scientific reports

OPEN

Subnetwork representation learning for discovering network biomarkers in predicting lymph node metastasis in early oral cancer

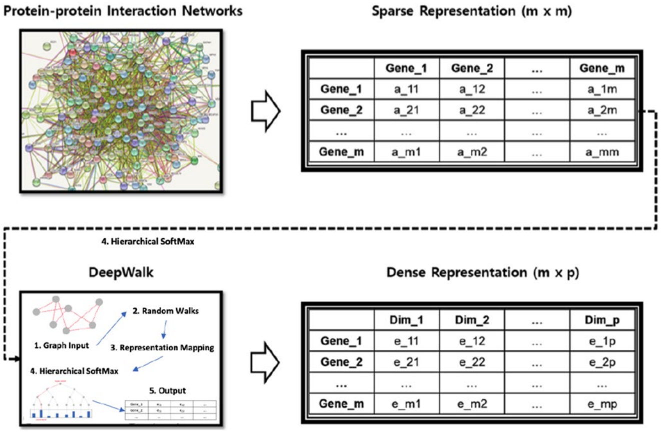
Minsu Kim[1], Sangseon Lee[2], Sangsoo Lim[3], Doh Young Lee[4]✉ & Sun Kim[3,5,6,7,8]✉

Cervical lymph node metastasis is the leading cause of poor prognosis in oral tongue squamous cell carcinoma and also occurs in the early stages. The current clinical diagnosis depends on a physical examination that is not enough to determine whether micrometastasis remains. The transcriptome profiling technique has shown great potential for predicting micrometastasis by capturing the dynamic activation state of genes. However, there are several technical challenges in using transcriptome data to model patient conditions: (1) An Insufficient number of samples compared to the number of genes, (2) Complex dependence between genes that govern the cancer phenotype, and (3) Heterogeneity between patients between cohorts that differ geographically and racially. We developed a computational framework to learn the subnetwork representation of the transcriptome to discover network biomarkers and determine the potential of metastasis in early oral tongue squamous cell carcinoma. Our method achieved high accuracy in predicting the potential of metastasis in two geographically and racially different groups of patients. The robustness of the model and the reproducibility of the discovered network biomarkers show great potential as a tool to diagnose lymph node metastasis in early oral cancer.

Oral tongue squamous cell carcinoma (OTSCC) is one of the most common malignant tumors in the oral cavity[1]. Cervical lymph node metastasis is a major factor in a poor prognosis for OTSCC and also occurs even in early stages[2]. Currently, clinical diagnosis relies on physical examinations such as palpation, ultrasonography, computed tomography (CT-scan), and magnetic resonance imaging (MRI). Unfortunately, these physical examinations are not accurate enough to determine if micrometastasis remains in the lesion. Micrometastasis indicates that a small number of cancer cells that have spread from the primary tumor to other parts of the body are too few to be detected by screening or physical examination. For this reason, clinicians recommend lymphadenectomy for patients who do not require resection[3]. Lymphadenectomy refers to surgery to remove lymph nodes, which can cause serious side effects. Therefore, being able to detect micrometastases with molecular-level data could be of significant benefit to patients with OTSCC.

Transcriptome data are whole genome-scale molecular profiles generated by high-throughput RNA profiling techniques such as microarrays and RNA sequencing (RNA-seq), which are known to have great potential to identify micrometastasis in cancer patients[4–6]. There are several challenges in modeling patient conditions using transcriptome data. First, despite advances in high-throughput RNA profiling technology, the cost of production per sample is still at a non-negligible level, and the number of genes to consider is relatively large compared to the number of samples, which is a challenge for many researchers. This problem is also referred to as the low sample high dimension problem[7]. In addition, cellular proteins rarely act individually and generally cooperate to perform specific functions and express a specific phenotype[8]. Therefore, the complex dependence between genes due to protein interactions should also be considered. Finally, heterogeneity between patient samples is

[1]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [2]Institute of Computer Technology, Seoul National University, Seoul 08826, Korea. [3]Bioinformatics Institute, Seoul National University, Seoul 08826, Korea. [4]Seoul National University College of Medicine, Seoul 03080, Korea. [5]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Korea. [6]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea. [7]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Korea. [8]AIgenDrug, Co., Ltd, Seoul, Korea. ✉email: gedo0212@naver.com; sunkim.bioinfo@snu.ac.kr


Protein-protein Interaction Networks
Sparse Representation (m x m)
4. Hierarchical SoftMax
DeepWalk
Dense Representation (m x p)
1. Graph Input
2. Random Walks
3. Representation Mapping
4. Hierarchical SoftMax
5. Output

	Gene_1	Gene_2	...	Gene_m
Gene_1	a_11	a_12	...	a_1m
Gene_2	a_21	a_22	...	a_2m
...	...	...	...	...
Gene_m	a_m1	a_m2	...	a_mm

	Dim_1	Dim_2	...	Dim_p
Gene_1	e_11	e_12	...	e_1p
Gene_2	e_21	e_22	...	e_2p
...	...	...	...	...
Gene_m	e_m1	e_m2	...	e_mp

Figure 1. Extracting subnetworks using graph embedding technique . This involves 1) generating an adjacency matrix from a given PPI network, 2) random work sampling from a given PPI graph, and 3) generating a word2vec representation of the sampled works to generate a dense representation of each gene.

known to have a significant impact on cohort studies due to the genetic diversity between individuals with different geographic and ethnic backgrounds[9, 10].

Subnetwork level representation (SLR) is one of the most promising ways to reduce the high dimensions of transcriptome data using biological networks. Studies have shown that the SLR approach using protein-protein interaction (PPI) network is excellent for predicting the clinical status of cancer patients in terms of robustness and effectiveness[11–14]. Additionally, the SLR approach can provide a comprehensive understanding of the underlying mechanisms by which the disease progresses and influences prognosis[15, 16].

The biggest challenge when using PPI networks is the sparse network representation[17, 18]. The integration of the gene expression matrix and the adjacent matrix is not an easy task even without sparsity because they have completely different shapes. Subsystem Activation Score (SAS) is one of the most effective ways to solve this problem[19]. SAS introduced a natural way to integrate PPI networks with the transcriptome. In a recent study, Lim et al.[20] compared several SLR methods including SAS in terms of 1) reproducibility of RNA-seq data characteristics, 2) robustness to noise, 3) classification for tumor versus normal information, 4) classification for survival information, and 5) classification for cancer subtypes. They devised various statistics to measure the performance of each method on each criterion. The study showed that SAS has the best overall performance compared to other SRL methods when evaluated for the above five criteria.

