: \mathcal{D} is grouped into a number of disjoint subsets C_i 's where, $\mathcal{D} = \cup_{i=1}^k C_i$ and where the clustering is guided by \mathbf{y} .

A. WSURFC: Weakly Supervised Random Forest Clustering

- **Step 1:** Given the feature vector of patient samples, \mathbf{X} and the clinical variable of interest, \mathbf{y} , train a random forest classifier to classify the clinical variable. We denote the random forest model with \mathcal{F} .
- **Step 2:** Calculate the patient similarities based on co-occurrence in the feature subspaces formed by the trees in \mathcal{F} . To calculate the similarity of patients i and j , draw a random depth, d , within a predefined depth range. Sort down i and j in the forest of trees and check whether i and j fall onto the same internal node at this randomly drawn depth d . Calculate the pairwise patient similarity based on the fraction of times the patients fall on the same internal node. Form the patient similarity matrix S and convert the similarity matrix to a distance matrix.
- **Step 3:** Use this distance matrix for clustering patients.

The detailed procedure is summarized in Algorithm 1 in detail and demonstrated in Figure 1.

1) *Step 1: Random Forest Classification:* WSURFC uses a random forest classifier to predict the clinical variable to obtain the feature subspaces. Random forest is an ensemble method that learns many decision trees and aggregates their results [21]. Each tree is independently trained using a bootstrap sample of the training examples. In growing a tree, at each split, m features are randomly selected from the global set of features, and the one that maximizes an impurity criterion, in this case, the Gini index, is selected. The input samples

Algorithm 2 Random Forest Subspace Similarity (RFSim)

Input: D set of n samples, B number of trees in the random forest, \mathcal{F} random forest model, Z_b b -th bootstrap sample by which tree T_b is trained with, $z_{i,j}$ number of bootstrap samples where i and j are in the bag.

Output: \mathbf{S} : $n \times n$ similarity matrix

1. For each i, j pair in D
 - i. For all bootstrap samples Z_b where i, j are both in Z_b
 - (a) Get tree T_b of B
 - (b) Get height h_b of T_b
 - (c) Sample d from $[h_b \times d_s, h_b \times d_e]$ uniformly at random
 - (d) Traverse i on T_b until depth d is reached and find the internal node p_i on which i falls at that depth
 - (e) Traverse j on T_b until depth d is reached and find the internal node p_j on which j falls
 - (f) **if** $p_i == p_j$ **then**
 $\mathbf{S}[i, j] \leftarrow \mathbf{S}[i, j] + 1$
 - ii. $\mathbf{S}[i, j] \leftarrow \frac{\mathbf{S}[i, j]}{z_{i, j}}$
2. return \mathbf{S}

that arrive at a particular node are further split based on the value of the selected best feature. A path from the root to a node includes a subset of features in a tree. These subsets of features and their combination forms a feature subspace as depicted in Figures 1. WSURFC assesses the similarities of the input samples under these formed feature subspaces based on whether they fall in the same subspace or not.

2) *Step 2: Calculation of Random Forest Subspace Similarity:* Using the random forest ensemble, we calculate a similarity metric, which we refer to as the *RFSim*. By sorting down the example pairs onto random depths in the tree and checking how often they arrive at the same internal node in the tree, we construct a similarity matrix of the patients. Different random depths provide different views of the samples under different feature combinations. Checking whether a pair of patients would fall on the same internal node translates into checking if they are in the same subspace created by a nonlinear combination of a subset of features. For the pairs that end up at the same node, their co-occurrence count is incremented by 1. This is repeated for all the trees, and the similarities are finally normalized with the number of bootstrap samples where both samples are in the bag. The steps of this calculation are presented in Algorithm 2.

The similarity metric is similar to random forest proximity [21] with a key difference. Random forest proximity is calculated based on how often the example pairs fall onto the same leaf node; thus, the feature combinations that predict the labels are used. On the other hand, we calculate the similarity based on internal nodes; these feature representations may not be strong enough to predict the label. However, they can still be descriptive enough to reveal similarities of the examples, as we illustrate in our experiments. We observe that a depth chosen from the mid-level of the tree is useful, as discussed

