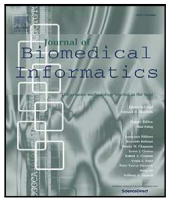
Journal of Biomedical Informatics 147 (2023) 104512

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Integration of incomplete multi-omics data using Knowledge Distillation and Supervised Variational Autoencoders for disease progression prediction

Sima Ranjbari, Suzan Arslanturk *

Department of Computer Science, Wayne State University, Detroit, 48202, MI, USA

ARTICLE INFO

Dataset link: http://cancergenome.nih.gov, https://github.com/Sima-Ranjbari

Keywords:
Survival prediction
Data integration
Knowledge distillation
Multi-modal data

ABSTRACT

Objective: The rapid advancement of high-throughput technologies in the biomedical field has resulted in the accumulation of diverse omics data types, such as mRNA expression, DNA methylation, and microRNA expression, for studying various diseases. Integrating these multi-omics datasets enables a comprehensive understanding of the molecular basis of cancer and facilitates accurate prediction of disease progression.
Methods: However, conventional approaches face challenges due to the dimensionality curse problem. This paper introduces a novel framework called Knowledge Distillation and Supervised Variational AutoEncoders utilizing View Correlation Discovery Network (KD-SVAE-VCDN) to address the integration of high-dimensional multi-omics data with limited common samples. Through our experimental evaluation, we demonstrate that the proposed KD-SVAE-VCDN architecture accurately predicts the progression of breast and kidney carcinoma by effectively classifying patients as long- or short-term survivors. Furthermore, our approach outperforms other state-of-the-art multi-omics integration models.
Results: Our findings highlight the efficacy of the KD-SVAE-VCDN architecture in predicting the disease progression of breast and kidney carcinoma. By enabling the classification of patients based on survival outcomes, our model contributes to personalized and targeted treatments. The favorable performance of our approach in comparison to several existing models suggests its potential to contribute to the advancement of cancer understanding and management.
Conclusion: The development of a robust predictive model capable of accurately forecasting disease progression at the time of diagnosis holds immense promise for advancing personalized medicine. By leveraging multi-omics data integration, our proposed KD-SVAE-VCDN framework offers an effective solution to this challenge, paving the way for more precise and tailored treatment strategies for patients with different types of cancer.

1. Introduction

The quick advancements of high-throughput technologies in biomedical domain led to the collection of a wide variety of "omics" data with unprecedented level of details. This provided the opportunity to use different genome-wide data with a variety of molecular functions, including mRNA expression, DNA methylation, and microRNA (miRNA) expression for diverse disease studies. Taken individually, each of these datasets offers solutions to important domain- and source-specific challenges. Collectively, they represent complementary views of related data entities with an aggregate information value often well exceeding the sum of its parts. Integrative analysis of multi-omics data has been proposed in many studies for a better understanding of cancers' molecular basis and accurate prediction of disease progression [1–4]. Some have employed statistical analysis and machine learning methods for multi-omics data integration for cancer survival prediction [5–12]. However, due to the dimensionality curse problem (limited number of samples with high dimensional feature space), conventional approaches typically faced limited ability to integrate multi-omics data effectively. Although different feature selection and dimensionality reduction methods have been proposed to tackle the dimensionality curse problem, this has led inevitably to a loss of valuable predictive information. More recently, deep learning approaches has achieved considerable success in multi-omics data integration on a variety of tasks, including cancer subtype prediction [13–16], disease progression prediction [17–19], pathway analysis and clustering [15, 20], and biomarker identification [21–24]. However, existing deep learning-based data integration approaches able to fuse data from different modalities still suffer from challenges as (i) inherent associations among multiple data, (ii) high dimensionality of the feature space, (iii) linearity assumptions and (iv) small number of samples

* Corresponding author.
E-mail addresses: sima.ranjbari@wayne.edu (S. Ranjbari), suzan.arslanturk@wayne.edu (S. Arslanturk).

https://doi.org/10.1016/j.jbi.2023.104512
Received 10 June 2023; Received in revised form 31 August 2023; Accepted 3 October 2023
Available online 7 October 2023

common among multiple modalities [25]. As a result, there is a critical need for a novel integrative analysis technique leveraging the complementary information available in multi-omics data with limited sample sizes overcoming the aforementioned challenges to better understand the biology of disease. In this paper, we have focused on the integration of multiple modalities with limited common samples by designing a novel framework referred to as Knowledge Distillation and Supervised Variational AutoEncoders utilizing View Correlation Discovery Network (KD-SVAE-VCDN). Here, we integrated multi-omics data using variational autoencoders, and used a knowledge distillation pipeline to unlock the full information potential among multiple modalities for a more accurate and robust understanding of disease progression.