According to BioGrid[21] , each gene has an average number of PPI interactions of 9.56, meaning each gene is linked to an average of 9.56 genes. Since there are at least 20 000 whole genomes, looking at the vector representation of the adjacency matrix of the PPI network, these vectors will look very sparse. That is, most values are 0 and very few (average 9.56) values are 1. As Perozzi et al.[18] stated, this sparsity can make generalizations in statistical learning models extremely difficult. DeepWalk[18] is a well-designed solution to this sparsity problem and we wanted to take advantage of it. The proposed method was largely motivated by the work of Perozzi et al.[18] using DeepWalk as a graph embedding method.

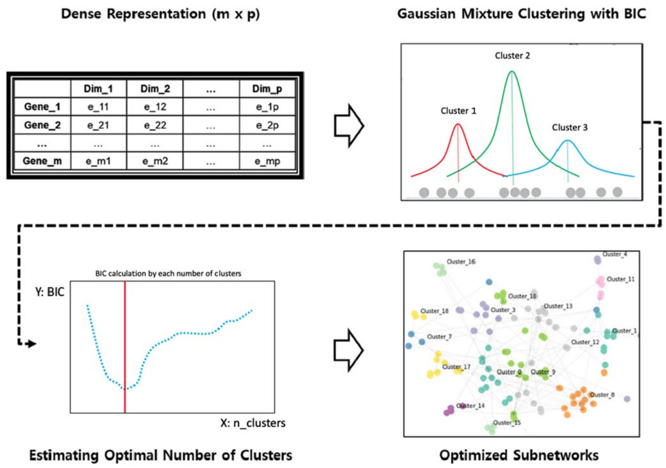
In this paper, we propose a method to discover network biomarkers and determine their metastasis potential in early OTSCCs designed to overcome the aforementioned challenges. To achieve this goal, we have developed three new computational techniques that are combined into a single computational framework, including a supervised subnetwork level representation learning system for extending SAS , a subnetwork extraction method using the DeepWalk graph embedding technique , and an attention-based classification system for integrating subnetwork level representations and discovering network biomarkers. In the following sections, we described (1) how the proposed method defined subnetworks for identifying network biomarkers, (2) how the problem is addressed as a machine-learning framework that calculated the representation of each subnetwork based on the given input and target variables, and (3) how it was applied to early oral cancer to predict lymph node metastasis.

Materials and methods

The proposed method works in three stages: (1) subnetwork extraction using graph embedding technique, (2) construction of subnetwork level representation, and (3) integration of subnetwork level representation into the master-level decision.

Subnetwork extraction using graph embedding technique.

Extracting subnetworks from a given PPI network , taking into account its biological significance, is an important task in constructing subnetwork-level representations. Essentially, the problem can be thought of as a clustering node within a PPI network represented in the form of an adjacency matrix (Fig. 1). The sparsity of network representation is useful for defining clusters, but at the same time is a huge challenge to the generalization of machine learning. DeepWalk is a powerful tool to deal with this problem, deploying representation learning techniques based on neural networks such as Word2Vec[18, 22]. It works as a graph embedding tool and shows good performance when used for node classification[18]. In the study of Perozzi et al.[18], DeepWalk was compared with five other methods in terms of the



	Dim_1	Dim_2	...	Dim_p
Gene_1	e_11	e_12	...	e_1p
Gene_2	e_21	e_22	...	e_2p
...	...	...	...	...
Gene_m	e_m1	e_m2	...	e_mp

Figure 2. Subnetwork clustering using latent representation. This involves (1) applying a Gaussian mixture model to a given dense representation of the PPI network using a wide range of the number of components as parameters, (2) evaluating each model by calculating the BIC criterion, and (3) choosing the best model to create a subnetwork for a given PPI network.

multi-label classification task, which is a problem of erasing some of the labels and guessing the erased labels through node clustering when given a graph with labels on each node. DeepWalk outperformed all other opponents under various experimental conditions. Based on the Macro-f1 score, DeepWalk's performance reaches up to 43.05%.

DeepWalk receives the sparse representation of the PPI network and generates a dense representation of the individual nodes encoding the relationship between each node in a continuous vector space with a reduced number of dimensions (Fig. 1). Using the encoded vector as a new representation of each node, we can solve the problem of extracting subnetworks by transforming it into a clustering problem. We used the Gaussian Mixture Model (GMM) and Bayesian Information Criteria (BIC) (Eq. 1) to estimate the optimal number of clusters for a given PPI network , and each resulting cluster can be considered a subnetwork (Fig. 2). For this step, the Python library `scikit-learn-0.19.2` was used[23].

$$BIC = \ln(n)(kd) - 2\ln(p(x \mid \widehat{\theta}, M)) \tag{1}$$

where x is the observed data, n is the number of data points in x, and k is the number of clusters. d is the number of dimensions of the latent representation generated by DeepWalk. $p(x \mid \widehat{\theta}, M)$ represents the maximum value of the GMM likelihood function. Where $\widehat{\theta}$ is the parameter value that maximizes the likelihood function. The model with the lowest BIC value is considered optimal.

In order to improve interpretability and reduce noise, we took the Hallmark Gene Set (HGS) from the molecular signature database (MSigDB)[24] to limit the gene space. HGS is a well-selected group of functional genes, in which genes associated with a common cancer phenotype are grouped into a set of genes. HGS has 50 genesets containing a total of 4,384 genes. For each geneset, we first generated a PPI network graph using the protein-protein interactions between the genes of each geneset. Here , PPI networks were extracted from BioGrid[21] using only high-confidence protein interactions . For each PPI network graph, we applied DeepWalk to create a vector space, then applied GMM to create subnetworks (Figs. 1 and 2). By defining sub-networks within each HGS geneset, the genes in each sub-network are not only closely linked in terms of PPI network , but also in terms of cancer phenotype. In summary, 279 subnetworks were obtained, each subnetwork assigned to one of the 50 HGS genesets.

Construction of subnetwork level representation.