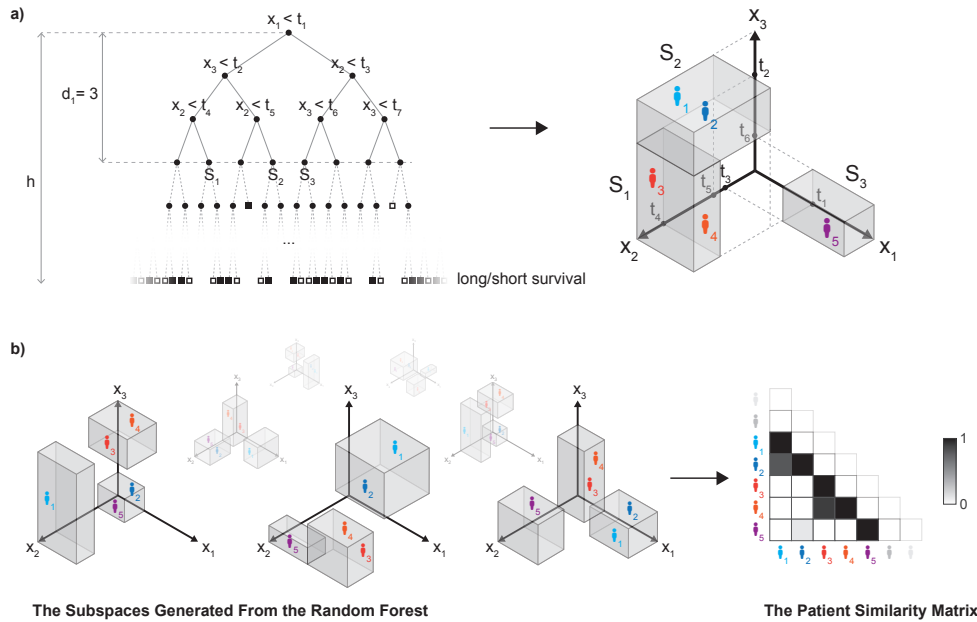


Fig. 1: Illustration of WSURFC algorithm. a) A tree in the random forest is trained to classify patients into long survivor or short survivor classes. t_1, t_2, t_3, t_4, t_5 are features of the patients. h is the height of the tree. At the random depth $d_1 = 3$, patients 3 and 4 shown with orange color fall into same node S_1 . Patients 1 and 2, shown in blue, fall on node S_2 , and patient 5 falls on node S_3 . At this level, three subspaces S_1, S_2 and S_3 arise. These subspaces are used to calculate the similarity of patient pairs. b) Different depth values to generate different partitions of subspaces are selected uniformly at random. The similarity of the two patients is calculated based on how many times they fall in the same subspace for the trees in which these two patients are both in the bootstrap samples. Patients 3 and 4 are in the same subspace three times, while patients 2 and 5 are in the same subspace for once. Therefore, the similarity value of patients 3 and 4 is closer to 1, and it is indicated with a darker color.

in the Results section.

3) *Step 3: The Final Clustering:* In the last step of the algorithm, we convert the similarity matrix to a distance matrix by subtracting the values from 1 and input this distance matrix to the clustering algorithm. For clustering, we use the hierarchical clustering algorithm with average linkage, but other clustering algorithms can be used at this stage.

B. Datasets

We first use MNIST handwritten digit dataset [22] to demonstrate WSURFC. The dataset contains images of 28×28 handwritten pixel digits. We sample 5000 examples from the entire set. Next, we apply our algorithm to breast cancer molecular data. We use data that contain solid primary tumors mRNA, microRNA, and protein expression levels that are made available through the Cancer Genome Atlas (TCGA) project [23]. The data is retrieved through the UCSC (University of California, Santa Cruz) Cancer Genomics Browser. The mRNA expression data contain 1196 patients and 20531 transcripts. miRNA expression data comprise information from 1194 patients and 1046 features. Protein expression data hold information on 747 patients and 131 proteins.

III. RESULTS

To demonstrate the method’s effectiveness, we perform experiments on a problem that is easier to inspect, the well-known MNIST handwritten digit dataset. Subsequently, results

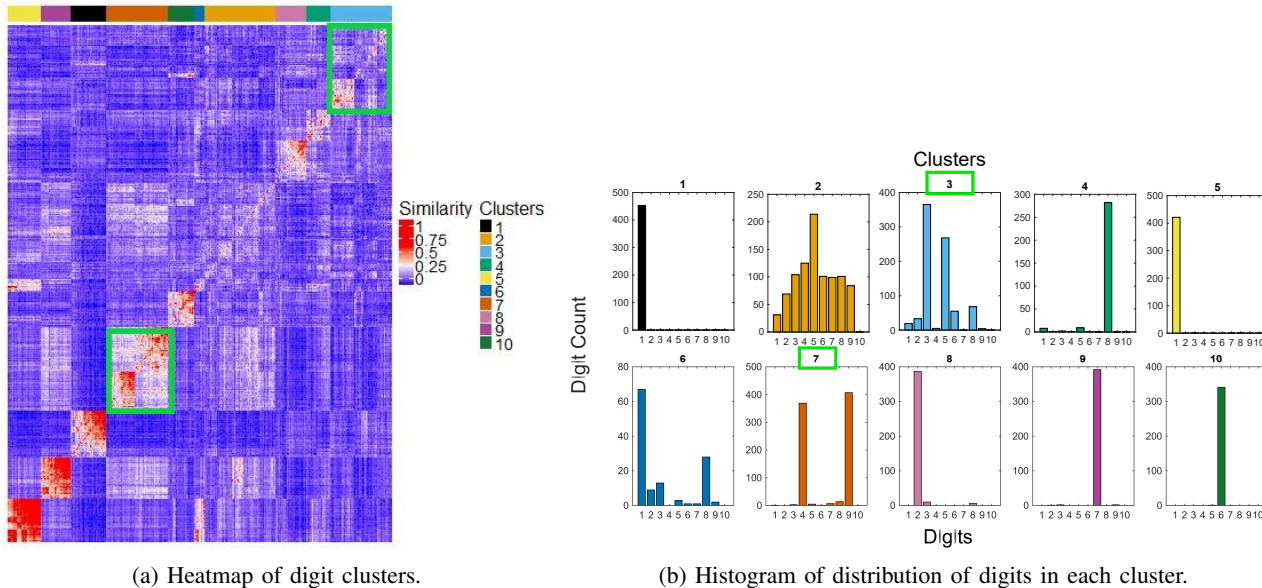
of applying WSURFC on the breast cancer patient dataset are presented.

