BMC Bioinformatics

RESEARCH Open Access

Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences

Younghoon Kim[1], Tao Wang[2,3], Danyi Xiong[4], Xinlei Wang[4] and Seongoh Park[5*]

*Correspondence:
spark6@sungshin.ac.kr

[1] Department of Industrial and Management Systems Engineering, Kyung Hee University, Yongin, Gyeonggi, Korea
[2] Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, USA
[3] Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA
[4] Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
[5] School of Mathematics, Statistics and Data Science, Sungshin Women's University, Seoul, Korea

Abstract

Early detection of cancers has been much explored due to its paramount importance in biomedical fields. Among different types of data used to answer this biological question, studies based on T cell receptors (TCRs) are under recent spotlight due to the growing appreciation of the roles of the host immunity system in tumor biology. However, the one-to-many correspondence between a patient and multiple TCR sequences hinders researchers from simply adopting classical statistical/machine learning methods. There were recent attempts to model this type of data in the context of multiple instance learning (MIL). Despite the novel application of MIL to cancer detection using TCR sequences and the demonstrated adequate performance in several tumor types, there is still room for improvement, especially for certain cancer types. Furthermore, explainable neural network models are not fully investigated for this application. In this article, we propose multiple instance neural networks based on sparse attention (MINN-SA) to enhance the performance in cancer detection and explainability. The sparse attention structure drops out uninformative instances in each bag, achieving both interpretability and better predictive performance in combination with the skip connection. Our experiments show that MINN-SA yields the highest area under the ROC curve scores on average measured across 10 different types of cancers, compared to existing MIL approaches. Moreover, we observe from the estimated attentions that MINN-SA can identify the TCRs that are specific for tumor antigens in the same T cell repertoire.

Keywords: Multiple instance learning, Instance selection, Primary instance, Sparsemax

Introduction

Multiple instance learning (MIL) is a supervised learning task that includes a special structure called a bag in each entity. In MIL, a set of instances in the same bag and their explanatory variables are observed. Though they share an observed bag-level response (bag label), instances may not have an individual instance-level response that one observes in traditional single-instance learning. Supervised MIL tasks can be classified by types of the label; multiple instance classification (MIC) if the label takes its value on a discrete space, and multiple instance regression (MIR) if on a set of (non-discrete) real numbers. Most MIL applications are MIC, including remote sensing [1, 2], computer vision [3], sentimental analysis [4], and especially biology [5–7]. In the meanwhile, MIR has relatively scarce literature. We refer readers to [8] for more examples in MIC and [9] in MIR. However, as pointed out in [8], benchmark MIL datasets have been still limited to predicting binding sites of molecules, image/text classification in literature. For instances, [8, 10] have conducted a comparative study using typical real data examples so-called Musk [11], Text [12], Speaker [13], Corel [14], Birds [15], Letters [16], and Tiger/Elephant/Fox (TEF) [12] , which other MIL researches have also used [17–23]. In the light of it, there were recent attempts to bring brand-new data under the MIL framework [7, 24]: T cell receptor sequencing data for cancer detection.

The problem to distinguish normal and cancerous tissues/patients has attracted much attention due to its great significance in biomedical fields for cancer prognosis. To address this biological problem, past works have used medical images [25, 26], gene expressions [27–30], and single nucleotide polymorphisms (SNPs) [31–33], etc. As the basis for forming predictive models, the recent appreciation of the roles of the host immunity system in tumor biology has motivated researchers to study T cell receptors (TCRs) [7, 24, 34–36]. [7] predicted the tumor status of patients using TCR sequences. As multiple TCR sequences (instances) are observed together in different T cells in the same patient (tumor or normal), the observations naturally fall into the category of MIL. The authors conducted a benchmarking study to compare MIC algorithms, but they did not treat deep neural networks in depth, thus leaving the performance of neural network models under-explored. Though [37]'s models are included in comparison, they are not complex enough to learn successfully the underlying structure of the TCR data, thus showing unsatisfactory performance. [24] investigated the multiple instance learning task that distinguishes T-cell repertoires between tumor and healthy tissues, but they assumed the standard MIL assumption which does not consider relative importance of instances. [34] proposed a deep learning model called DeepCat that utilizes tumor-specific or non-cancer TCRs, but DeepCat ignores the bag-structure as well as possibly different contributions of TCRs. Another deep learning model, named DeepLION, is proposed by [38], but DeepLION cannot completely remove unimportant instances in explaining the bag labels.

Towards bridging this gap, we introduce a novel neural network model, titled MINN-SA (Multiple Instance Neural Network based on Sparse Attentions), for cancer detection based on TCR sequences. The salient part of the proposal is the sparse attention structure that flexibly drops out uninformative instances, thus rendering model interpretation more achievable. Recent works from [39–42] also employed an attention structure in their neural networks. However, their attention scores are dense, meaning none of them is exactly 0, so that irrelevant information for bag classification could be involved in extracted features. Moreover, the sparsity pattern considered in [42] is based on a simple heuristic that keeps top-N instances with the largest scores, which lacks of optimality and stability in results. In contrast, equipped with the sparsemax function by [43], MINN-SA adaptively discovers the pattern of sparsity in attention scores. This flexibility is also beneficial to predictive performance, which can be further enhanced by adding the skip connection by [44]. With this state-of-the-art architecture, we achieve the highest overall AUC scores both in balanced and imbalanced datasets in the cancer detection problem. Our main contributions are summarized as follows:

- We propose a sparse attention-based neural network that drops out uninformative instance per bag, achieving model interpretability.
- MINN-SA outperforms comparative methods in cancer detection based on TCR sequences, achieving the highest AUC scores both in balanced and imbalanced datasets.

The remainder of this paper is organized as follows. in "Methods" Section gives the details of our method. in "Result" Section presents the experimental setup and results of TCR dataset. Finally, we end in "Discussion" Section with a summary and discussion.

Methods

Multiple instance learning for cancer detection using TCR sequences

In MIL, an observational unit is a bag (a sample). In bag i ($i = 1, \ldots, n$), each of multiple instances is characterized by a vector x_{ij} of p features in $\mathbb{R}^p$, $j = 1, \ldots, m_i$, and a single label y_i is tagged on it. The goal of MIL is to estimate a function f that predicts a bag-level label from a set of instances. Note that this function takes a set of instances as an input, so it should be adaptive to different number of instances for each bag.

In our application, tissue samples are collected either from normal or cancer patients and a set of TCR sequences are identified in each sample by using next generation sequencing technologies [7, 45]. The main task is to determine whether a tissue is cancerous or not based on its TCR sequences. Here, we treat the tissue type as a bag-level label and the set of TCR sequences as multiple instances, all of which are contained together in a patient, or a bag. Under this context, we focus on a binary MIC task and thus restrict a bag label in $\{0, 1\}$; for example, 0 (a negative bag) is non-cancer and 1 (a positive bag) is cancer. The binary classification could be easily extended to the multi-class setup simply by changing a score function to the softmax function and a loss function to the multi-class cross-entropy. To associate a series of unlabeled instances to a bag label, we adopt the primary instance assumption. In other words, it is assumed that a portion of instances, or primary instances, can explain the label while the remaining instances, or non-primary instances, are irrelevant to it. In our contexts, those selected TCRs represent specialized T cells that the human immune system develops against the tumor cells. More specifically, the TCRs recognize the tumor-associated antigens [46, 47] or tumor neoantigens [48–50] presented on the surface of the tumor cells, which are markers of the tumor cells and distinguish them from normal epithelial cells.