Constructing a subnetwork-level representation of cancer transcriptome requires the integration of gene expression levels and PPI networks between genes so that the activity of each subnetwork can be quantified. SAS is one of the most effective tools to this end[19]. SAS uses RNA-seq samples and subnetworks generated from PPI networks as inputs to quantify subnetwork level activation for each sample. As explained in the Eqs. (2a), (2b), (2c) and (2d), SAS is a single value called Subnetwork Activation Score for the subnetwork level representation of the transcriptome. It is defined as a nonlinear combination of gene expression using the closeness centrality of each gene with a coefficient defined by a given PPI network .

$$ACT_{i,j} = N_{i,j} * \frac{(c_i r_i + c_j r_j)^2}{2(r_i + r_j)} \tag{2a}$$

$$SAS = \sum_i \sum_j ACT_{i,j} \tag{2b}$$

$$N_{i,j} = \frac{a_{ij}}{\sum_s \sum_t a_{st}} \tag{2c}$$

$$a_{ij} = \begin{cases} 1 & \text{if gene i and j are connected} \\ 0 & \text{otherwise} \end{cases} \tag{2d}$$

$ACT_{i,j}$ represents the edge level activation score between the two genes i and j. r_i represents the expression level of the gene i (TPM in this case). c_i represents the closeness centrality of the gene i within a given subnetwork PPI network . SAS is the total activation score for the subnetwork. $a_{i,j}$ is an indicator of whether two genes are linked within a given PPI network , and $N_{i,j}$ is the normalized term for $a_{i,j}$.

While SAS does not use sample labels when calibrating subnetwork representations, our goal is to predict metastasis potential in early OTSCC, so we modified SAS to better serve this purpose and named it supervised SAS (sSAS). sSAS inherited the basic idea of SAS, but calculated the coefficients in different ways (e.g. c_i and $N_{i,j}$). Rather than defining the coefficients directly in the network topology, they were estimated by maximizing the log-likelihood function (Eq. 3f) designed to be considered as a latent variable and minimize prediction errors for the labels of each sample. As shown in Eq. (3a) and (3b), $sSAS$ is defined as a logit in a logistic regression problem rather than a single activation score. x is defined as a vector containing a nonlinear combination of gene expression combined by paired combinations of genes, and θ is the latent weight corresponding to x. The problem definition is as follows.

First, the $ACT_{i,j}$ term is divided into three parts: $\frac{r_i^2}{r_i+r_j}$, $\frac{r_j^2}{r_i+r_j}$, and $\frac{r_i r_j}{r_i+r_j}$. Then, all coefficients are considered as latent variables such as w_{ij1}, w_{ij2}, and w_{ij3}. Then the linear combination of the three division terms replaces $ACT_{i,j}$ (Eq. 3a). We named it $sACT_{i,j}$ as a supervised $ACT_{i,j}$. Then the term SAS is also changed to a supervised format (e.g. $sSAS$) to estimate the latent weights by target variable (ie, sample label) (Eq. 3b). The original observations are transformed into a nonlinear combinatorial vector of gene expression x (Eq. 3c) and their weights are defined by the model parameter θ (Eq. 3d). Based on this, the logistic function $q_k(x)$ is defined to represent the estimated probability of observation x with target label k (Eq. 3e). Finally, a log-likelihood function $l(\theta_k)$ is defined so that the model parameter θ_k can be estimated by maximizing $l(\theta_k)$ (Eq. 3f and 3g).

$$sACT_{i,j} = w_{ij1}(\frac{r_i^2}{r_i + r_j}) + w_{ij2}(\frac{r_j^2}{r_i + r_j}) + w_{ij3}(\frac{r_i r_j}{r_i + r_j}) \tag{3a}$$

$$sSAS = \ln \frac{q}{1-q} = \sum_i \sum_j sACT_{i,j} \tag{3b}$$

$$x = \left\langle \frac{r_1^2}{r_1 + r_2}, \frac{r_2^2}{r_1 + r_2}, \frac{r_1 r_2}{r_1 + r_2}, \ldots \right\rangle \tag{3c}$$

$$\theta = \langle w_{ij1}, w_{ij2}, w_{ij3}, \ldots \rangle \tag{3d}$$

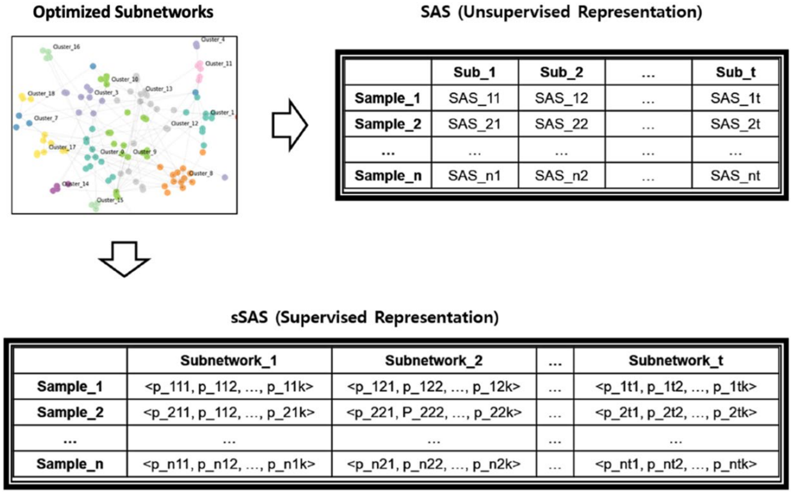
$$q_k(x) = \frac{1}{1 + e^{-\theta_k^T x}} \tag{3e}$$

$$l(\theta_k) = \sum_m y_{mk} \ln q_k(x_m) + (1 - y_{mk}) \ln(1 - q_k(x_m)) \tag{3f}$$

$$y_{mk} = \begin{cases} 1 & \text{if the label of sample m is k} \\ 0 & \text{otherwise} \end{cases} \tag{3g}$$

The representation of a subnetwork t of a sample m is defined in Eq. (4a) and (4b). In the case of multiple classes, the model parameter for each class θ_k is independently estimated in a "one versus the rest" way, then consolidated into p_{mtk} as in Eq. (4b). In our scheme, therefore, the subnetwork level representation of a sample is probability distribution estimated from the given data at each subnetwork by logistic regression model (Eq. 4c and Fig. 3). For example, if RNA-seq samples have k classes of labels then each RNA-seq sample will have a vector with dimensions of $279 * k$ because we used 279 subnetworks in this study. `scikit-learn-0.19.2` was used for this step[23].

$$q_{mtk} = \frac{1}{1 + e^{-\theta_k^T x_m}} \tag{4a}$$



SAS (Unsupervised Representation)

	Sub_1	Sub_2	…	Sub_t
Sample_1	SAS_11	SAS_12	…	SAS_1t
Sample_2	SAS_21	SAS_22	…	SAS_2t
…	…	…	…	…
Sample_n	SAS_n1	SAS_n2	…	SAS_nt

sSAS (Supervised Representation)