A. Results on MNIST Digit Dataset

The random forest classifier is generated with 200 trees and trained with digit labels as class labels. In constructing the similarity matrix, we sample from $[\frac{h}{3}, \frac{2h}{3}]$ interval depth. The digits are then clustered into 10 clusters by inputting the corresponding similarity matrix to the hierarchical clustering algorithm.

Figure 2 shows a heatmap of the digit similarities and the clusters. The histogram shows the digit label distribution in each cluster. The aim of clustering here is not to arrive at clusters that represent purely single digits. Because we use subspaces not close to the leaves; however, we hope to reveal structurally similar digits. WSURFC finds these similarities that are not input. For example, Cluster 7 contains mostly digit pairs 4 and 9. Both of these digits have very similar structures. Similarly, cluster 3 reveals that 3 and 5 are similar, and additionally, 8 exhibits similarities to this digit pair. These results indicate that although the classifier is trained with 10 digit labels, WS-RFClust can uncover structural similarities between individual digits. Supplementary Figure 1 shows the silhouette width of the examples of these clusterings.

1) *The effect of Sampling from Interval Nodes at Different Depths:* In applying WSURFC, the depth levels in each tree



(a) Heatmap of digit clusters.

(b) Histogram of distribution of digits in each cluster.

Fig. 2: Clustering results of handwritten digit dataset. a) Colors on the heatmap represent similarities computed for sample pairs (reds indicate high similarity, blue indicates low). The bars on top indicate different clustering. b) Each subplot that bears the same color on the histogram displays the digit content of the clusters based on their true class labels. The x-axis of a histogram represents digits, and the y-axis represents the number of observed samples in each digit. The two interesting clusters, 3 and 7, are marked with green boxes.

are sampled randomly but within a predefined depth range. To understand the effect of the sampling depth, we experiment with different depth intervals on the digit dataset: let h be the height of a tree in the forest. We experiment with selecting d from the interval lower part of the tree that is from $(0, \frac{h}{3}]$, the middle part from $[\frac{h}{3}, \frac{2h}{3}]$ and third interval is from $[\frac{2h}{3}, h]$. To speed up calculations, for training random forest classifiers, we sample 1500 digits in these experiments.

Supplementary Figure 2a displays the results where the intervals are selected from the interval nodes closer to the root. In cluster 8, we observe that the digits 4 and 9 are in the same cluster; these are digit pairs with very similar shapes. In cluster 2, in Supplementary Figure 2b, where the intervals are sampled in the medium part of the trees, the grouping of digits 4 and 9 is clearer (cluster 9). Similarly, cluster 2 reveals that 3 and 5 are similar and that digit 8 is similar to these digits.

B. Application of WSURFC to Breast Cancer

We apply WSURFC to breast cancer patient data with three different input types, mRNA, miRNA, and protein expressions (RPPA). In each case, we dichotomize the survival time of patients into two classes. Patients with survival time shorter than the 25% quartile are labeled as low survivors, whereas patients with survival times longer than the 75% quartile are labeled as high survivors. These labels constitute the class labels for the classification step. The number of patients is different in each case as the number of available patient samples for each data type is different. We apply different feature selection criteria including, ttest, ROC, Entropy, Chernoff, and Wilcoxon statistical tests to reduce the number of features. We experiment with different feature set size values and select the

one that produces the best 5-fold cross-validation accuracy. We apply 5-fold cross-validation to the training data 10 times and form decisions based on the average accuracy over 10 runs. We apply WSURFC by sampling uniformly at random from depths from the interval $[\frac{h}{3}, \frac{2h}{3}]$, where h is the height of the trees. We also run the widely adapted method NMF-Consensus clustering on each of these datasets for comparison.