It has long been understood that identifying patients' disease progression (short-term survival vs. long-term survival) at the time of diagnosis will lead to a more personalized and targeted treatment. While many attempts to achieve this based on integrating multiple types of high-throughput data have been undertaken, these efforts yielded only modest success so far due to the heterogeneity of cancer with multifactorial etiology. The technology proposed here can discover disease progression patterns at the time of diagnosis by integrating collective information available through multiple modalities with heterogeneous data types (mRNA, miRNA, DNA Methylation, etc.) and limited number of common samples. The results of this paper can help optimize treatment by separating the patients with aggressive disease from those with less aggressive disease, as well as to increase the success of clinical trials by separating the respondents vs. non-respondents to treatments.

2. Related work

Several studies have focused on omics-data (without integration of multiple modalities) using statistical analysis to discover associations among clinical and biological features [21,22,26]. However, the carcinogenesis and progression of disease may be a result of complex mechanisms and changes at different levels, such as genome, proteome, and transcriptome. Therefore, integration of omics data provides better opportunities to understand the biology of cancer [27]. Generally, data integration can fall into three different categories [28]: (i) late or output integration: each data is modeled separately and the final outputs are combined subsequently (ii) intermediate or partial integration: this refers to a joint model that learns from multi modalities simultaneously and (iii) early or complete integration: this integration method focuses on combining data before the learning process, either by simple concatenation or by learning a joint latent representation. Though successful, integration of multiple modalities suffers from challenges including the curse of dimensionality, data heterogeneity, inconsistent data distributions, scaling, small number of common samples among distinct modalities.

There has been several machine learning and deep learning algorithms able to integrate multi-modal data by overcoming several aforementioned challenges including the high dimensionality [13,14,23,25,29–32]. Some of these proposed models are better suited than others for integration of various kinds of data, such as autoencoders (AE). Simidjievski, et al. [28] have proposed a network based on different variational autoencoders (VAEs) for data integration to classify patients into breast cancer subtypes. They have designed and tested different network architectures and reported that the performance of X-shaped VAEs (that learns to reconstruct the input data from a single shared homogenous latent representation that is built from different heterogenous data sources) and Hierarchical VAEs (that learns a high-level representation with the input of low-level representations that are built on each single data, separately) outperformed all other network architectures. Arslanturk et al. have proposed a data integration methodology to identify subtypes of cancer using multiple data types (mRNA, methylation, microRNA and somatic variants) and different data scales that come from different platforms (microarray, sequencing, etc.) [33]. Their proposed data integration and disease subtyping approach accurately identifies novel subgroups of patients with significantly different survival profiles. Ma et al. [34] proposed an autoencoder based architecture for cancer progression and survival prediction integrating multi omics data. They built hidden latent representations of the data separately and then utilized them to calculate the patients' similarity matrix to feed to a neural network classifier. In another study, Mitchel et al. [35] pre-selected the important features based on mutual information gain and then applied principal component analysis to the selected features and fed them into a neural network for subsequent breast cancer survival prediction. They employed gene expression, DNA methylation, miRNA expression, and copy number variations as the input. Wang et al. [22] have proposed a model based on graph convolutional networks (GCNs) to integrate omics data for cancer detection. After training GCNs on each individual data separately, they performed the classification task and fused the probabilities in a cross-omics discovery network to feed into the final classifier. This way, they were able to measure the correlation among modalities besides feature extraction to boost the classification performances.