	Subnetwork_1	Subnetwork_2	…	Subnetwork_t
Sample_1	<p_111, p_112, …, p_11k>	<p_121, p_122, …, p_12k>	…	<p_1t1, p_1t2, …, p_1tk>
Sample_2	<p_211, p_112, …, p_21k>	<p_221, P_222, …, p_22k>	…	<p_2t1, p_2t2, …, p_2tk>
…	…	…	…	…
Sample_n	<p_n11, p_n12, …, p_n1k>	<p_n21, p_n22, …, p_n2k>	…	<p_nt1, p_nt2, …, p_ntk>

Figure 3. Constructing subnetwork level representation. This includes (1) calculating the sSAS representation for each optimized subnetwork and (2) integrating the representations into the subnetwork-level representation for each sample.

$$p_{mtk} = \frac{q_{mtk}}{\sum_r q_{mtr}} \tag{4b}$$

$$Sub_{mt} = \langle p_{mt1}, p_{mt2}, \ldots, p_{mtk} \rangle \tag{4c}$$

Integration of subnetwork level representation into master-level decision. The remaining problem is to incorporate the constructed subnetwork level representation into a single master level decision. We solved this ensemble learning problem using the attention layer built into the neural network. The attention mechanism stems from the problem of sequence-to-sequence mapping in machine translation[25]. In the work of Bahdanau et al., the attention layer was inserted between the encoder and decoder layer to act as memory[25]. In other words, they are trained to dictate which context to focus on at a specific point in time and which context to not. The attention mechanism has been applied to various tasks and has been shown to exhibit excellent performance[26, 27]. Also, Choi et al.[27] suggested that the attention mechanism can be used to make models more explainable. In this model, the attention layer acts as a master-level decision agent trained to decide which subnetwork to focus on based on the certainty computed with each sub-network level representation (Fig. S1). As shown in Eq. (5a) and (5b), the attention layer takes the negative Shannon's entropy[28] of each subnetwork level representation. Since the entropy of a given probability distribution represents the level of uncertainty, negative entropy was used to quantify how certain each subnetwork level predictor is for a given classification task.

The negative entropy values of each subnetwork are concatenated into a single vector (Eq. 5b) and passed to the next fully concatenated layer (FC). Then softmax activation is applied, resulting in a proportional distribution that is the attention layer (Eq. 6a and 6b). Hence, the actual parameter that can be learned here is the W matrix (Eq. 6a), which learns to decide which subnetworks to focus on based on the C vector. The actual decision-making process is described in Eq. (6c) and (6d) (Fig. S2). It is basically the weighted sum of subnetwork level representation for each class, where the weights are learned by the attention mechanism. The prioritization of features by the model is an instance-wise process, so each sample gets a different attention value depending on their subnetwork level representation. Python libraries `tensorflow-1.10.0`[29] and `keras-2.2.2`[30] were used for this step.

$$c_t = \sum_k p_{tk} \ln p_{tk} \tag{5a}$$

$$C = \langle c_1, c_2, \ldots, c_t \rangle \tag{5b}$$

$$H = softmax(WC^T) \tag{6a}$$

$$H = \langle h_1, h_2, \ldots, h_t \rangle \tag{6b}$$

$$d_k = \sum_t h_t * p_{tk} \tag{6c}$$

$$f_k = \frac{e^{d_k}}{\sum_s e^{d_s}} \tag{6d}$$

Evaluation design

Two sets of experiments were prepared to evaluate the proposed method. 1) The first used breast invasive carcinoma (BRCA) cohort data from the Cancer Genome Atlas (TCGA) consortium[31] (referred to as BRCA-case). 2) The second used squamous cell carcinoma of the head and neck (HNSC) cohort data from TCGA[32] and proprietary data provided by SMG-SNU Boramae Medical Center (referred to as ORAL-case). The purpose of the first experiment was to test the model's performance with a well-known dataset so that the results could be compared to previously studied knowledge that corresponds to the case. The second experiment was the main subject of the study.

Data description: BRCA-case. For BRCA-case, we collected 981 RNA-seq samples from TCGA, where each of them is labeled with PAM50 classes[33]. PAM50 is a standard *de facto* method for identifying the molecular status of breast cancer, which has five subtypes: lumen A (LumA), lumen B (LumB), HER2-enriched (HER2), basal (Basal), and normal (Normal) subtypes, which was initially defined by unsupervised clustering analysis using a whole-genome scale gene expression profile. There were 499 LumA, 197 LumB, 78 HER2, 171 Basal, and 36 Normal . Note that they were all primary tumors and samples with the Normal subtype were excluded. Note that all the RNA-seq samples have gene expression levels measured by the Transcripts Per Million (TPM) scale for 20,501 genes.

Data description: ORAL-case. For ORAL-case, we collected 97 RNA-seq samples from both TCGA and SMG-SNU Boramae Medical Center, where 64 RNA-seq samples (will be referred to as TCGA-ORAL samples) were from TCGA and 33 RNA-seq samples (will be referred to as SNUH-ORAL samples) were from SMG-SNU Borame Medical Center. Each of them was primary tumors with oral tongue origin and cancer stages I and II. They were labeled with their lymph node metastatic status as Positive and Negative. For TCGA-ORAL samples, there were 31 samples labeled with Positive and 33 samples labeled with Negative, while there were 11 samples labeled with Positive and 22 samples labeled with Negative in SNUH-ORAL samples. The gene expression levels were measured the same as in the BRCA-case.

Validation. For each experiment, we divided the dataset into train and test sets to validate the performance of the method. In BRCA-case, we randomly sampled 30% of the overall samples with class label balanced and considered them as a test set. In summary, 689 samples (LumA: 350 , LumB: 138 , Her2: 55, Basal: 120, and Normal: 26) were used as a train set, while 292 samples (LumA: 149 , LumB: 59 , Her2: 23, Basal: 51, and Normal: 10) were used as a test set. In ORAL-case, the 64 TCGA-ORAL samples (Positive: 31 and Negative: 33) were used as a train set, while the 33 SNUH samples (Negative: 22, Positive: 11) were used as a test set. Note that the SNUH samples had completely different geographical and ethnic compositions (ie, Korean) compared to the TCGA samples in ORAL-case. Three types of metrics were used in the evaluation. 1) Averaged area under the curve (mAUC) , 2) Accuracy (ACC), and 3) F1 score (F1). mAUC is the adjusted value of AUC for class imbalance in a multiclass classification problem. mAUC is the average AUC for each class when treated as a binary classification (i.e. one versus the rest). The mAUC, ACC, and F1 measured with the training set are $mAUC_{train}$, ACC_{train}, and $F1_{train}$ and the values measured with the test set are $mAUC_{test}$, ACC_{test}, and $F1_test$.