To inspect clusters in each case, we use other available clinical data in breast cancer. We compare clusters in terms of survival, tumor stage distributions, and PAM50 subtypes. We used the Kaplan-Meier curves and the log-rank test to check if clusters' separations in terms of survival distributions. χ^2 test of independence is used to test the association of the clusters with the tumor stages and PAM50 subtypes.

1) *WSURFC with mRNA Expression Data*: There are 1196 patient samples with mRNA expression data. The number of long survivor patients is 299, while the number of short survivors is 300. We apply WSURFC on these 1196 patients. Let k denote the number of clusters, we try clustering with $k = 2, 3, 4, 5, 6$. Finally, we use the trained model to cluster all the samples. The best clustering resulted in $k = 5$ clusters. Figure 3a shows the Kaplan-Meier survival curves for each of the clusters; we apply the log-rank test on the survival distributions of the clusters. For $k = 5$, the p-value of the test is $4.5e-05$, indicating that survival distributions of clusters are distinct at significance level 0.05. The silhouette width plot for this clustering is shown in Figure 3b. Supplementary Figures 3 and 4 show silhouette width and survival plots for all k values, respectively.

Figure 4a demonstrates the Kaplan-Meier survival plots of the clusters when consensus NMF is applied to the mRNA data. Smallest p-value is achieved when $k = 6$ is p -value=