We utilize the sparse attention in the multiple instance neural networks to detect such meaningful TCRs. The proposed layer selectively reflects instances' information to an extracted feature vector for final classification. The sparsity enhances the classifier's performance and the explainability of classification results. Moreover, the proposed method is computationally efficient. The details of the proposed method are presented in the following subsections.

Numeric embeddings of TCR sequences

We describe the process of numeric embedding of TCR sequences carried out in [7]. TCR sequence is a text string comprising a series of amino acids, which is actually a text string. According to [51], each amino acid can be converted to five Atchley (latent) factors that sufficiently represent the attributes of the amino acid. This conversion of a set of TCR sequences returns a Atchley matrix, which is inserted into the TCR encoding algorithm [36, 45]. The key part of the algorithm named TESSA is a stacked auto-encoder that takes Atchley matrices and returns a set of vectors of fixed length (30 dimensions) determined by the number of neurons in the bottleneck layer. The encoded numeric representation facilitates the usage of TCR sequence data. Most MIL algorithms are only compatible with the numeric type of data, especially for those methods calculating a distance between instances (or bags). We refer to [7] for more information about the data and processing details. In particular, we do not claim any original contribution to the data.

We mention that there are other sequence processing methods such as [52]. The auto-encoder TESSA offers satisfactory performance in our datasets. However, practitioners can always experiment on other methods for potentially better performance.

Neural networks based on sparse attention

We propose a neural network based on sparse attention to solve multiple instance classification problems. The overall structure of our model is illustrated in Fig. 1a, and we give details of each component in our neural network below.

The input size is fixed by $m^* \times p$ for each of n bags where m^* is the hyperparameter that decides the number of instances to be included in modeling. If a bag has fewer instances than m^*, then an empty part is padded by zeros. Some of bags have larger instance size than m^*. In the case, we determine the first m^* instances as our bag instances. Also, we keep using this masking information across the whole process. For ease of handling, one can easily set m^* by the largest bag size in data. In "Results" Section, we conduct a sensitivity analysis for choice of the size m^*.

Each layer consists of $m^* \times p$ neurons, but they are not fully connected to the activation functions in the previous layer as in the usual fully-connected layer. Instead, we build the full connections between neurons within each instance, but not across different instances (see Fig. 1b). The weights for a fully connected layer are shared across the instances to handle the variable number of instances in each bag. Hence, the two consecutive layers are connected by a $p \times p$ weight matrix (or $(p + 1) \times (p + 1)$ if a bias term is included). The network deals with the non-linearity in data with the rectified linear unit (ReLU) [53]. Note that we used fully connected layers rather than convolution layers because the input features are not locally correlated between adjacent features.

We employed the dropout [54] and skip connection [44] to enhance the predictive performance. The dropout layer is attached after each locally fully-connected layer appears. It randomly forces the output variables to be zero with probability 0.3 while training the network. It is well known that this layer prevents the complex neural network from overfitting on training data. The residual learning framework eases the training of networks with stacked layers. The skip connections in the framework explicitly let the layers fit a residual mapping instead of hoping stacked layers to directly fit a desired underlying mapping denoted by $\mathcal{H}(x)$. We let the stacked nonlinear layers fit another mapping of $\mathcal{F}(x) := \mathcal{H}(x) - x$. The bypassing path for a gradient mitigates the vanishing gradient issue in neural networks. The empirical studies show that it is easier to optimize the residual mapping than to optimize the original. Thus, we employ the skip connection to enhance the performance of the locally fully-connected neural network with multi-layers. The effect of residual connection is demonstrated in "Ablation study" Section.

The attention layer combines the attention weights $\{\alpha_j\}_{j=1}^{m^*}$ returned by the sparsemax function [43] with the feature matrix $Z \in \mathbb{R}^{m^* \times p}$ obtained from the last network layer, ending up with a weighted feature vector $\tilde{z} = \sum_{j=1}^{m^*} \alpha_j z_j$ where z_j is the j-th row of Z and $\tilde{z} \in \mathbb{R}^{1 \times p}$ (see Fig. 1c). The commonly used softmax function does not pursue exact zeros in the output, but the sparsemax function permits zeros in it. Let $\Delta^{K-1} := \left\{p \in \mathbb{R}^K | 1^T p = 1, p \geq 0\right\}$ be the $(K-1)$-simplex. The sparsemax is a function mapping vectors in $\mathbb{R}^K$ to probability distributions in Δ^{K-1}:

$$\text{sparsemax}(z) := \underset{p \in \Delta^{K-1}}{\text{argmin}} \|p - z\|^2. \tag{1}$$

In our context, the sparsemax function takes attention scores and encourages some of them to be zero if they do not exceed some thresholding value. As shown in [43], the threshold is adaptively determined from the scores, not manually by users like hard/soft-thresholding functions. The manual adjustment of the attention threshold uses an additional hyperparameter to determine redundant attention scores forced to be zero. If some of the attention scores are less than the threshold, we manually set the attention score equal to zero. A grid search algorithm usually optimizes the hyperparameter and depending on the grid, the optimal value can vary. However, adaptive thresholding with the sparsemax function finds theoretically optimal value without heuristic search, leading to better generalization performance of the prediction model by training with the optimal hyperparameter. Moreover, adaptive thresholding is computationally more efficient than manual thresholding because it requires no validation procedure to optimize the threshold value.

For each bag, two scores, which correspond to classes of tumor and normal tissues, are converted to probabilities, which is easier to interpret. The transition is done by the sigmoid function. The input of classification layer is batch-normalized feature vector $\tilde{z}^* \in \mathbb{R}^{1 \times p}$ and the output is a scalar. The aggregated feature vector $\tilde{z}$ may have different means and variances across components, which could hamper the network from learning data stably. Thus, we apply the batch normalization [55] to overcome this difficulty.

We should mention the difference between the proposed attention structure and the others from the previous works. Derived from [39–42], the attention scores are strictly larger than zero. Thus, it is challenging to discard unimportant or non-primary instances that have little to do with bag classification. In contrast, MINN-SA allows the attention scores to be sparse or have many zeros, meaning that strictly positive weights are only given to the primary instances responsible to bag classification. This is an attractive instance selection because it transparently shows which instances are chosen in model training and thus facilitates one to interpret classification results only depending on the selected instances. Moreover, our selection is more advanced than the elementary top-N rule [42] and free from tuning parameters often required in thresholding operators.

Table 1 The sample size for each cancer type

Cancer	BRCA	DLBC	ESCA	KIRC	LUAD
Balance	404	90	332	404	404
Imbalance	225	225	225	225	225
Cancer	**LUSC**	**OV**	**SKCM**	**STAD**	**THYM**
Balance	404	404	404	404	216
Imbalance	225	225	225	225	225

Half of the samples are tumor tissue samples for the balanced case, while about a tenth of them are for the imbalanced case

Computation

We implement the method with a deep learning framework, PyTorch 1.8.1 [56] with CUDA 11.1 [57]. The computation system consists of Intel i9-10900 CPU, 32GB RAM, and RTX 3090 GPU. It takes 0.01 seconds for each epoch in the training stage. We stop the training after 100 epochs and determine the best performing model during the training procedure. Therefore, the whole training process approximately takes 1 second in our computation system setting. Referring to the computational time results from [7], the proposed method is more efficient than other multiple instances learning methods in computation.