Prediction power evaluation. For evaluation, we first set the baseline performance using existing machine learning methods such as Logistic Regression (LR), K Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM) , and Multi-Layer Perceptron (MLP) in each case (ie, BRCA-case and ORAL-case). The experimental setup was prepared with all possible combinations of parameters listed in Table S2. The total number of combinations was 25,664. We chose the model with the best performance in terms of mAUC and mACC and used it as the baseline performance to evaluate the proposed method. `scikit-learn-0.19.2` was used for this test[23]. Similarly, we evaluated all the possible combinations of the parameters listed in Table S1 to select the parameters to be used in the proposed method. The total number of combinations was 4,900. By comparing the best performances between the conventional methods and our method after searching for each parameter space with comparable sizes, we can estimate the extent to which our method can perform relatively better than conventional methods. The python library `tensorflow-1.10.0`[29] and `keras-2.2.2`[30] were used for this setup.

Network biomarker evaluation. The attention layer assigns weights to each subnetwork level prediction (Eqs. 6a, 6b and 6c), where the weights are the probability distribution due to softmax activation. Therefore, each weight can be considered the importance of the features learned in the decision model (Fig. S2). The weights are defined instance-wise, so the overall feature importance (i.e. FI_t) was defined as the average value of all samples (Eq.7). Then, each subnetwork was ranked in ascending order. The feature ranking of the subnetwork t is called FIR_t.

$$FI_t = \frac{\sum_{n=1}^{N} h_{nt}}{N} \tag{7}$$

Note that the h_{nt} indicates the attention value of the subnetwork t of sample n.

We conducted a test to evaluate how well the decision model prioritizes features. In BRCA-case, since the PAM50 subtyping is based on the 50 genes[33], we can use this information to design a permutation test. First, we defined a function that defines a score for the reference geneset (ie, PAM50 genes) against a given feature rankings of subnetworks (Eqs. 8a, 8b, and 8c). Then, the score of PAM50 genes with the feature rankings provided by the decision model was set as a baseline score (Eq. 8a). Next, a permutation test was performed with feature rankings shuffled 1-million times (Eqs. 8d, 8e, 8f, and 8g). The purpose of shuffling is to simulate the null hypothesis by generating randomized rankings. Then, the number of randomized settings that exceed the baseline score was counted for calculating empirical p-value (Eq. 8d). The purpose of this test was to evaluate how significantly the feature rankings learned by the attention layer reproduce the prior knowledge corresponding to the given data (ie, PAM50 genes).

$$SCORE = \frac{\sum_{g=1}^{50} RANK_g}{50} \tag{8a}$$

$$RANK_g = \frac{\sum_t SUB_{gt} * FIR_t}{\sum_t SUB_{gt}} \tag{8b}$$

$$SUB_{gt} = \begin{cases} 1 & \text{if gene g is included in subnetwork t} \\ 0 & \text{otherwise} \end{cases} \tag{8c}$$

$$p\text{-}value = \frac{\sum_{iter=1}^{1,000,000} I_{iter}}{1,000,000} \tag{8d}$$

$$I_{iter} = \begin{cases} 1 & \text{if } SCORE_{iter}^{permute} > SCORE \\ 0 & \text{otherwise} \end{cases} \tag{8e}$$

$$SCORE_{iter}^{permute} = \frac{\sum_{g=1}^{50} RANK_{g,iter}^{permute}}{50} \tag{8f}$$

$$RANK_{g,iter}^{permute} = \frac{\sum_t SUB_{gt} * FIR_{t,iter}^{permute}}{\sum_t SUB_{gt}} \tag{8g}$$

Results

BRCA-case. As mentioned, we designed an evaluation scheme for comparing the baseline methods and the proposed method in terms of mAUC and mACC. As noted, the figures below are the results of PAM50 label predictions for the TCGA-BRCA cohort. In BRCA-case, the SVM algorithm with cosine kernel PCA with 18 components, TPM legalization, and no feature selection showed the best performance ($mAUC_{test}$: 0.8700, $mACC_{test}$: 0.8390, and $F1_{test}$: 0.8320) for the baseline performance. Our method showed a better performance ($mAUC_{test}$: 0.9006, $mACC_{test}$: 0.8664, and $F1_{test}$: 0.8623) (Fig. 4), using adagrad, squared hinge loss, 10% dropout, 25% split, and with feature selection. Also, we conducted the permutation test over the feature rankings that were generated by the attention layer of the best model, which showed that PAM50 genes are significantly highly ranked in the resulted attention layer (empirical p-value: 1.1e–05). See the resulting confusion matrix of our approach in Supplementary Table S3.

ORAL-case. In ORAL-case, the RF algorithm with linear kernel PCA, dimension reduction size 4, TPM logarization, and using feature selection showed the best performance ($mAUC_{test}$: 0.7045, $mACC_{test}$: 0.7576, and $F1_{test}$: 0.7570) for the baseline performance. This is the result of predicting lymph node metastasis in OTSCC. Our method showed better performance ($mAUC_{test}$: 0.9174 , $mACC_{test}$: 0.8864 , and $F1$: 0.8333) (Fig. 5), using AdaGrad, Mean absolute percentage error, 50% dropout, 10% split, and a feature selection. The more detailed metrics are in Table 1 and the top-5 subnetworks highlighted by the attention layer are in Table 2. See the resulting confusion matrix of our approach in Supplementary Table S3.

Discussion

As described, the proposed method outperformed the baseline methods in both BRCA-case and ORAL-case, and the attention map of the best model was found to greatly reproduce the prior knowledge in BRCA-case. This indicates that our method can construct a computational model for predicting patient status based on the subnetwork level representation of transcriptome. Also, the proposed method outperformed the baseline method in ORAL-case, showing that the model constructed by our method is robust enough to be reproduced between two groups of different geographical and ethnic origins.