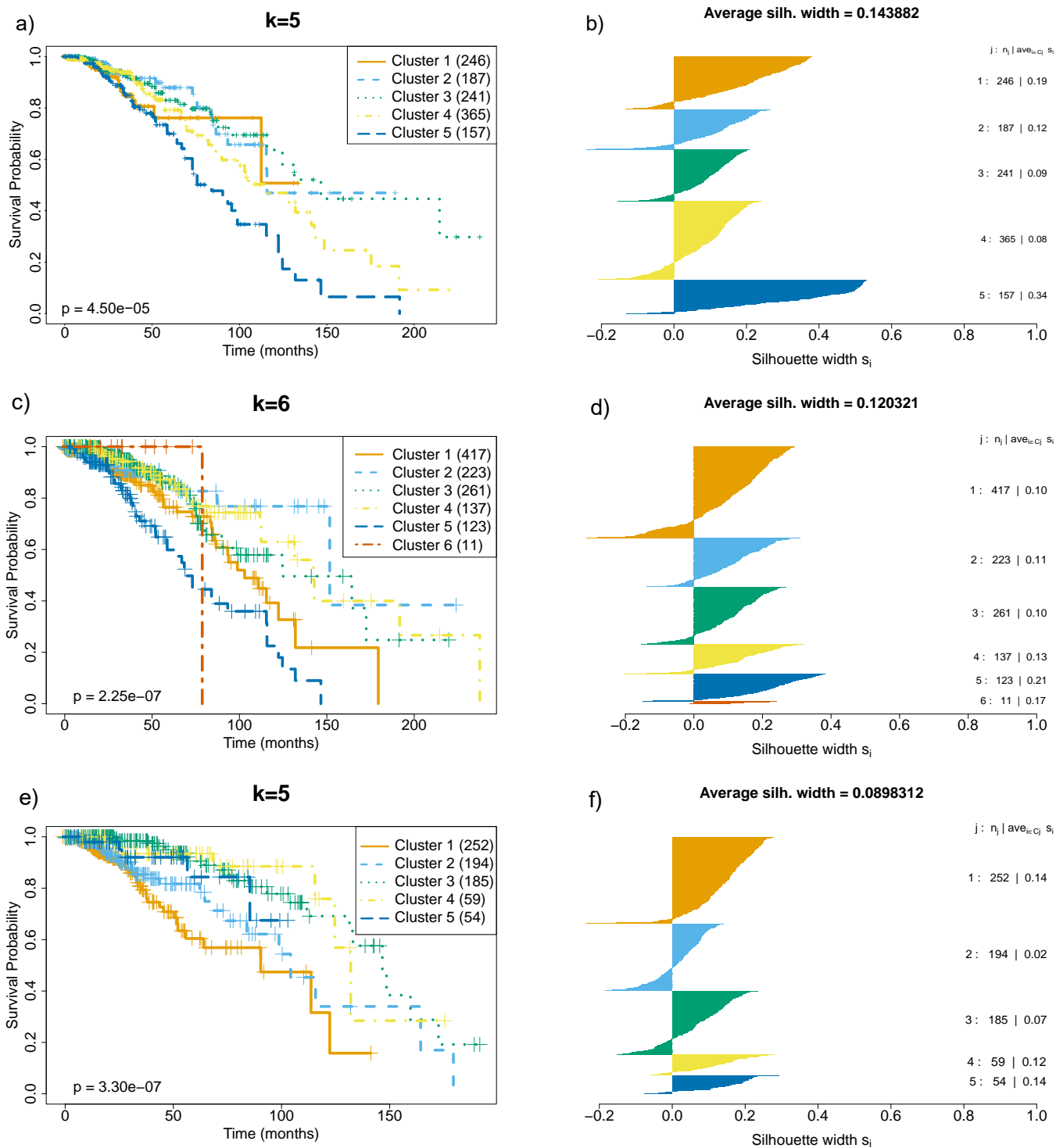


Fig. 3: In survival plots of each data type (a, c, and e), the x-axis shows the survival time in months, and the y-axis shows survival probability at a given time. In the silhouette graph of each data type (b,d,f), the x-axis is the ruler that shows the width of each cluster. j is cluster id, n_j is number of patients in cluster C_j and S_i is the silhouette width of C_j . The y-axis shows $j : n_j | \text{ave}_{i \in C_j} S_i$ for each cluster. Average silhouette width is the overall average of all clusters. a) Survival plot for mRNA clusters. b) Silhouette width plot for mRNA clusters. c) Survival plot of microRNA clusters. d) Silhouette width plot for miRNA clusters. e) Survival plot of RPPA clusters. f) Silhouette width plot for RPPA clusters.

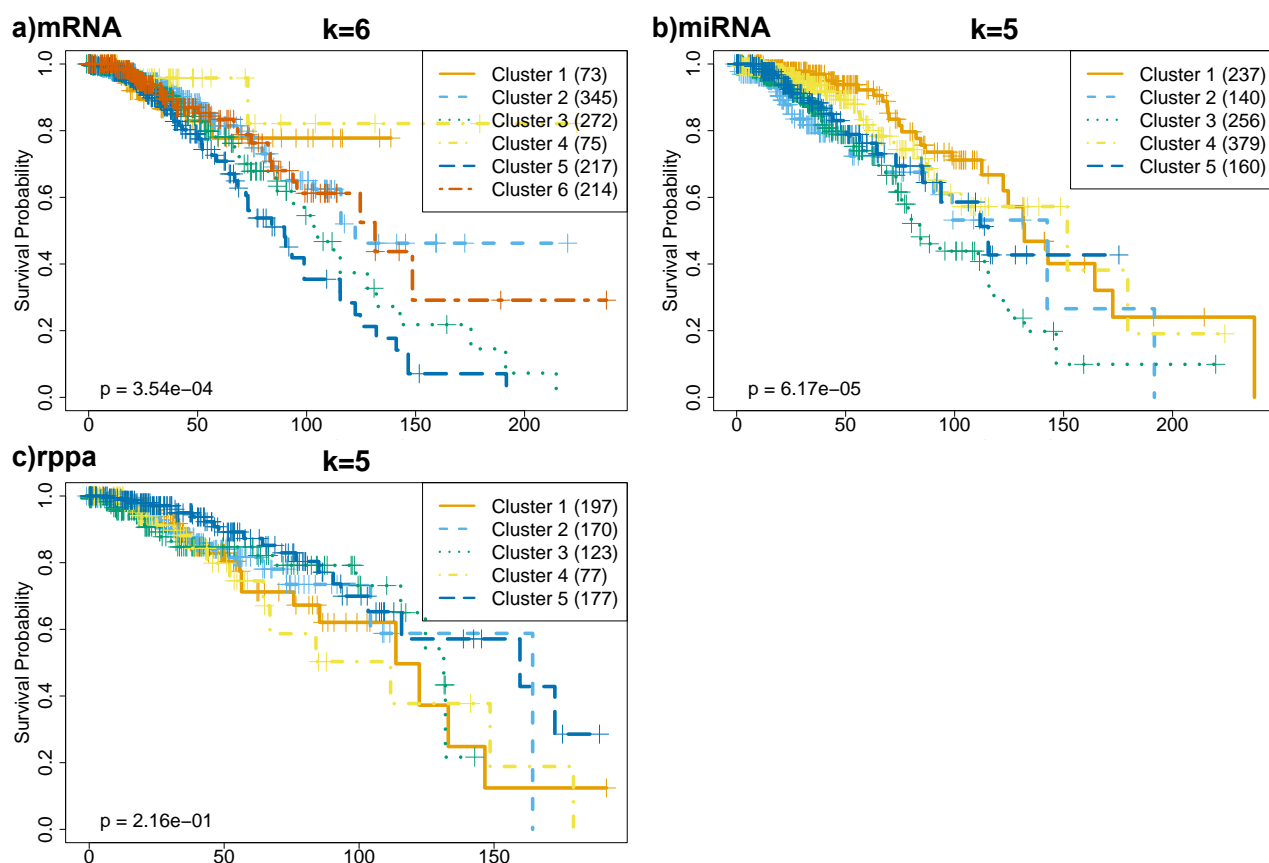


Fig. 4: a) Survival plot of consensus NMF method on mRNA data. Smallest p-value is achieved at $k = 6$. b) Survival plot of consensus NMF method on microRNA data. Smallest p-value is achieved at $k = 5$. c) Survival plot of consensus NMF method on RPPA data. Smallest p-value is achieved at $k = 5$