Results

In this section, we showcase our novel model in distinguishing tumor and normal tissue samples from different types of cancers. Different existing methods are compared against our model in terms of predictive accuracy. Moreover, we provide the instance selection result derived from the estimated attention weights. Lastly, we conduct an ablation study to examine contributions of individual components to our model.

The real datasets we analyze are from The Cancer Genome Atlas (TCGA) database, in whole generated by the TCGA Research Network[1]. Normal and healthy tissues are collected for 10 types of cancers listed in Table 1. As mentioned in "Numeric embeddings of TCR sequences" Section, these samples are processed through the next generation sequencing, TCR reconstruction techniques, and TCR encoding algorithms so that the genomic data from the donors are converted in numeric vectors.

Setting

To figure out how they behave in different scenarios, we test the comparative models under two scenarios: (1) the balanced case and (2) the imbalanced case. The former has an equal number of positive (tumor) and negative (normal) bags, while the latter sets about 10% of bags to be positive. The balanced data is commonly used and often preferred in machine learning literature. The other one is to capture characteristics of large population cancer screening where few patients have tumors [7, 58, 59]. To create a dataset for each cancer type, we subsample normal and tumor tissues to keep the aimed proportion of positive (tumor) bags. The sample size of each dataset is tabulated in Table 1.

[1] https://www.cancer.gov/tcga

Table 2 Abbreviations of 18 comparative methods including the proposed method "MINN-SA" and their original references

18 comparative methods		
ADeep [22]	BoW [10]	CCE [60]
CkNN [61]	EMD-SVM [62]	EMDD [63]
SI-kNN [8]	MI-SVM [64]	miGraph [65]
MILBoost [66]	MILES [67]	MInD [68]
mi-Net [69]	MI-Net [69]	mi-SVM [64]
NSK-SVM [70]	SI-SVM [71]	MINN-SA

For model training and validation, we conduct 10-fold cross validation (CV) to split training and testing datasets. On the testing dataset, the Area Under the Curve (AUC) of each method is calculated based on the Receiver Operating Characteristic (ROC) curve. We follow the same experimental design in the preceding work [7] for fair comparison.

Benchmarking on cancer detection

To benchmark the proposed model, 18 MIC methods are considered, which are listed in Table 2. We refer to Section 3 of [7] for a detailed exposition about these methods.

Figure 2 shows average AUC values by methods for the ten cancer types. Remarkably, MINN-SA dominates all methods in most of cancer types. Out of 10, MINN-SA wins in 7 types (BRCA, ESCA, KIRC, LUAD, OV, SKCM, STAD) for the balanced case and in different 7 types (BRCA, ESCA, KIRC, LUSC, OV, SKCM, THYM) for the imbalanced case. KIRC, LUAD, SKCM, LUSC are the four most immunogenic cancer types [72], meaning they have a lot of T cell infiltrations. It makes sense these cancer types are among the ones [72] for which our model performs the best, which investigates TCRs of T cells for classification. Generally, when the class distribution is balanced, each model shows more stable performance [72]. The gap between MINN-SA and the second best method is considerably big in BRCA and KIRC datasets for the imbalanced case.

Figure 3 shows the boxplots of all methods in an ascending order of median AUC. MINN-SA tops in both balanced imbalanced cases, with medians 74.40 and 81.80, respectively, followed by EMD-SVM with 72.20 and 75.40. To demonstrate the superiority of the proposed method over comparative methods, we conducted Wilcoxon signed-rank test on the best and second-best methods with rank statistics. Our proposed method achieves the best performance in terms of average rank over cancers. The average ranks of the proposed method are 2.3 and 2.2 in balanced and imbalanced cases, respectively. The second best method is EMD-SVM, whose average ranks are 2.7 and 5.2. In the balanced case, the p-value is 0.0372, and the p-value for the imbalanced case is 0.0043. Our proposed method outperforms other approaches in both cases with statistical significance. For detailed AUC values averaged over cancer types, See Table 3.

Moreover, the proposed MINN-SA enjoys interpretable results from selected instances, which EMD-SVM does not afford. Also, MINN-SA is not very sensitive to class imbalance. Contrary to it, most MIL methods have degraded performance for the imbalanced case, calling for further modifications; for example, data generation or modifying the loss function of the classification model [73–75].

Table 3 The AUC for various methods averaged over cancer types

Method	**ADeep**	**BoW**	**CCE**	**CkNN**	**EMD-SVM**
Balance	66.44 (8.18)	68.74 (1.60)	69.89 (1.17)	51.97 (1.85)	**72.67 (1.23)**
Imbalance	**70.48 (15.33)**	64.28 (4.90)	61.91 (3.82)	52.33 (3.65)	**70.19 (2.04)**
Method	EMDD	mi-Net	mi-SVM	MI-SVM	miGraph
Balance	56.98 (3.81)	67.66 (9.03)	67.08 (1.48)	67.87 (1.58)	63.88 (1.49)
Imbalance	62.67 (4.67)	63.79 (17.66)	64.69 (3.50)	66.55 (2.94)	59.50 (2.62)
Method	MILBoost	MILES	MInD	MINet	MINN-SA
Balance	50.83 (2.31)	68.94 (1.40)	67.03 (1.66)	58.07 (14.25)	**73.88 (8.82)**
Imbalance	59.89 (2.45)	66.05 (3.90)	65.94 (2.85)	35.28 (16.10)	**79.20 (17.18)**
Method	**NSK-SVM**	**SI-kNN**	**SI-SVM**	**Average**	
Balance	**70.00 (1.31)**	66.50 (1.48)	66.72 (1.61)	64.78 (12.39)	
Imbalance	63.43 (3.12)	57.63 (5.16)	65.05 (2.94)	61.74 (16.80)	

The number in parenthesis is the average standard deviation of AUC from 10-fold CV across cancer types. The last method ("Average") shows the mean and the standard deviation of AUC values from all methods except MINN-SA. Numbers in boldface are the best 3 methods in each of balanced and imbalanced cases

Attention of instances

In Fig. 4, attention weights of instances are displayed in heatmap. The heatmap is given in a $n \times m^*$ matrix form where the weights are colored in blue-white spectrum and the masking area (no instances) in gray. The visual inspection demonstrates that the attention weights estimated by MINN-SA are sparser than those by the softmax-based method. For the case based on the softmax function, all instances have strictly positive weights, which is depicted by smooth patterns in the heatmap. Consequently, the dense weights makes the aggregated feature vector from the attention layer depend on redundant information for classification. On the other hand, MINN-SA forces the attention weights of insignificant instances to be exactly zero, which makes decisions of MINN-SA independent of them. In our data, the selected instances are likely the TCRs that are specific for tumor antigens, such as tumor neoantigens or tumor associated antigens, presented on the surface of the tumor cells.