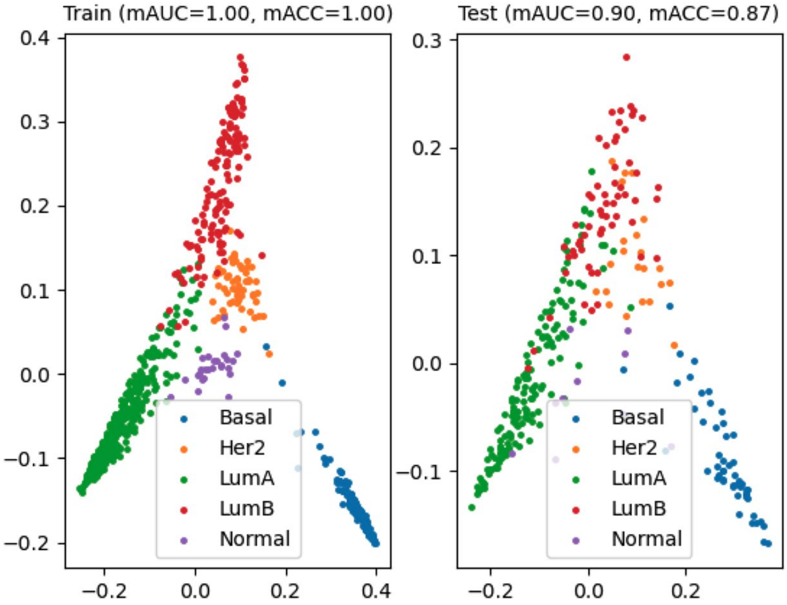


Figure 4. Performance evaluation results for PAM50 subtype prediction in breast cancer. The color-coding indicates the actual class label of samples.

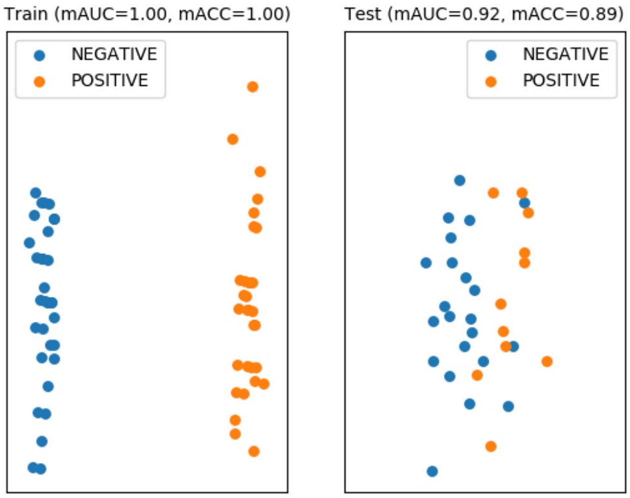


Figure 5. Performance evaluation results for lymph node metastasis prediction in early oral tongue cancer. The color-coding indicates the actual class label of samples.

Robustness of the method. Since there were not many samples available (64 for ORAL), two additional evaluations were performed instead of cross-validation to determine if the method was overfitting. In BRCA-case, we performed a permutation test to determine how the PAM50 gene was enriched in the resulting attention map of the best model, as described in the "Network biomarker evaluation" section. The PAM50 subtype is based on 50 genes, so if the model truly reflects the underlying biology and is not overfitted to noise, those 50 genes should rank higher than the others. In this regard, the results in the" BRCA-case" section clearly show that the model did not overfit in the BRCA-case (permutation test p-value: 1.1e−05).

In ORAL-case, we performed an extensive literature search to find out how the high-rank sub-networks in ORAL-case are associated with lymph node metastasis in OTSCC. The literature search results strongly suggest that the top-ranked subnetworks are closely linked with lymph node metastasis. This also implies that the model is not overfitting. The attention layer were able to capture core mechanisms known to be associated with lymph node metastasis in OTSCC as well as other cancer types, which are listed as follows.

Measure	Value*	Description
TP	10	# of True positive
TN	19	# of True negative
FP	3	# of False positive
FN	1	# of False negative
Sensitivity	0.9091	$TPR = \frac{TP}{TP+FN}$
Specificity	0.8636	$SPC = \frac{TN}{FP+TN}$
Precision	0.7692	$PPV = \frac{TP}{TP+FP}$
Negative predictive value	0.9500	$NPV = \frac{TN}{TN+FN}$
False positive rate	0.1364	$FPR = \frac{FP}{FP+TN}$
False discovery rate	0.2308	$FDR = \frac{FP}{FP+TP}$
False negative rate	0.0909	$FNR = \frac{FN}{FN+TP}$
Accuracy	0.8788	$ACC = \frac{TP+TN}{P+N}$
F1 score	0.8333	$F1 = \frac{2TP}{2TP+FP+FN}$

Table 1. Detailed metrics for lymph node metastasis prediction in early oral tongue cancer. * Note that the metrics are not class-balanced.

Subnetwork[1]	Attention[2] (%)	Genes
EPITHELIAL_MESENCHYMAL		
_TRANSITION_2	9.12	BMP1,DAB2,FBLN5,GADD45A,GEM,LOXL1,LUM,SNAI2,TPM1, VEGFA
E2F_TARGETS_5	9.06	AURKB,BRCA1,CCNE1,CDC20,CDKN2C,EXOSC8,GINS3,IPO7, MAD2L1,MCM7,POLA2,PRIM2,PTTG1,RAD1,RAD21,RANBP1, SYNCRIP,TK1,TUBB,XPO1
MYOGENESIS_6	7.97	ADAM12,CDKN1A,HSPB8,KIFC3,MB,MYOZ1,PSEN2,TNNC2, TNNT3,TPM3
TNFA_SIGNALING_VIA_NFKB_6	7.45	CCND1,CEBPB,CFLAR,ETS2,FOS,GADD45A,MAP3K8,MYC, NFE2L2,SMAD3,SPHK1,TNF,TRAF1,TRIB1
MITOTIC_SPINDLE_2	4.78	ARHGAP27,ARHGEF11,CENPE,CEP250,KIF4A,KIF5B,KIFAP3, LLGL1,LMNB1,RACGAP1,RASA1,TUBA4A

Table 2. Attention map of the best model using the proposed method. Note that the top-5 subnetworks are listed. [1]SUBNETWORK column indicates the name of subnetworks, where the prefix is the name of HGS geneset and the postfix (the integer number) is the index. [2]ATTENTION column indicates the attention values generated by the best model of the proposed method averaged by all samples (including the TCGA and SNUH samples).

(1) Epithelial-Mesenchymal Transition (EMT, 9.12%): EMT is a series of critical events observed during cancer progression including invasion and metastasis caused by the acquisition of fibroblast-like phenotype of cancer cells, which is the core mechanism of lymph node metastasis in various types of carcinoma including OTSCC[34].

(2) E2F Targets (9.06%): E2F is a class of transcription factors that regulates the expression of genes associated with cell proliferation[35], which has been known to affect the Disease-Free Survival (DFS) in oral cancer[36].