$3e-04$. This is 100 x fold larger than the p-value we obtain in WSURFC and indicates that better clusters are obtained with WSURFC. We test the null hypothesis that there is no association between the tumor stages and the clusters at the significance level 0.05 and reject the null hypothesis with $p = 0.03$. In this test, we exclude stages Stage IB, Stage II, Stage III, Stage Tis, Stage X, Stage IV; because only a few patients belong to these stages. We next check whether the identified clusters are related to the PAM50 subtypes. We conclude WSURFC clusters have a strong association with the intrinsic molecular subtypes. The corresponding test is $p < 2.2e-16$.

WSURFC with miRNA Expression Data: microRNA expression matrix contains 1172 available patients and 1046 miRNA expression level features. We follow the same steps in training the mRNA expression data. We select only the low and high survivor 587 patients from miRNA expression data and select 200 features with the t-test. We divide these patients into 476 training and 116 test examples. Test examples are classified with a random forest. Then, all the patients (1172) that are available in the dataset are input to train the model, and WSURFC constructs a similarity matrix of patients. We apply hierarchical clustering for $k = 2, 3, 4, 5, 6$. Best clustering is achieved with $k = 6$. Figure 3b shows the Kaplan-Meier plot for $k = 6$, which yields a very low p-value of $2.25e-07$

and the silhouette width plot for this clustering is shown in Figure 3d. Supplementary Figures 5 and 6 show silhouette width and survival plots for all k values, respectively.

We applied Consensus NMF to the microRNA dataset to compare the clustering performance of WSURFC with Consensus NMF. We run the consensus NMF algorithm dataset with 1172 samples containing all the patients. Figure 4 b) demonstrates Kaplan-Meier survival plot for the best clustering, $k = 5$ with p-value ($p = 6.17e-05$). This value is larger than p-value which we obtain in WSURFC as $p = 2.25e-07$ in $k=5$. Therefore, we conclude that WSURFC provides a better separation of clusters.

We next checked the association of the clusters found with the tumor stages, the χ^2 test results with $p = 0.002 < 0.05$, therefore we reject the null hypothesis in favor of the alternative hypothesis, which states that the tumor stages are associated with WSURFC subtypes in miRNA dataset. Note that we excluded stages Stage IB, Stage II, Stage III, Stage IIIB, Stage Tis, Stage X, Stage IV due to small numbers of patients belonging to those stages. Finally, we tabulate the data into PAM50 cluster ids and WSURFC subtypes and apply the χ^2 test of independence for the clustering results with $k = 6$. The resulting $p < 2.2e-16$ of test is considerably smaller than the significance level 0.05.

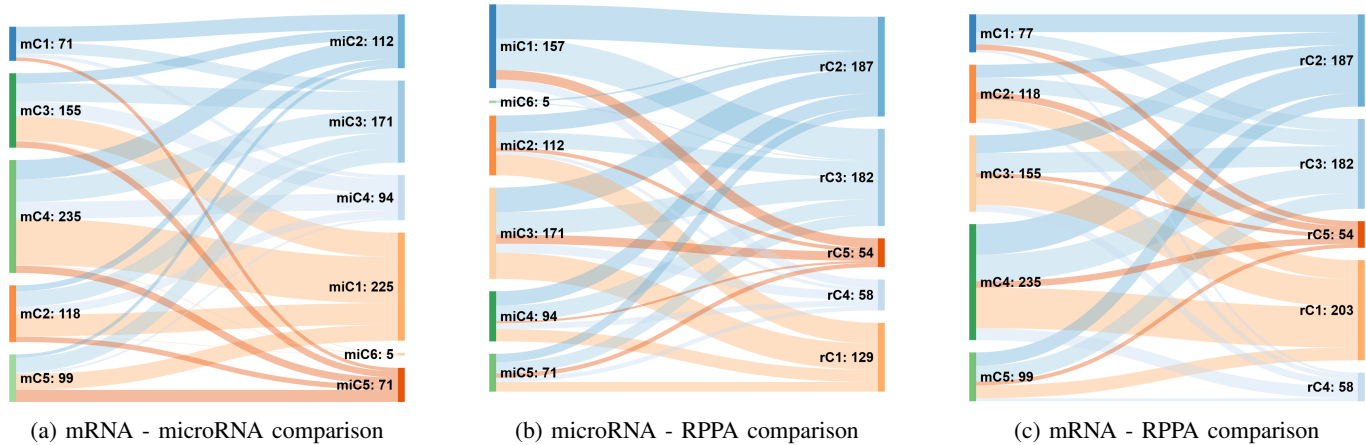


Fig. 5: Overlap between clusters obtained with different molecular data as inputs. The bars in the left and right columns are results of two different clusterings; the flow between a pair of clusters indicates the overlapping patients in those clusters. The diagrams are created using the SankeyMATIC tool.