Multi-modal learning generally aims to integrate information from distinct modalities with heterogeneous features that describe the same set of subjects. Utilizing the information available from different modalities will lead to an enhanced performance as compared to learning with the information available from only one modality. Though successful, a common drawback of multi-modal learning is to only utilize the shared information of multiple modalities. Different modalities may have distinct sample sizes with only a limited number of samples in common among modalities. For instance, the Kidney Renal Clear Cell Carcinoma (KIRC) dataset available at TCGA has around 530 patients with gene expression, copy number variation, and DNA Methylation data but only 339 patients with single nucleotide variant (SNV) data. This may result in (i) either excluding the SNV from the analysis and developing models using the remaining datatypes or (ii) using only the 339 patients that are in common across all data types within the analysis. Both solutions would lead to a significant information loss and hence can affect the model's ability to demonstrate optimal performance. Several studies have discarded the samples with missing modalities and only focused on samples common across all modalities [5,9,13,22,36]. However, discarding such valuable information is a major concern especially with small samples as the amount of information from the onset is already limited making each observation crucial to preserve. One solution to this problem is to use missing value imputation techniques to replace the missing entries with an estimated value based on other available information, however this may introduce bias that can affect the subsequent prediction tasks [27,28,37,38]. Zhou et al. proposed a method that incorporates the missing modalities into the training network [25]. They initially used only complete samples of all modalities to get the shared latent representation to capture the intra-correlation among data and then used the incomplete data separately to learn each data's features as accurately as possible. Finally, they combined all representations to map into the label space. They used neuroimaging and genetic data (single nucleotide polymorphism) for diagnosing Alzheimer's disease. Wang et al. [39] proposed a framework that has inspired and given us the basis on which we have built our KD-SVAE-VCDN architecture. In their proposed framework, they have used a knowledge distillation-based model that is able to utilize the supplementary information of all modalities, and hence preventing large amounts of data to be wasted. Knowledge distillation allows the transfer of knowledge encoded in the pseudolikelihoods assigned to the output of a large model (i.e., a teacher model) to a smaller model (i.e., a student model). Learning the distribution of likelihoods among classes for a sample during training through the large model, and then distilling such knowledge into the smaller model results in a better ability to learn concise knowledge representations. In their proposed approach, they initially train models on each modality separately using

all the available data. Then the trained models are used as teachers by transferring the concise knowledge representations to the student model, which is trained with only those samples having complete modalities. In this paper, we proposed a novel approach referred to as KD-SVAE-VCDN for cancer progression prediction. Our contributions are summarized as follows:

- We have defined an end-to-end pipeline able to integrate data from multiple modalities for subsequent disease progression prediction tasks. Compared to various information fusion strategies, our model is able to enrich the study population through a knowledge distillation-based model that is able to utilize the supplementary information of all modalities, and hence preventing large amounts of data to be wasted.
- In addition, through utilizing cross-omics correlation tensors and VCDN, we obtained the intra-correlation among multi-modalities in the latent representation space to be included in the classification task.
- Our proposed network outperformed other state of the art multi-omics integration models.

3. Methodology

3.1. Supervised variational AutoEncoders

Recently, many machine learning algorithms have been employed to enhance treatment and to better understand disease progression for patients with cancer. Some of these models are better suited than others for integration of various kinds of data [28]. In our study, we use variational autoencoders (VAEs) due to the fact that they are generative, non-linear, and capable of learning meaningful information as well as integrating different types of high dimensional data modalities.

In general, an autoencoder encompasses two networks, an encoder and a decoder, that perform (i) encoding, *i.e.*, transforming input data with high dimensions into a latent representation with lower dimensions and (ii) decoding, *i.e.*, reconstructing the input data from the embedding output of the encoder with minimal loss [28]. The model includes an encoder function $e(.)$ and a decoder function $d(.)$ parameterized by ϕ and θ, respectively. The lower dimensional representation learned from an input x is referred to as $e(x)$ and the reconstructed input is $x' = d(e_\phi(x))$.

The key problem in designing an autoencoder is that it is highly affected by its input data. A VAE, one of the recent variants of autoencoders is capable of addressing the aforementioned problem. The VAE uses variational inference to estimate the underlying probability distribution of the data, in the form of latent variables z. In a probabilistic framework, a VAE draws the high-dimensional data x from a random variable with distribution p_{data} (x). The hidden representation space (also referred to as a 'bottleneck') is stochastic with a gaussian probability density. Let us denote the encoder output as q_θ (x | z) so the VAE tries to estimate the true posterior p_θ (x | z) with true parameters ϕ by adopting a recognition model with trainable parameters θ of a fully connected neural network.