(3) TNFα Signaling via NFκB (7.45%): Tumour Necrosis Factor Alpha (TNFα) is an import inflammatory factor that has a critical role in proliferation, migration, invasion, and angiogenesis, which frequently collaborates with the Nuclear Factor Kappa B (NFκB), inducing tumor cell invasion and metastasis[37]. The TNFα and NFκB signaling have been known to be associated with invasion and metastasis in oral cancer[38].

The highlighted subnetwork of the attention map showed a strong association with the results of previous studies, suggesting that the model is very reliable in terms of consistency with prior knowledge. This also suggests that other subnetworks (eg, Mitotic spindle and Myogenesis) may be unknown regulators of lymph node metastasis in OTSCC. A complete list of the attention maps for BRCA-cases and ORAL-cases is listed in Table S3.

Biological significance of the subnetworks of PPI network extracted by DeepWalk.

In our method, a neural network-based graph embedding technique DeepWalk[18] was used for extracting subnetworks from a given PPI network . DeepWalk generates vectors of real numbers for each protein in a given PPI network , where vector distances between adjacent proteins in the PPI network are smaller than the distant proteins. Hence, clusters generated by using the DeepWalk representation can be considered as well-optimized collections of interacting proteins in terms of the PPI network . It has been known to show better performances than the classical graph clustering approaches such as SpectralClsutering and Modularity-based clustering[18]. The strength of DeepWalk comes from the random-walk-based estimation of topological distances between two proteins, meaning that it considers not only the direct edges between two proteins but also considers indirect connections implied in the neighborhood information between two proteins. It allows the model to capture the hidden relationships between two proteins, which might not be explicitly specified in the PPI network due to the lack of knowledge.

In addition, we used MSigDB Hallmark Gene Set (HGS) to annotate sub-networks, a set of cancer hallmark genes, so the framework is currently only valid for cancer tissues. It is designed to work for any cancer tissues, as it has been shown to be effective in both oral and breast cancers.

Advantages of the supervised setting for constructing subnetwork level representations. Originally, the SAS[19] method was not designed for a classification problem. Hence, we replaced the coefficients in the SAS framework with latent variables estimated by using each sample label (Eq. 3f). It virtually rewires edges within each subnetwork to be suited for solving the given classification problem. There are three advantages to this setting. First, it can fill the gap of knowledge from the data, such as unknown interactions between proteins. Secondly, it can calibrate the edge weights to reflect condition-specific interactions or broken interactions that are specific to given data. Lastly, it can reduce the weights of passenger interactions, redistributing to the drivers[39].

There are many dimensionality reduction techniques such as principal component analysis (PCA)[40] . The main difference between PCA and SLR is the interpretability of the results each model produces. For PCA, the output is just a coefficient assigned to each gene to transform each sample's gene expression vector into a reduced vector embedding space. It is difficult to infer biological knowledge or therapeutic targets from the results. On the other hand, our approach can generate subnetwork-level attention maps highlighting the subnetworks that are important for predicting specific cancer phenotypes. It is much more intuitive and informative in terms of biological and clinical applications.

Clinical implication of the decision making process by the attention layer. In our method, the final prediction was made by combining each subnetwork level representation, which can be considered a type of multimodal learning. The multimodal learning approach has been applied to cancer genomics as a tool to integrate heterogeneous data sources such as multi-omics integration[41]. In our model, subnetworks that are optimally defined in terms of PPI network and HGS are considered multimodal units, meaning that each subnetwork level prediction has been independently generated to recognize patient status. This is not a new concept on the clinical bench. Clinicians routinely use heterogeneous sources of information to make well-adjusted decisions[42]. Thus, the human clinician's decision-making process is inherently multimodal. Our model can be considered a computationally well-optimized tool for simulating the decision-making process of human clinicians based on transcriptome data. In addition, explainability is a key challenge for maximizing the utility of a machine learning model[43]. Reducing high dimension transcriptome data into a much smaller but biologically meaningful subnetwork space will be helpful to explain the prediction result in a more acceptable way for both clinicians and patients. Moreover, since the attention layer operates in an instance-wise manner, the attention map of each patient represents the highlights of the importance of each subnetwork customized to each patient, which can be useful for the personalized medicine[44]. Lastly, even though the attention map mostly focused on the subnetworks well-known to be associated with metastasis, some of them (eg, Mitotic spindle and Myogenesis) were relatively not studied for their connection to the metastasis. It suggests that the attention model can be used as a tool for discovering previously unknown network biomarkers, which can lead to a new understanding of cancer biology or new therapeutic targets.

Limitations and future works. As stated in "Biological significance of the subnetworks of PPI network extracted by DeepWalk", the HGS geneset is pre-customized for each cancer phenotype, so the proposed method is essentially tailored only to cancer samples. Although it is designed to act on all cancer tissues, it has only been tested for oral and breast cancers, so its applicability to other cancer types is unknown. To clarify this, future experiments need to be conducted to apply the proposed method to a wide range of cancer samples, such as the pan-cancer project [?] data.

Conclusion

The proposed method is a computational framework that learns subnetwork representations of the transcriptome to discover network biomarkers and determine metastatic potential in early oral tongue squamous cell carcinoma. This method achieved high accuracy in predicting the likelihood of metastasis in two geographically and racially different groups of patients. The robustness of the model and the reproducibility of the discovered network biomarkers show great potential as a tool to diagnose lymph node metastasis in early oral cancer. Our contribution can be summarized as follows .

(1) Developed a method to define optimized subnetworks from a given PPI network using a state-of-the-art graph embedding technique.
(2) Developed a supervised subnetwork representation learning system to successfully reduce the input dimension of transcriptome data by considering complex dependencies between genes, leading to robust prediction models with good performances.
(3) Developed an attention-based classification system to integrate the subnetwork level representations, creating an attention map that highlights important network biomarkers tailored to each patient, leading to feature rankings that significantly reproduced the prior knowledge.

Data availability

TPM-quantified gene expression levels for TCGA-BRCA and TCGA-HNSC were downloaded from http://firebrowse.org/, where the archive name is illuminahiseq_rnaseqv2-RSEM_genes in the mRNASeq section. The PAM50 label of TCGA-BRCA was extracted from https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018. The OTSCC annotation information such as tissue origin and the tumor stage is based on the clinical information of the TCGA-HNSC cohort downloaded from http://firebrowse.org/, where the archive name is Clinical_Pick_Tier1 in the Clinical section. A cleaned dataset is deposited on Zenodo, whose link is "https://zenodo.org/record/5485336#.YTfuPp5Kiqk". All the other related data and source codes are available at https://github.com/mdy89/subnet-learn. Supplementary Tables S1, S2, Supplementary Figs. S1 and S2 are included in the supplementary information section.