WSURFC with RPPA Expression Data: As the final input, we experiment with protein expression. Protein expression data is collected on 744 available patients and 131 features. There are only 131 features in the RPPA dataset; therefore, we use all the features without any feature selection. The protein expression data contains 373 low and high survivor patients. We utilize 299 of them as the training set and 74 of them as the test set. We apply hierarchical clustering for $k = 2, 3, 4, 5, 6$. The resulting clusters are compared with respect to survival rate, tumor stage, and PAM50 subtypes.

Supplementary Figure 7 indicates silhouette width graph of clustered patients in the protein expression dataset. Supplementary Figures 7 and 8 show silhouette width and survival plots for all k values, respectively. The best clustering is achieved when $k = 5$ with a p -value of $3.30e-07$ when $k=5$. Supplementary Figures 7 and 8 show silhouette width and survival plots for all k values, respectively.

We apply Consensus NMF to the protein expression dataset to compare the clustering performance of WSURFC with Consensus NMF. We run the consensus NMF algorithm dataset with 744 samples containing all the patients. Figure 4c) demonstrates Kaplan-Meier survival plots for each k value when consensus NMF is applied. p -value is $p = 0.055$ when $k=5$, overall p -value range is between $0.01 - 0.2$. Consensus NMF results are not confidently below $\alpha = 0.05$. Therefore we conclude that Consensus NMF clusters are not significantly different in terms of survival rate. WSURFC outperforms Consensus NMF in terms of survival rate differentiation between subgroups.

We exclude stages Stage IB, Stage II, Stage III, Stage IIIB, Stage Tis, Stage X, Stage IV; because there are only a few patients who belong in these stages. The χ^2 test of independence yields to p -value= 0.02; therefore, we reject the null hypothesis at a significance level of 0.05. We conclude that tumor stages are randomly distributed in the WSURFC subtypes. To compare PAM50 subtypes and WSURFC subtypes, we apply the χ^2 test of independence for $k = 5$. $p < 2.2e-16$ of the test is considerably smaller than 0.05.

Therefore WS-RClust clusters have a strong correlation with the PAM50 molecular subtypes.

C. Overlap among clusters

When used as input, each molecular input data yields a different clustering of patients. We analyzed the overlap among these clustering results. Figure 5 shows the pairwise overlap between the clustering results. We conducted this analysis with the patients whose profiles contain data pertinent to these molecular types. Some of the clusters have a large overlap, such as C4 which is obtained with mRNA and C1 of miRNA data, as indicated by the large flow size (Figure 5a). Similarly, the miRNA C1 cluster is mainly composed of patients that belong to the RPPA C2 and C3 clusters (Figure 5b). mRNA C4 cluster with RPPA C1 cluster and mRNA C3 with RPPA C1 cluster share a considerable number of patients (Figure 5c). These results also indicate that different molecular data types can bring different views into clusterings, and a multi-view approach could be useful to integrate the clusters. We leave this direction as a future work.

IV. CONCLUSION

Inaccurate grouping of patients hinders the development of effective targeted therapies. Identifying patient subgroups with similar molecular profiles can reveal the unique molecular characteristics that shape them and open up possibilities for targeted therapeutics. Traditionally, unsupervised clustering analysis is applied to the genomic data of tumor samples, and the patient clusters are considered interesting if they can be associated with a clinical outcome variable such as the survival rate of patients [5], [12], [24], [25]. We propose a weakly supervised clustering framework (WSURFC) in place of this unsupervised framework. In this approach, the clustering partitions are weakly guided with the clinical outcome of interest. We achieve this by using the similarity of patients under subsets of features created in a random forest ensemble which is trained with a label of interest.

We apply WSURFC to handwritten digit datasets to understand the effect of several parameters. To understand how the sampling from different levels of the tree would affect clustering, we vary the interval range from which we sample random depths. We observe that if the depths are close to the tree height, the resulting partitions are found to be close to the leaves, and therefore, these clusters correspond to the classes. If we choose depths near the root, the structure information is lost. Thus, we conclude that sampling from a medium-range is critically important to attain the best trade-off between predictive accuracy and speed.