Generally, the VAE model assumes that the latent representation follows a centered isotropic multivariate gaussian distribution denoted as $p_\phi(z) = N(z; 0, I)$ and it will be necessary for the variational approximate posterior to have a multivariate gaussian structure as formulated by the following equation:

$$q_\phi(z|x^{(i)}) = N(z; \mu^{(i)}, \sigma^{(i)} I) \tag{1}$$

where $\mu^{(i)}$, and $\sigma^{(i)}$ represent the mean and variance vectors. The difference between p_θ (z) and q_ϕ $(z|x^{(i)})$ can be easily computed and discriminated as both are normally distributed. Therefore, the loss function for a datapoint $x^{(i)}$ can be written as:

$$l_i(\theta, \phi) = -E_{q_\theta(z|x^{(i)})}[log p_\theta(x|z)] + KL(q_\theta(z|x^{(i)} \parallel p_\theta(z))] \tag{2}$$

Here, the first term is the reconstruction loss which refers to the decoder's output to regenerate the datapoint with minimal information loss. The expected negative log-likelihood is taken into consideration with regard to the distribution of the encoder over the representations. The second term, Kullback–Leibler divergence (KL-divergence), encourages the difference between the true prior p(z) and posterior distributions q(z | x) to be minimized. The latent variables z generated by the encoder can be fed to a classifier. Therefore, the output layer of the encoder is connected to a neural network classifier. The classification performance can be heavily affected by the quality of the generated features by the encoder. As a result, we specify the total loss for the SVAE as the following:

$$l_i(\theta, \phi) = -E_{q_\theta(z|x^{(i)})}[log p_\theta(x|z)] + KL(q_\theta(z|x^{(i)} \parallel p_\theta(z))) - [y_i.log(p(y_i)) + (1 - y_i).log(1 - p(y_i))] \tag{3}$$

Here, the third term represents the binary cross entropy loss (BCE), so the latent representations generated by the encoder influence the classification criteria to make a better inference.

3.2. Knowledge distillation

Transferring knowledge from a teacher to a student is considered as knowledge distillation. At first, the teacher model is trained using a single modality on a given dataset $D = \{\{X_1, y_1\}, \{X_2, y_2\}\{X_N, y_N\}\}$, where X_i representing the data features for the ith sample with one-hot actual labels y_i and learning parameters ϕ, denoted as Te ϕ. The prediction model will then generate logits z_i for each sample i. Then, the student model tries to replicate the teacher's output. Assuming there are C classes, (2 in our case study), the generated labels are given by:

$$z_i = Te(X_i, \phi) \tag{4}$$

Afterwards, the student model is trained with both one-hot actual labels $\{y_1, y_2, \ldots, y_N\}$ and the logits $\{z_1, z_2, \ldots, \}$ which are softened using temperature scaling, denoted by the following:

$$\sigma_j^{\sim t}(x) = \frac{e^{x_j/t}}{\Sigma_{k=1}^{C} e^{\frac{x_k}{t}}} for j = 1, 2, \ldots, C, and t > 1 \tag{5}$$

Here, $\sigma_j^{\sim t}(x)$ represents the softened class probability distribution produced by the model Te (ϕ). The main idea behind using soft labels in knowledge distillation is that it is more informative about a data sample than the peaky probability distributions. For example, if there are multiple classes and the predicted probabilities for all classes are high, it means that the sample of interest might lie on the decision boundary. Therefore, forcing a student to imitate these probabilities should encourage the network to absorb some of the teacher's knowledge in addition to what is contained in the true labels alone. Assume the student model is trained by parameter ϕ, denoted as St(ϕ) which takes the input X_i. In the student network, the loss function for training phase is defined as follows:

$$\min_{\varphi} l = \Sigma_i^n l_c(X_i, y_i; \varphi) + l_d(X_i, z_i; \varphi) \tag{6}$$

where l_c is related to a classification loss with the genuine label, which its formula has been provided in Eq. (3). l_d denotes the distillation loss which can take the form of negative cross-entropy loss or KL-divergence. In this paper, the KL-divergence loss is used as the distillation loss. The formula is represented as:

$$l_d(X_i, z_i; \phi) = D_{KL}(\sigma^t(St(X_i, \phi); t), \sigma^t(Te(X_i, \phi); t)) \tag{7}$$

where $\sigma^t(Te(X_i, \phi); t)$ is the generated probabilities with temperature t rescaled by SoftMax. Also, the term σ_{Te}^t $(z_i; t)$ refers to soft labels of the teacher network for the same sample X_i. It is worth noting that the higher the temperature t, the more smoothed the output probability.