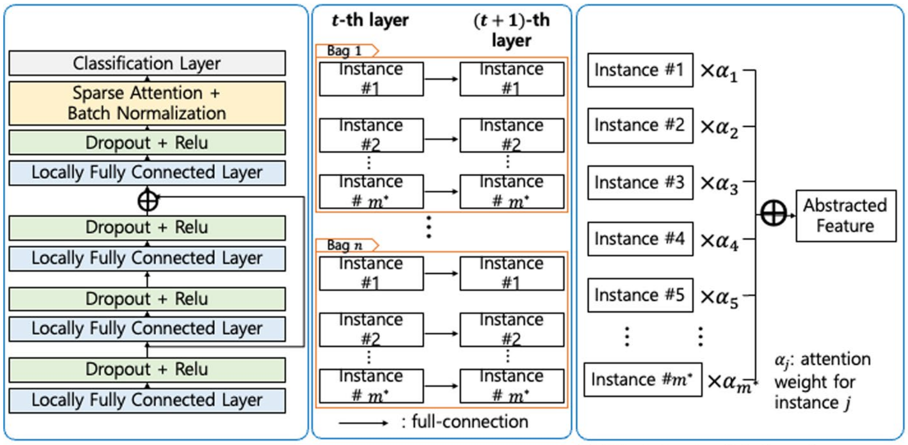
To prove this claim, another validation experiment is conducted. We extract all virus-related TCRs from http://friedmanlab.weizmann.ac.il/McPAS-TCR/. We embed the 31315 TCRs in 30-dimensional space via the TCR encoder TESSA mentioned in "Numeric embeddings of TCR sequences" Section. Then, we measure Euclidean distances between all pairs of TCRs from the new database and TCGA database. By averaging the distances for each TCR from tumor/normal samples of TCGA, we can compare how close each TCR is to the virus-related ones. We focus on the three cancer types (ESCA, OV, STAD) where the classifiers considered show best performance. We show boxplots to compare distances. Recall that MINN-SA estimated weights of instances, so we know which are primary/non-primary.

In Fig. 5 above, tumor bags (Label=1) have larger distances in primary instances compared with non-primary instances (all p-values close to 0), which implies the primary instances responsible for bag classification are not virus-specific, but tumor-specific. This is opposite in normal samples (Label=0), where the primary instances are specific to non-tumor immunologic events, such as virus infection but of course

Table 4 Comparison of the four models in the ablation study

	FC	Skip	Sparse	Proposed
Skip connection	×	✓	×	✓
Sparsemax layer	×	×	✓	✓
Balance	66.44	69.61	70.33	73.87
Imbalance	70.47	73.79	75.61	79.19

Structural difference is checked in the second and third rows where "×" means absence of such structure and "✓" means presence of it. Their average AUC values are summarized in the last two rows


Classification Layer
Sparse Attention + Batch Normalization
Dropout + Relu
Locally Fully Connected Layer
⊕
Dropout + Relu
Locally Fully Connected Layer
Dropout + Relu
Locally Fully Connected Layer
Dropout + Relu
Locally Fully Connected Layer
t-th layer
$(t+1)$-th layer
Bag 1
Instance #1
Instance #1
Instance #2
Instance #2
Instance # m^*
Instance # m^*
Bag n
Instance #1
Instance #1
Instance #2
Instance #2
Instance # m^*
Instance # m^*
: full-connection
Instance #1 $\times\alpha_1$
Instance #2 $\times\alpha_2$
Instance #3 $\times\alpha_3$
Instance #4 $\times\alpha_4$
Instance #5 $\times\alpha_5$
Instance #m^* $\times\alpha_{m^*}$
⊕
Abstracted Feature
α_j: attention weight for instance j

(a) An architecture of the proposed network

(b) Layer connection

(c) Calculating input features for classification layer by weighted average of instance features with attention weights

Fig. 1 Description of the proposed neural network

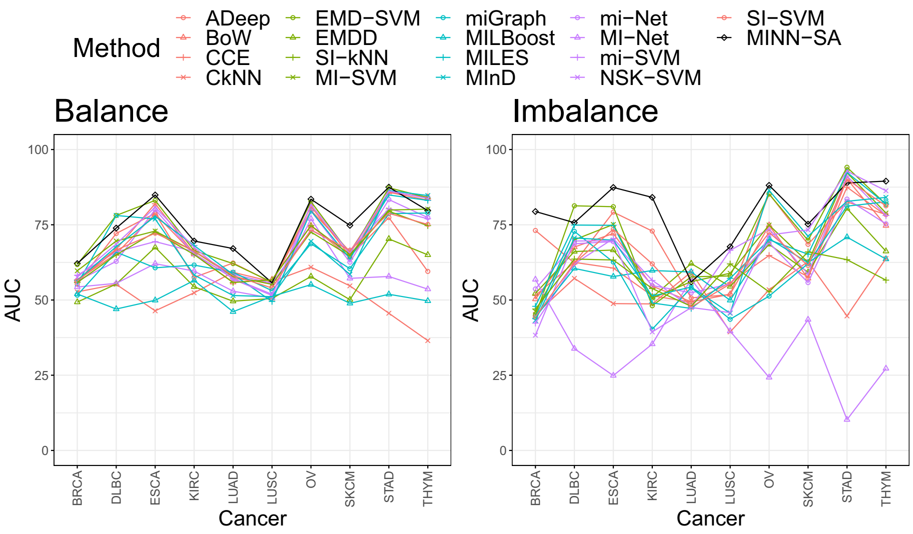


Fig. 2 The AUC of all models in comparison across different cancer types. Each point is the average AUC values over 10-fold CV

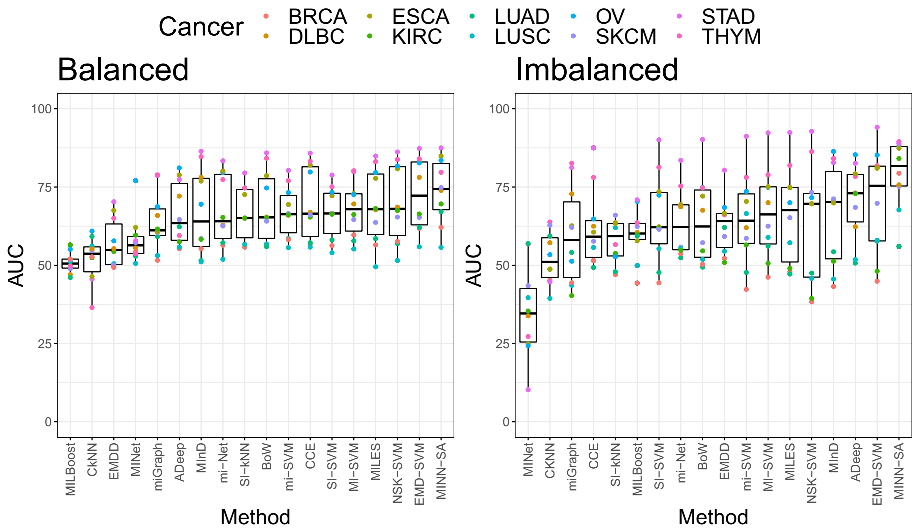


Fig. 3 Boxplots of the AUC values across different methods in comparison. They are ordered by median AUC. Each point is the average AUC values over 10-fold CV and different colors are used to show types of cancer

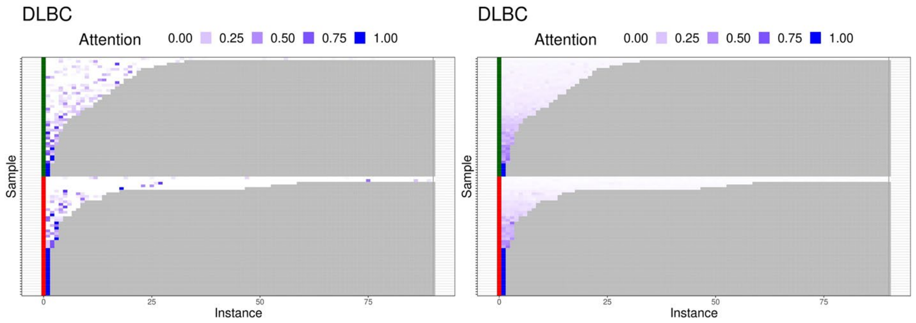


Fig. 4 Heatmap of estimated attention weights via the sparsemax (left) and softmax (right) function. Samples are reordered by the number of instances and the masking area (no instances) is colored in gray. Red and green labels indicate tumor and normal samples, respectively. Here, we show the balanced case for DLBC cancer data

other events as well. As a result, this additional analysis proves well the performance of instance selection using MINN-SA, which is directly related to its interpretability.