Received: 1 June 2021; Accepted: 18 November 2021
Published online: 14 December 2021

References

1. Vigneswaran, N. & Williams, M. D. Epidemiologic trends in head and neck cancer and aids in diagnosis. *Oral Maxillofac. Surg. Clin.* **26**, 123–141 (2014).
2. Wangsa, D. *et al.* Ki-67 expression predicts locoregional recurrence in stage I oral tongue carcinoma. *Br. J. Cancer* **99**, 1121 (2008).
3. Leusink, F. K. J. *Molecular Markers of Lymph Node Metastases in Oral Cancer*. Ph.D. thesis (Utrecht University, 2017).
4. Smeets, A. *et al.* Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and mirnas. *Breast Cancer Res. Treat.* **129**, 767–776 (2011).
5. Prado, K., Zhang, K. X., Pellegrini, M. & Chin, A. I. Sequencing of cancer cell subpopulations identifies micrometastases in a bladder cancer patient. *Oncotarget* **8**, 45619 (2017).
6. Zheng, Y. *et al.* Cox-2 mediates tumor-stromal prolactin signaling to initiate tumorigenesis. *Proc. Natl. Acad. Sci. USA* **116**, 5223–5232 (2019).
7. Safo, S. E. *Design and Analysis Issues in High Dimension, Low Sample Size Problems*. Ph.D. thesis (University of Georgia, 2014).
8. Suchkov, S. & Herrera, A. S. The role of human photosynthesis in predictive, preventive and personalized medicine. *EPMA J.* **5**, A146 (2014).
9. Chiu, A. M., Mitra, M., Boymoushakian, L. & Coller, H. A. Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer. *Sci. Repo.* **8**, 1–14 (2018).
10. Kopper, O. *et al.* An organoid platform for ovarian cancer captures intra-and interpatient heterogeneity. *Nat. Med.* **25**, 838–849 (2019).
11. Dao, P. *et al.* Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics* **26**, i625–i631 (2010).
12. Xiao, H. *et al.* Protein-protein interaction analysis to identify biomarker networks for endometriosis. *Exp. Ther. Med.* **14**, 4647–4654 (2017).
13. Rezaei-Tavirani, M., Rezaei-Tavirani, S., Mansouri, V., Rostami-Nejad, M. & Rezaei-Tavirani, M. Protein-protein interaction network analysis for a biomarker panel related to human esophageal adenocarcinoma. *Asian Pac. J. Cancer Prev.* **18**, 3357 (2017).
14. Yuan, X. *et al.* Network biomarkers constructed from gene expression and protein-protein interaction data for accurate prediction of leukemia. *J. Cancer* **8**, 278 (2017).
15. Gaire, R. K. *et al.* Discovery and analysis of consistent active sub-networks in cancers. *BMC Bioinform.* **14**, S7 (2013).
16. Inavolu, S. M. *et al.* Iodne: An integrated optimization method for identifying the deregulated subnetwork for precision medicine in cancer. *CPT* **6**, 168–176 (2017).
17. Zhang, J., Wang, Y., Tang, J. & Ding, M. Spectral network embedding: A fast and scalable method via sparsity. arXiv preprint arXiv:1806.02623 (2018).
18. Perozzi, B., Al-Rfou, R. & Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710 (ACM, 2014).
19. Lim, S. *et al.* Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods* **110**, 81–89 (2016).
20. Lim, S., Lee, S., Jung, I., Rhee, S. & Kim, S. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Brief. Bioinform.*https://doi.org/10.1093/bib/bby097 *(2018).*
21. Chatr-Aryamontri, A. *et al.* The biogrid interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
22. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
23. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
24. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
25. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
26. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems* 577–585 (2015).
27. Choi, E. et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems* 3504–3512 (2016).
28. Shannon, C. E. A mathematical theory of communication, part II. *Bell Syst. Tech. J.* **27**, 623–656 (1948).
29. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015).
30. Chollet, F. keras. https://github.com/fchollet/keras (2015).
31. Network, C. G. A. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (2012).
32. Network, C. G. A. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576 (2015).
33. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160 (2009).
34. Onoue, T. *et al.* Epithelial-mesenchymal transition induced by the stromal cell-derived factor-1/cxcr4 system in oral squamous cell carcinoma cells. *Int. J. Oncol.* **29**, 1133–1138 (2006).
35. Johnson, D. G. & Schneider-Broussard, R. Role of e2f in cell cycle control and cancer. *Front. Biosci.* **3**, d447–d448 (1998).
36. Kwong, R. A. *et al.* Overexpression of e2f–1 is associated with increased disease-free survival in squamous cell carcinoma of the anterior tongue. *Clin. Cancer Res.* **9**, 3705–3711 (2003).
37. Wu, Y.-D. & Zhou, B. Tnf-α/nf-κb/snail pathway in cancer cell migration and invasion. *Br. J. Cancer* **102**, 639–644 (2010).
38. Tang, D. *et al.* Tnf-alpha promotes invasion and metastasis via nf-kappa b pathway in oral squamous cell carcinoma. *Med. Sci. Monit. Basic Res.* **23**, 141 (2017).
39. Haber, D. A. & Settleman, J. Drivers and passengers. *Nature* **446**, 145–146 (2007).

40. Pearson, K. L. III. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. **2**, 559–572 (1901).
41. Lee, G., Nho, K., Kang, B., Sohn, K.-A. & Kim, D. Predicting Alzheimer-s disease progression using multi-modal deep learning approach. *Sci. Rep.* **9**, 1–12 (2019).
42. Kazi, A. et al. Graph convolution based attention model for personalized disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 122–130 (Springer, 2019).
43. Gunning, D. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web **2** (2017).
44. NCI. *Precision medicine in cancer treatment*. https://www.cancer.gov/about-cancer/treatment/types/precision-medicine (2017).

Acknowledgements

This work was supported by the Post-Genome Program by Ministry of Science, ICT & Future Planning (2014M3C9A3063541), the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT(NRF-2019M3E5D307337511), a grant (DY0002258224) from Ministry of Food and Drug Safety in 2020, National Research Foundation of Korea (https://www.nrf.re.kr/index). This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Author contributions: M.K. developed the method, carried out evaluations, interpreted results, and wrote the manuscript. S.Le. and S.Li. both reviewed the manuscript. D.L and S.K. supervised the project and revised the manuscript.

CompetingInterests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03333-5.

Correspondence and requests for materials should be addressed to D.Y.L. or S.K.