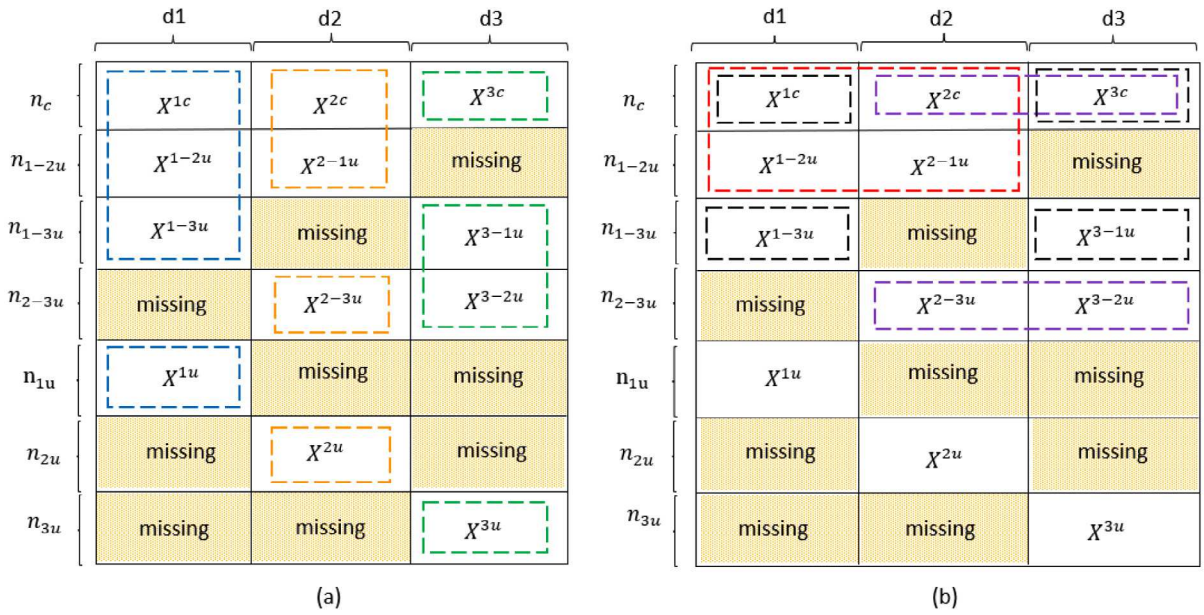


Fig. 1. (a) represents the structure of data used for first-level teachers. Samples in blue dashed-line box are utilized to train the first teacher model, namely Te1, the orange dashed-line box are for the second teacher model, namely Te_2, and samples included in green dashed-line box are for the third teacher model, namely Te_3. (b) shows the structure of the data used for second-level teachers. Samples in red dashed-line box are included in training of the model Te_{1-2}, samples in black dashed-line box are for Te_{1-3}, and samples in purple dashed-line box are used for training of Te_{2-3}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Multi-modal learning with missing modalities

It is rather typical for some samples in multimodal learning to lack some modalities. In our case study, we have three modalities with their data features and actual labels denoted as $\{X^1 \in R^{n_1 \times d_1}, X^2 \in R^{n_2 \times d_2}, X^3 \in R^{n_3 \times d_3}\}$ with n_i and d_i representing the sample size and feature dimensions of the ith modality, respectively. Data samples having all modalities present (or complete) are indicated as $X^{1c} \in R^{n_c \times d_1}, X^{2c} \in R^{n_c \times d_2}$, and $X^{3c} \in R^{n_c \times d_3}$ for the first, second, and third modalities, respectively. Samples having two complete modalities (which are the first and second modalities) are denoted as $X^{1-2u} \in R^{n_{1-2u} \times d_1}$ and $X^{2-1u} \in R^{n_{2-1u} \times d_2}$. To clarify, it is worth mentioning that $n_{1-2u} = n_{2-1u}$, since those notations each represent the same set of samples that are common among the first and second modalities. Similarly, the data with the first and third modalities present are denoted as $X^{1-3u} \in R^{n_{1-3u} \times d_1}$ and $X^{3-1u} \in R^{n_{3-1u} \times d_3}$ and the data with their second and third modalities present are indicated as $X^{2-3u} \in R^{n_{2-3u} \times d_2}$ and $X^{3-2u} \in R^{n_{3-2u} \times d_3}$. Moreover, samples having only one modality present are denoted as $X^1u \in R^{n_{1u} \times d_1}, X^{2u} \in R^{n_{2u} \times d_2}$, and $X^{3u} \in R^{n_{3u} \times d_3 2}$. Note that, $n_1 = n_c + n_{1u} + n_{1-2u}(or\ n_{2-1u}) + n_{1-3u}(or\ n_{3-1u}), n_2 = n_c + n_{2u} + n_{2-1u}(or\ n_{1-2u}) + n_{2-3u}(or\ n_{3-2u}), and n_3 = n_c + n_{3u} + n_{3-1u}(or\ n_{1-3u}) + n_{3-2u}(or\ n_{2-3u})$.