Extracted features

Figure 6 shows the extracted features before the last classification layer. The heatmap is given in a $n \times p$ matrix form colored in a blue-yellow spectrum. It can be seen that the extracted features using the sparsemax function (left) are more activated than using the softmax function (right). Hence, the difference between the features in each observation of the sparsemax case is more distinct than the softmax case. The results demonstrate that the extracted features using the sparsemax function are more informative to characterize the characteristics of each sample. We believe that the sparsity in the proposed attention structure distinguishes the instances responsible for the bag classification so that the aggregated features can accurately discriminate bags in the classification layer.

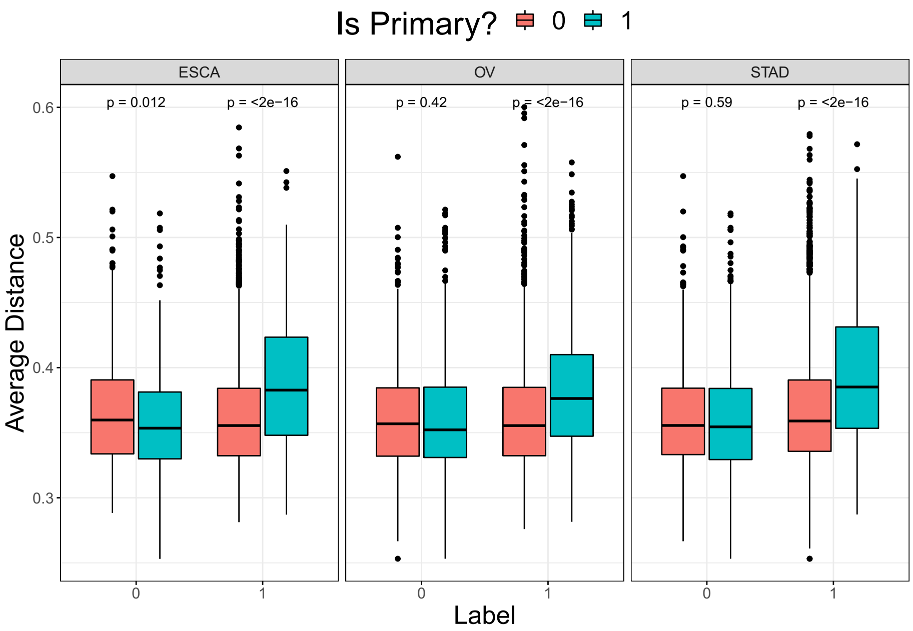


Fig. 5 Boxplot of Euclidean distances (*y*-axis) averaged across virus-related TCRs. In *x*-axis, we indicate the bag label (1:tumor, 0:normal). The mean difference between primary/non-primary instances are tested by Wilcoxon rank-sum test, whose (raw) p-values are given

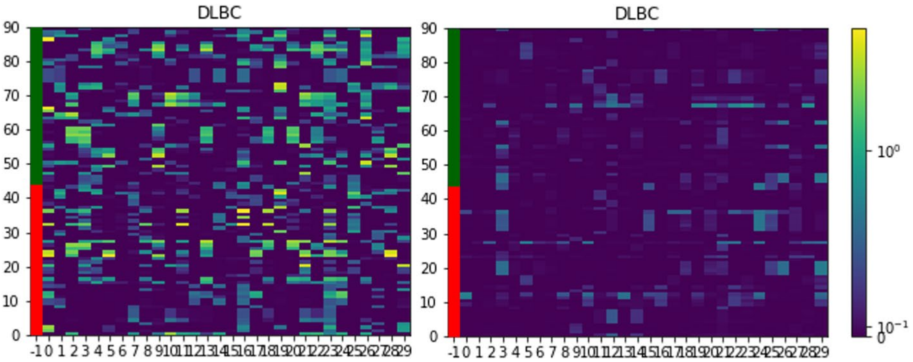


Fig. 6 Heatmap of extracted features via the sparsemax (left) and softmax (right) function. Values are min-max normalized and transformed in log-scale. Red and green labels indicate tumor and normal samples, respectively. Here, we show the balanced case for DLBC cancer data

Ablation study

Firstly, we perform an ablation study to measure contributions of the two components to our neural network: (1) the skip connection and (2) the sparsemax function. Thus, we set a baseline model, denoted by "FC" (short for "fully-connected"), by removing the two components from MINN-SA, and we add each component one after another to "FC". This leads to the four comparative models shown in Table 4. Note that "FC" is the model used in [22], and "Proposed" is MINN-SA.

Figure 7 shows AUC values of the four models for different cancer types. "Proposed" outperforms the others in most cases; otherwise it makes a close second. The superiority of "Proposed" is less distinct in the imbalanced case, but it always takes the first

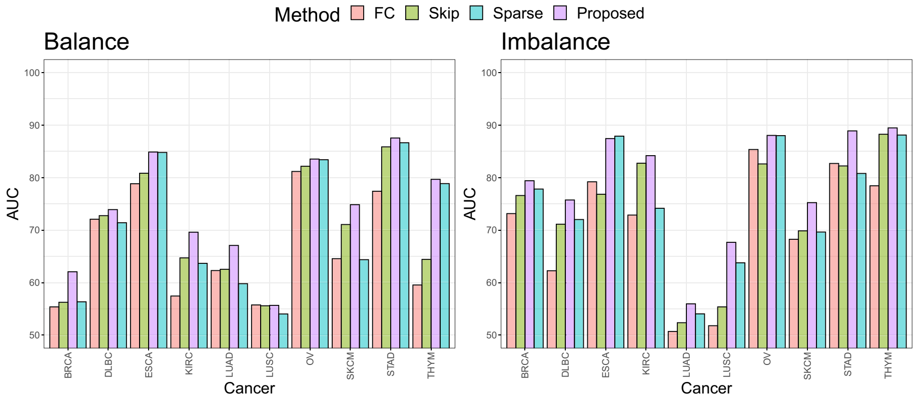


Fig. 7 The AUC of four models across different cancer types. Each bar denotes the average AUC values over 10-fold CV

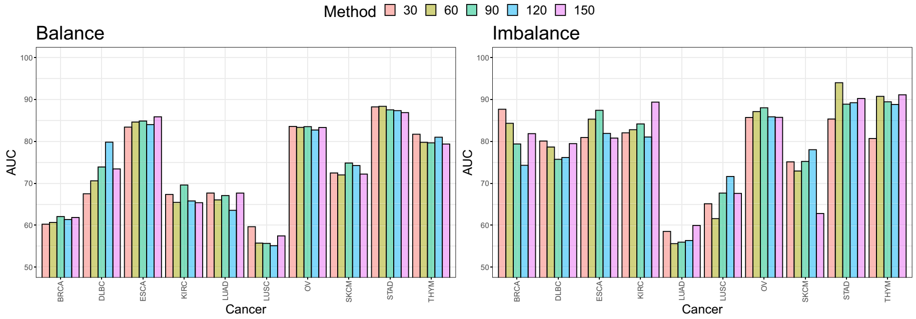


Fig. 8 The AUC of the proposed models having different capacity (bag size) of instances across different cancer types. Each bar denotes the average AUC values over 10-fold CV

or second place, and thus achieves the highest AUC value 79.19 in average followed by 75.61 (see Table 4). These results lend strong support to the proposed method against the existing multiple instance neural network model by [22], a fully-connected neural network based on the softmax function without the skip connection. According to the known phenomenon of immunodominance in the field of immunology [76, 77], not all instances own necessary information and thus such instances would be better removed for classification by the sparse attention structure. The skip connection also proves valuable for this specific application. Interestingly, the performance is the best when both skip connection and sparsemax function are utilized together. This phenomenon aligns with the previous study [78] that shows combining regularization tricks can achieve the best classification performance.