A widely adopted technique in cancer is using the Consensus NMF clustering approach. We apply WSURFC to TCGA breast cancer miRNA, mRNA, and protein expression datasets separately to identify breast cancer subtypes. We also run the Consensus NMF approach on the same datasets to see if we can capture better subgroupings of patients. We vary the number of clusters and analyze these clusters in terms of internal cluster validity metrics, such as silhouette width, and external clinical data, such as tumor stage and PAM50 classification. When the data are clustered into 5 or 6 subgroups, the resulting survival rates of subgroups significantly differ from each other and are better in terms of survival separation compared to the consensus NMF approach. Regardless of the input expression data type, the method performs well, and the resulting clusters are found to be associated with the tumor stages and PAM50 subtypes. Although the different clusterings obtained with different molecular data types have large overlap, they are different. A multi-view clustering approach, where each view is obtained with a molecular data type, could be interesting.

In this study, we have limited our analysis to breast cancer, but the approach presented herein can be applied to any cancer type and any clinical variable of interest. This work assumes that the target variable y is discrete; however, the approach can easily be extended to the cases where y is a continuous variable, replacing the random forest classifier with a regressor. Alternatively, the target variable can be cast as a survival variable, and a random survival forest can be adapted.

ACKNOWLEDGMENT

The results shown here are in part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

REFERENCES

- [1] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (p4) cancer medicine," *Nature reviews Clinical oncology*, vol. 8, no. 3, p. 184, 2011.
- [2] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature reviews Clinical oncology*, vol. 15, no. 2, pp. 81–94, 2018.
- [3] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Verizer, J. McMichael, L. Fulton, D. Dooling, L. Ding, E. Mardis *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.
- [4] P. S. B. Joel S. Parker, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, p. 1160–1167, 2009.
- [5] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*," *Cancer cell*, vol. 17, no. 1, pp. 98–110, 2010.
- [6] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.
- [7] A. Ally, M. Balasundaram, R. Carlsen, E. Chuah, A. Clarke, N. Dhalla, R. A. Holt, S. J. Jones, D. Lee, Y. Ma *et al.*, "Comprehensive and integrative genomic characterization of hepatocellular carcinoma," *Cell*, vol. 169, no. 7, pp. 1327–1341, 2017.
- [8] The Cancer Genome Atlas Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, pp. 609–615, 2011.
- [9] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2014.
- [10] F. Chen, D. S. Chandrashekar, S. Varambally, and C. J. Creighton, "Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers," *Nature communications*, vol. 10, no. 1, pp. 1–15, 2019.
- [11] M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta *et al.*, "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma," *Cell*, vol. 164, no. 3, pp. 550–563, 2016.
- [12] N. Rappoport and R. Shamir, "Multi-omic and multi-view clustering algorithms: review and cancer benchmark," *Nucleic acids research*, vol. 46, no. 20, pp. 10546–10562, 2018.
- [13] L. Cantini, P. Zakeri, C. Hernandez, A. Naldi, D. Thieffry, E. Remy, and A. Baudot, "Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.
- [14] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.
- [15] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.
- [16] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "Pinsplus: a tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019.
- [17] Y. I. Tepeli, A. B. Ünal, F. M. Akdemir, and O. Tastan, "Pamogk: a pathway graph kernel-based multimomics approach for patient clustering," *Bioinformatics*, vol. 36, no. 21, pp. 5237–5246, 2020.
- [18] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLOS Biology*, vol. 2, no. 4, April 2004.
- [19] D. C. Koestler *et al.*, "Semi-supervised recursively partitioned mixture models for identifying cancer subtypes," *Bioinformatics*, vol. 26, no. 20, pp. 2578–85, 2010.
- [20] E. A. Houseman *et al.*, "Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions," *BMC Bioinformatics*, vol. 9, p. 365, 2008.
- [21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [22] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits." [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [23] National Cancer Institute. (2011) The cancer genome atlas. [Online]. Available: <http://cancergenome.nih.gov/>
- [24] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, vol. 10, no. 11, pp. 1108–1115, 2013.
- [25] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, vol. 31, no. 12, pp. i268–i275, 2015.



Duygu Ozcelik received her B.Sc. degree in Computer Science from Bilkent University of Ankara, Turkey (2011) and her M.Sc. in Computer Science from Bilkent University of Ankara, Turkey (2016). She has been working as a Senior Software Engineer at Havelsan since 2015.



Oznur Tastan obtained her B.Sc. degree in Biological Sciences and Bioengineering from Sabanci University (2004) and her M.Sc. and Ph.D. degrees in Computer Science from Carnegie Mellon University (2007 and 2011), respectively. She worked as a postdoctoral researcher at Microsoft Research New England. She worked as an assistant professor at Computer Engineering Department of Bilkent University from 2012 to 2017. Since 2017, she has been an assistant professor at the Faculty of Engineering and Natural Sciences of Sabanci University,

affiliated with the Computer Science and Engineering and Molecular Biology Genetics and Bioengineering programs.