Fig. 1 shows the structure of the data used in our case study. There are two steps to train the teacher models, Fig. 1(a) shows the data used for first-level single modal models acting as teachers for the subsequent step. Therefore, in this step we are using all the available samples including the data with missing modalities. Here, we construct these teacher models as three SVAEs, namely $SVAE_1$ (ϕ_1), $SVAE_2$ (ϕ_2), and $SVAE_3$ (ϕ_3) with parameters ϕ_1, ϕ_2, and ϕ_3 using the data from $X^1 = [X^{1c}, X^{1-2u}, X^{1-3u}, X^{1u}], X^2 = [X^{2c}, X^{2-1u}, X^{2-3u}, X^{2u}]$, and $X^3 = [X^{3c}, X^{3-1u}, X^{3-2u}, X^{3u}]$, respectively. The teacher j is trained by minimizing the equations of losses for SVAEs by using the following equation.

$$Te_j(\phi_j) = min_{\phi_j} \underbrace{\Sigma_i^{n_j} BCE(d(e(X_i^j;\theta);\phi), X_i^j) + KL(q_\theta(z|X_i^j) \parallel p(z))}_{l_{VAE}} + \underbrace{\alpha * clf(SVAE_j(X_i^j;\phi_j)), y_i)}_{l_{clf}}, j \in \{1,2,3\} \quad (8)$$

The l_{VAE} part of Eq. (8) represents the VAE's loss and the l_{clf} part represents the classification loss with α defined as the coefficient. Next, we use these three first-level teacher models to label the next step's samples. The logits z_i^j, representing the labels defined by teacher j on the ith sample can be denoted as:

$$z_i^j = Te_j(X^{k-t};\phi_j)k, t, j \in \{1,2,3\}, k < t \quad (9)$$

Fig. 1(b) shows the data samples utilized for second-level teachers. In this step we are using all the data available in each pair of modalities. We construct the second-level teacher models as three SVAEs, namely, $SVAE_{1-2}$ (ϕ_{1-2}), $SVAE_{1-3}$ (ϕ_{1-3}), and $SVAE_{2-3}$ (ϕ_{2-3}) with parameters ϕ_{1-2}, ϕ_{1-3}, and ϕ_{2-3} using the data from $X^{1-2} = [X^{1c}, X^{2c}, X^{1-2u}, X^{2-1u}], X^{1-3} = [X^{2c}, X^{3c}, X^{1-3u}, X^{3-1u},]$, *and* $X^{2-3} = [X^{2c}, X^{3c}, X^{2-3u}, X^{3-2u}]$, respectively.

Note that, here we also use the aforementioned first-level teacher models' logit outputs to train the second-level teacher models so the loss for these three teachers can be minimized as defined in the following equation:

$$\begin{aligned} Te_{k-t}(\phi_{k-t}) = min_{\phi_{k-t}} \Sigma_i^{n_{k-t}} l_{VAE}^{k-t} + l_{clf}^{k-t} + \beta D_{KL}(\sigma^t(SVAE_{k-t} \\ \times (X^{k-t}, \phi_{k-t}); t), \sigma^t(Te_k(X^{k-t}, \phi_k); t)) + \\ \gamma D_{KL}(\sigma^t(SVAE_{k-t}(X^{k-t}, \phi_{k-t}); t), \\ \sigma^t(Te_t(X^{k-t}, \phi_t); t))k, t \in \{1,2,3\}, k < t \end{aligned} \quad (10)$$

In the above formula, D_{KL} denotes the distillation loss and β and γ are tunable parameters that can determine how much knowledge can be distilled from the previous teachers' network to the current teacher network. For instance, if k = 1 and t = 2, the knowledge of previous step teachers, namely Te_k (Te_1) and Te_t (Te_2), will be distilled to $Te_{k-t}(Te_{1-2})$ which is trained on the modalities indexed by k(1) and t(2) using the $SVAE_{k-t}(\phi_{k-t})(SVAE_{1-2}(\phi_{1-2}))$ models. We use these teachers to label the ith sample which is common among all three modalities X^{1c}, X^{2c}, and X^{3c}, so the logits produced by the teachers k-t are as the following:

$$z_i^{k-t} = Te_{k-t}([X^{kc}, X^{tc}];\phi_{k-t})k, t \in \{1,2,3\}, k < t \quad (11)$$