In the following ablation study, we assess the sensitivity to the maximum number of instances per bag. We have tried different maximum numbers by $m^* = 30, 60, \ldots, 150$ and their AUC values are reported in Fig. 8.

The results show that there is no significant difference in classification performance according to the hyperparmeter m^*. This implies that the absolute amount of data increases as m^* increases, but useful information learned for classification could be limited. However, as shown in Fig. 9, bags with more than 30 instances belong to the

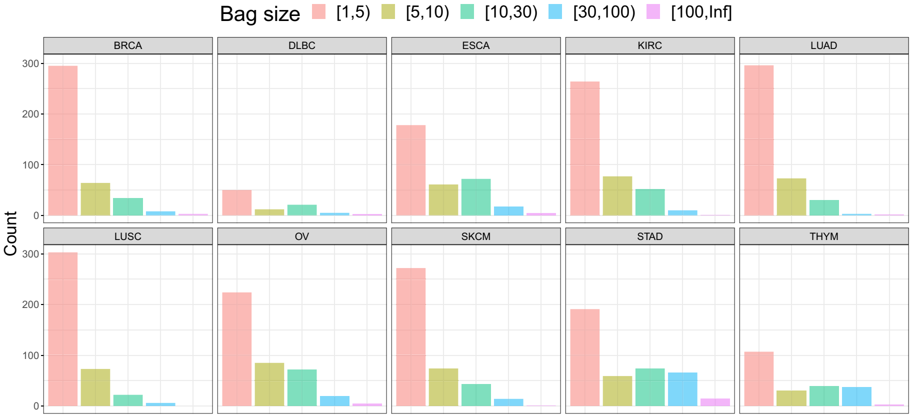


Fig. 9 The distribution of bag sizes in the balanced dataset. Most of bags have less than 5 instances, expressing heavy-tailed features

minority of cases, calling for caution about generalization of this result to other applications. Hence, one should take characteristics of data in hand into account to decide which range of m^* would be explored. The bag size m^* is a hyperparameter for the machine learning model and can be optimized by validation procedures such as k-fold cross-validation.

Discussion

This paper shows that MINN-SA has improved performance on most cancer types compared to the existing methods in predicting tumor vs. normal tissue samples using TCRs. The results imply that a deep neural network is suitable for multiple instance learning. In contrast, traditional statistical approaches may not work that well, and this is because the deep learning approaches can capture very complicated bag-instance structures. The flexible structures of deep neural networks reflect the bag-instance information efficiently through purely data-driven approaches. For statistical approaches, such bag-instance relationships have to be assumed and sophisticatedly specified. We believe the hand-made specifications are vulnerable to data variations such as cancer types.

The prediction performance can be varied depending on data distributions which are different for each cancer type. The proposed method describes some cancer data distributions well, but in some other cases, it shows worse performance than other methods. It is closely related to the “No Free Lunch” theorem in machine learning, [79]; that is, no algorithm can always perform better than others for any data distribution.

The proposed method achieves better performance in imbalanced cases. We believe that the sparse attention leads to the improvement. When the data are imbalanced, it is harder to determine the appropriate decision boundary for classification, and the boundary between classes could be vague in empirical cases. The sparse attention removes noisy instances from blurring the decision boundary. We conjecture that such feature improvement effects are more significant in the imbalanced cases.

Setting m^* as a hyperparameter is closely related to computational efficiency. Figure 9 of the revised manuscript shows that most bags have 30 or fewer instances. In this situation, if m^* is set large, most of the bags have redundant padding instances, resulting in inefficiency in both memory and computation. Thus, our model is designed to allow the adjustment of m* according to an actual situation. The performance comparison results on the different m^* values can be seen in Fig. 8 of the revised manuscript. Overall, the performance difference does not seem to be significant, and the performance variation pattern is unclear. Therefore, it is recommended to optimize performance by choosing an appropriate m^*. Note that users can set it to the maximum instance number if they do not wish to tune m^*.

Potentially, the proposed model can be applied to various MIC problems where instances are naturally arranged in sets and weakly annotated. The applications include biology and chemistry, computer vision, document classification, web mining, reinforcement learning, speech recognition, and time series classification [8]. Specifically to biology and bioinformatics, bags can consist of complex chemical or biological entities; e.g. protein-protein interactions [80–82] where a bag can be defined between two objects (proteins or protein complexes) and instances defined by pairwise combinations of sub-units in each object. Thus, a seemingly non-MIL problem can be reformulated into a MIL problem. We conjecture that MINN SA might be adapted to other prediction problems in bioinformatics, such as (lncRNA-miRNA association prediction [83]; drug-disease association prediction [84, 85]). Certainly, there is ample space for future research.

The instance selection procedure by the sparse attention could be considered a regularization technique commonly used in statistical learning [86]. The LASSO [87] is a representative method to select essential features with sparsity. Regression coefficients of LASSO have non-zero or exact zero values by a shrinkage constraint. The features with non-zero coefficients affect the response variations, but the features with zero coefficients have no influence. The sparsity removes redundant feature information in calculating responses. The proposed method and the LASSO share a similar concept of selecting important information from data. The difference is that LASSO selects important features, but the proposed method selects important instances in a bag. The relationship between the sparse and basic attention for MIL is demonstrated by the relationship between LASSO and Ridge regression [88]. Ridge regression is also a shrinkage method, but the coefficients are not forced to be exact zeros. Thus, it is hard to interpret the Ridge regression results regarding feature selection. In the MIL problem, previous methods based on attention are inappropriate for explaining classification results because all the attention values are non-zero.

Acknowledgements
In loving memory of our late friend Ze Zhang, who was a great company and colleague, we would like to express our deepest gratitude for her kindness and support.

Authors Contribution
All authors read and approved the final manuscript.

Funding
Younghoon Kim is partially supported by NRF-2020R1G1A1101853, Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)), and Translational Research and Development Program on Smart Rehabilitation Exercises under Grant TRSRE-Eq01; in part by the National Rehabilitation Center, Ministry of Health and Welfare, South Korea, Tao Wang and Xinlei Wang are funded by NIH (R01CA258584), Seongoh Park is supported by the Sungshin Women's University Research Grant of H20200131.

Availability of data and materials
All Python code and the pre-processed datasets used in the paper are available on GitHub (https://github.com/ykim-code/MINN-SA.git)

Declarations

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Received: 18 August 2022 Accepted: 26 October 2022
Published online: 08 November 2022

References

1. Wang Z, Radosavljevic V, Han B, Obradovic Z, Vucetic S. Aerosol optical depth prediction from satellite observations by multiple instance regression; 2008. pp. 165–176 .
2. Trabelsi M, Frigui H. Robust fuzzy clustering for multiple instance regression. Pattern Recogn. 2019;90:424–35.
3. Sun M, Han TX, Liu M-C, Khodayari-Rostamabad A. Multiple instance learning convolutional neural networks for object recognition. In: 2016 23rd International Conference on Pattern Recognition (ICPR); 2016. pp. 3270–3275.
4. Angelidis S, Lapata M. Multiple instance learning networks for fine-grained sentiment analysis. Trans Assoc Comput Linguist. 2018;6:17–31.
5. Bandyopadhyay S, Ghosh D, Mitra R, Zhao Z. Mbstar: multiple instance learning for predicting specific functional binding sites in microrna targets. Sci Rep. 2015; 5(1).
6. Gao Z, Ruan J. Computational modeling of in vivo and in vitro protein-DNA interactions by multiple instance learning. Bioinformatics. 2017;33(14):2097–105.
7. Xiong D, Zhang Z, Wang T, Wang X. A comparative study of multiple instance learning methods for cancer detection using t-cell receptor sequences. Comput Struct Biotechnol J. 2021;19:3255–68.
8. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recogn. 2018;77:329–53.
9. Park S, Wang X, Lim J, Xiao G, Lu T, Wang T. Bayesian multiple instance regression for modeling immunogenic neoantigens. Stat Methods Med Res. 2020;29(10):3032–47 (**PMID: 32401701**).
10. Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artif Intell. 2013;201:81–105.
11. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89(1):31–71.
12. Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. In: Becker S, Thrun S, Obermayer K, editors. Adv Neural Inf Process Syst, vol. 15. Vancouver, British Columbia, Canada: MIT Press; 2003.
13. Sanderson C, Lovell BC. Multi-region probabilistic histograms for robust and scalable identity inference. In: Tistarelli M, Nixon MS, editors. Adv Biom. Berlin, Heidelberg: Springer; 2009. p. 199–208.
14. Chen Y, Bi J, Wang JZ. Miles: multiple-instance learning via embedded instance selection. IEEE Trans Pattern Anal Mach Intell. 2006;28(12):1931–47.
15. Briggs F, Fern XZ, Raich R. Rank-loss support instance machines for miml instance annotation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12. Association for Computing Machinery, New York, NY, USA; 2012. pp. 534–542.
16. Frey PW, Slate DJ. Letter recognition using holland-style adaptive classifiers. Mach Learn. 1991;6(2):161–82.
17. Kim M, Torre FDL. Multiple instance learning via gaussian processes. Data Min Knowl Disc. 2014;28(4):1078–106.
18. Cheung P-M, Kwok JT. A regularization framework for multiple-instance learning. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06. ACM, New York, NY, USA; 2006. pp. 193–200.
19. Raykar VC, Krishnapuram B, Bi J, Dundar M, Rao RB. Bayesian multiple instance learning: Automatic feature selection and inductive transfer. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08. Association for Computing Machinery, New York, NY, USA; 2008. pp. 808–815.
20. Bergeron C, Moore G, Zaretzki J, Breneman CM, Bennett KP. Fast bundle algorithm for multiple-instance learning. IEEE Trans Pattern Anal Mach Intell. 2012;34(6):1068–79.
21. Cheplygina V, Tax DMJ, Loog M. Multiple instance learning with bag dissimilarities. Pattern Recogn. 2015;48(1):264–75.
22. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Dy J, Krause, editors. Proceedings of the 35th International Conference on Machine Learning, vol. 80; 2018. pp. 2127–2136.
23. Asif A, ul Amir Afsar Minhas F. An embarrassingly simple approach to neural multiple instance classification. Pattern Recogn Lett. 2019;128:474–9.
24. Ostmeyer J, Christley S, Toby IT, Cowell LG. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. Cancer Res. 2019;79(7):1671–80.
25. Saba T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. J Infect Public Health. 2020;13(9):1274–89.
26. Yan R, Zhang F, Rao X, Lv Z, Li J, Zhang L, Liang S, Li Y, Ren F, Zheng C, et al. Richer fusion network for breast cancer classification based on multimodal data. BMC Med Inform Decis Mak. 2021;21(1):1–15.
27. Lu Y, Han J. Cancer classification using gene expression data. Inf Syst. 2003;28(4):243–68.
28. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, Li L. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC Genomics. 2017;18(1):1–13.

29. Verda D, Parodi S, Ferrari E, Muselli M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. BMC Bioinform. 2019;20(9):1–13.
30. Mostavi M, Chiu Y-C, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics. 2020;13(5):1–13.
31. Hajiloo M, Damavandi B, HooshSadat M, Sangi F, Mackey JR, Cass CE, Greiner R, Damaraju S. Breast cancer prediction using genome wide single nucleotide polymorphism data. BMC Bioinform. 2013;14(13):1–10.
32. Batnyam N, Gantulga A, Oh S. An efficient classification for single nucleotide polymorphism (snp) dataset. In: Computer and Information Science; 2013. pp. 171–185.
33. Boutorh A, Guessoum A. Classication of snps for breast cancer diagnosis using neural-network-based association rules. In: 2015 12th International Symposium on Programming and Systems (ISPS); 2015. pp. 1–9.
34. Beshnova D, Ye J, Onabolu O, Moon B, Zheng W, Fu Y-X, Brugarolas J, Lea J, Li B. De novo prediction of cancer-associated t cell receptors for noninvasive cancer detection. Sci Transl Med. 2020;12(557):3738.
35. Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, Bethune MT, Fischer S, Yang X, Gomez-Eerland R, Bingham DB, Sibener LV, Fernandes RA, Velasco A, Baltimore D, Schumacher TN, Khatri P, Quake SR, Davis MM, Garcia KC. Antigen identification for orphan t cell receptors expressed on tumor-infiltrating lymphocytes. Cell. 2018;172(3):549–56316. https://doi.org/10.1016/j.cell.2017.11.043.
36. Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, Bernatchez C, Heymach JV, Gibbons DL, Wang J, Xu L, Reuben A, Wang T. Deep learning-based prediction of the t cell receptor-antigen binding specificity. Nat Mach Intell. 2021;3(10):864–75. https://doi.org/10.1038/s42256-021-00383-2.
37. Wang X, Yan Y, Tang P, Bai X, Liu W. Revisiting multiple instance neural networks. Pattern Recogn. 2018;74:15–24.
38. Xu Y, Qian X, Zhang X, Lai X, Liu Y, Wang J. Deeplion: deep multi-instance learning improves the prediction of cancer-associated t cell receptors for accurate cancer detection. Front Genet. 2022. https://doi.org/10.3389/fgene.2022.860510.
39. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, Holzleitner M, Brandstetter J, Sandve GK, Greiff V, Hochreiter S, et al. Modern hopfield networks and attention for immune repertoire classification. Adv Neural Inf Process Syst. 2020;33:18832–45.
40. Tourniaire P, Ilie M, Hofman P, Ayache N, DelingetteH. Attention-based multiple instance learning with mixed supervision on the camelyon16 dataset. In: MICCAI Workshop on Computational Pathology; 2021. pp. 216–226.
41. Rymarczyk D, Borowa A, Tabor J, Zielinski B. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. pp. 1721–1730.
42. Lu M, Pan Y, Nie D, Liu F, Shi F, Xia Y, Shen D. Smile: sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In: MICCAI Workshop on Computational Pathology; 2021. pp. 159–169.
43. Martins A, Astudillo R. From softmax to sparsemax: a sparse model of attention and multi-label classification. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning. vol. 48; 2016. pp. 1614–1623.
44. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 770–778.
45. Zhang Z, Xiong D, Wang X, Liu H, Wang T. Mapping the functional landscape of t cell receptor repertoires by single-t cell transcriptomics. Nat Methods. 2021;18(1):92–9.
46. Lee PP, Yee C, Savage PA, Fong L, Brockstedt D, Weber JS, Johnson D, Swetter S, Thompson J, Greenberg PD, et al. Characterization of circulating t cells specific for tumor-associated antigens in melanoma patients. Nat Med. 1999;5(6):677–85.
47. Lewis JD, Reilly BD, Bright RK. Tumor-associated antigens: from discovery to immunity. Int Rev Immunol. 2003;22(2):81–112.
48. Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. J Clin Investig. 2015;125(9):3413–21. https://doi.org/10.1172/jci80008.
49. Stevanović S, Pasetto A, Helman SR, Gartner JJ, Prickett TD, Howie B, Robins HS, Robbins PF, Klebanoff CA, Rosenberg SA, Hinrichs CS. Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer. Science. 2017;356(6334):200–5. https://doi.org/10.1126/science.aak9510.
50. Lu T, Wang S, Xu L, Zhou Q, Singla N, Gao J, Manna S, Pop L, Xie Z, Chen M, Luke JJ, Brugarolas J, Hannan R, Wang T. Tumor neoantigenicity assessment with csin score incorporates clonality and immunogenicity to predict immunotherapy outcomes. Sci Immunol. 2020;5(44):3199.
51. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci. 2005;102(18):6395–400.
52. Hu L, Yang S, Luo X, Yuan H, Sedraoui K, Zhou M. A distributed framework for large-scale protein-protein interaction data analysis and prediction using mapreduce. IEEE/CAA J Autom Sin. 2022;9(1):160–72. https://doi.org/10.1109/JAS.2021.1004198.
53. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: ICML; 2010. pp. 807–814 .
54. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(56):1929–58.
55. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning; 2015. pp. 448–456 .
56. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8026–37.
57. Garland M, Le Grand S, Nickolls J, Anderson J, Hardwick J, Morton S, Phillips E, Zhang Y, Volkov V. Parallel computing experiences with cuda. IEEE Micro. 2008;28(4):13–27.
58. Lin W-J, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform. 2012;14(1):13–26.
59. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform. 2019;90: 103089.