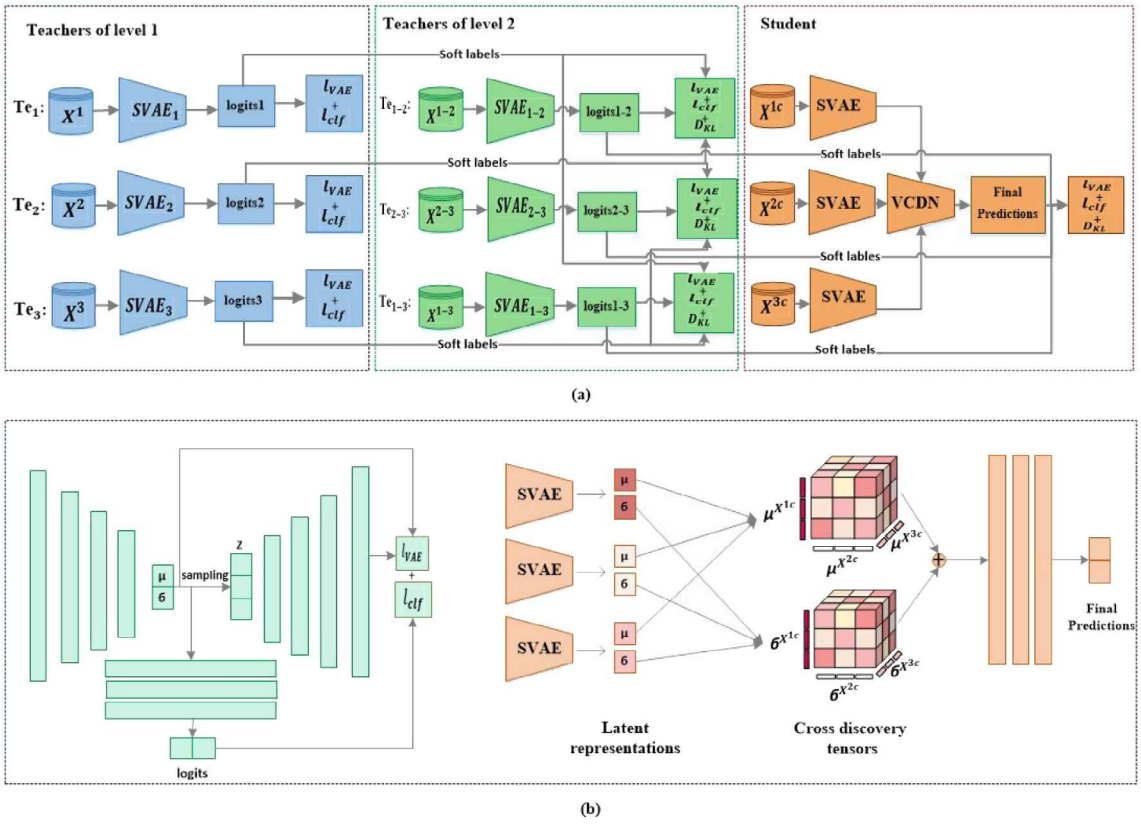


Fig. 2. Proposed KD-SVAE architecture. (a) shows the teacher models in each level along with the student model, (b) represents the architecture of the SVAE model (left) and the architecture of the student model (right).

Finally, we use the second-level teachers to train the student model St(φ) by minimizing the following loss function:

$$\begin{aligned} St(\varphi) = min_{\varphi} \Sigma_i^{n_c} l_{VAE} + l_{clf} + aD_{KL}(\sigma^t(SVAE([X^{1c}, X^{2c}, X^{3c}], \varphi); t), \\ \sigma^t(Te_{1-2}([X^{1c}, X^{2c}], \phi_{1-2}); t)) + \\ bD_{KL}(\sigma^t(SVAE([X^{1c}, X^{2c}, X^{3c}], \varphi); t), \\ \sigma^t(Te_{1-3}([X^{1c}, X^{3c}], \phi_{1-3}); t)) + \\ cD_{KL}(\sigma^t(SVAE([X^{1c}, X^{2c}, X^{3c}], \varphi); t), \\ \sigma^t(Te_{2-3}([X^{2c}, X^{3c}], \phi_{2-3}); t)) \end{aligned} \tag{12}$$

The student model is an SVAE with parameter φ, denoted as SVAE(φ), which uses data from the samples that have all three modalities present, $X^{1c}, X^{2c}, and X^{3c}$. The hyperparameters a,b, and c are the tunable in order to control the amount of knowledge to be distilled to the student model from the previous teachers.