60. Zhou Z-H, Zhang M-L. Solving multi-instance problems with classifier ensemble based on constructive clustering. Knowl Inf Syst. 2007;11(2):155–70.
61. Wang J, Zucker J-D. Solving multiple-instance problem: A lazy learning approach; 2000.
62. Zhang J, Marszałek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vision. 2007;73(2):213–38.
63. Zhang Q, Goldman SA. Em-dd: an improved multiple-instance learning technique. In: Advances in Neural Information Processing Systems; 2002. pp. 1073–1080.
64. Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems; 2003. pp. 577–584.
65. Zhou Z-H, Sun Y-Y, Li Y-F. Multi-instance learning by treating instances as non-iid samples. In: Proceedings of the 26th Annual International Conference on Machine Learning; 2009. pp. 1249–1256.
66. Babenko B, Dollár P, Tu Z, Belongie S. Simultaneous learning and alignment: multi-instance and multi-pose learning. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition; 2008.
67. Chen Y, Bi J, Wang JZ. Miles: multiple-instance learning via embedded instance selection. IEEE Trans Pattern Anal Mach Intell. 2006;28(12):1931–47.
68. Cheplygina V, Tax DM, Loog M. Multiple instance learning with bag dissimilarities. Pattern Recogn. 2015;48(1):264–75.
69. Wang X, Yan Y, Tang P, Bai X, Liu W. Revisiting multiple instance neural networks. Pattern Recogn. 2018;74:15–24.
70. Gärtner T, Flach PA, Kowalczyk A, Smola AJ. Multi-instance kernels. ICML. 2002;2:7.
71. Ray S, Craven M. Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of the 22nd International Conference on Machine Learning; 2005. pp. 697–704.
72. Wang T, Lu R, Kapur P, Jaiswal BS, Hannan R, Zhang Z, Pedrosa I, Luke JJ, Zhang H, Goldstein LD, Yousuf Q, Gu Y-F, McKenzie T, Joyce A, Kim MS, Wang X, Luo D, Onabolu O, Stevens C, Xie Z, Chen M, Filatenkov A, Torrealba J, Luo X, Guo W, He J, Stawiski E, Modrusan Z, Durinck S, Seshagiri S, Brugarolas J. An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. Cancer Discov. 2018;8(9):1142–55.
73. Huang C, Li Y, Loy CC, Tang X. Learning deep representation for imbalanced classification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. pp. 5375–5384.
74. Sundin I, Schulam P, Siivola E, Vehtari A, Saria S, Kaski S. Active learning for decision-making from imbalanced observational data. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97. Proceedings of Machine Learning Research; 2019). pp. 6046–6055.
75. Yang Y, Xu Z. Rethinking the value of labels for improving class-imbalanced learning. In: Conference on Neural Information Processing Systems (NeurIPS); 2020.
76. Akram A, Inman RD. Immunodominance: a pivotal principle in host response to viral infections. Clin Immunol. 2012;143(2):99–115. https://doi.org/10.1016/j.clim.2012.01.015.
77. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class i-restricted t lymphocyte responses. Annu Rev Immunol. 1999;17(1):51–88.
78. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. pp. 558–567.
79. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Comput. 1996;8(7):1341–90.
80. Yamakawa H, Maruhashi K, Nakao Y. Predicting types of protein-protein interactions using a multiple-instance learning model. In: Washio T, Satoh K, Takeda H, Inokuchi A, editors. New frontiers in artificial intelligence. Berlin, Heidelberg: Springer; 2007. p. 42–53.
81. Zhang Y-P, Zha Y, Li X, Zhao S, Du X. Using the multi-instance learning method to predict protein-protein interactions with domain information. In: Miao D, Pedrycz W, Ślzak D, Peters G, Hu Q, Wang R, editors. Rough sets and knowledge technology. Cham: Springer; 2014. p. 249–59.
82. Wang X, Yang W, Yang Y, He Y, Zhang J, Wang L, Hu L. Ppisb: a novel network-based algorithm of predicting protein-protein interactions with mixed membership stochastic blockmodel. IEEE/ACM Trans Comput Biol Bioinform. 2022. https://doi.org/10.1109/TCBB.2022.3196336.
83. Hu P, Huang Y-A, Chan KCC, You Z-H. Learning multimodal networks from heterogeneous data for prediction of lncrna-mirna interactions. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(5):1516–24. https://doi.org/10.1109/TCBB.2019.2957094.
84. Hu P, Huang Y-A, Mei J, Leung H, Chen Z-H, Kuang Z-M, You Z-H, Hu L. Learning from low-rank multimodal representations for predicting disease-drug associations. BMC Med Inform Decis Mak. 2021;21(1):308. https://doi.org/10.1186/s12911-021-01648-x.
85. Zhao B-W, Hu L, You Z-H, Wang L, Su X-R. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab515.
86. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer; 2009.
87. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B Methodol. 1996;58(1):267–88.
88. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.