3.4. The proposed KD-SVAE-VCDN architecture

When we have multiple modalities and desire to integrate them for a subsequent classification task, useful information could be discarded if we were to only use the common samples among modalities with limited sample sizes. Here, we propose a 3-fold KD-SVAE architecture, as shown in Fig. 2.

The first fold (referred to as Teachers of level 1) includes the three 1st-level teacher models, namely Te_1, Te_2, and Te_3, each taking a single modality as an input. Next, the mean and variance vectors identified through the SVAE models were used to generate the output logits denoted as z^1, z^2, and z^3. Note that, each teacher models' loss has been calculated and optimized differently to focus specifically on each individual modality using Eq. (8). The second fold (referred to as Teachers of level 2) of the architecture includes the 2nd-level teacher models, namely $Te_{1-2}, Te_{2-3}, and\ Te_{1-3}$. Here, we use each pair of modalities as an input in an effort to learn the features of the integrated data among multiple modalities to produce the logits $z^{1-2}, z^{2-3}, and\ z^{1-3}$. The models are optimized for each teacher model separately using Eq. (10). The softened labels generated in the previous step are then used as actual labels for defining the distillation losses. Finally, the common samples among all three modalities are integrated within the student model which consists of three VAEs. In order to build an effective multi-omics data integration framework, we obtained the cross-omics correlation. To specify, the mean and variance vectors generated through the VAEs for each modality were utilized differently through a View Correlation Discovery Network (VCDN) at the latent representation space to generate two discovery tensors (See Fig. 2b, right). Then the tensors were directly concatenated and fed to the neural network layers to get the final prediction labels. Note that, if the latent representations were used to generate only one cross discovery tensor, it would result in a larger computational complexity. Therefore, we generated two different tensors using the mean and variance vectors, separately. By using Eq. (12), the loss was calculated and each SVAE was optimized. Also, the previous teacher models' softened labels were utilized as actual labels in the calculation of distillation losses within the student model. The pseudocode of the proposed method is described in Algorithm 1.

The training has been performed on a computer equipped with Intel(R) Core(TM) i5-4300U CPU @ 1.90 GHz 2.50 GHz using 8.00 GB of RAM. The model complexities for KD_SVAE_VCDN and KD_SVAE_NN are listed in Table S1.

Algorithm 1. The proposed KD-SVAE-VCDN model for three modalities

Initialization

Inputs: X^j, y^j, X^{k-t}, and $y^{k-t}(j,k,t \in \{1,2,3\}, k<t)$, X^{1c}, X^{2c}, X^{3c}, y^c, α, β, Y, a, b, and c

Training teacher models in level 1:

1: For number of training iterations **do**

2: Train teachers Te_j with $\{[X^j, y^j]\}$ using Equation (8);

3: end **For**

4: Obtain soft labels for X^{k-t} using Equation (9);

Training teacher models in level 2:

5: For number of training iterations **do**

6: Train teachers Te_{k-t} with $\{[X^{k-t}, y^{k-t}]\}$ using Equation (10);

7: end **For**

8: Obtain soft labels for X^{1c}, X^{2c}, X^{3c} employing Equation (11);

Training student model:

9: For number of training iterations **do**

10: Train student St with $\{[X^{1c}, X^{2c}, X^{3c}, y^c]\}$ using Equation (12);

11: end **For**

3.5. Dataset

We conducted our study using multi omics data (including mRNA expression, miRNA, and DNA methylation) from breast carcinoma (BRCA) and pan-kidney cohort (KIPAN) samples available at The Cancer Genome Atlas (TCGA). For BRCA, using the survival days available within the clinical data, patients were stratified into two groups namely short- vs. long-term survivors (defined based on survival < 3 years vs. survival > 5 years). The five year cut-off was determined using an Expectation Maximization (EM) algorithm. Through EM, we were able to fit two Gaussian distributions as shown in Fig. 3 that were well separated using 1079 breast cancer samples. The two distributions were representing two separate clusters (short- vs. long-term survival) intersecting at approximately 1800 days (≈5 years). Therefore, we refer to patients with survival greater than five years as long-term survivors. Patients whose survival days were between three and five years were excluded from this study due to the uncertainty of their progression status. This exclusion prevents any potential bias affecting models ability to correctly predict patient progression. Similarly, patients with survival less than three years were referred to as short-term survivors. Hence, the labeling of data was achieved through an unsupervised clustering approach. Note that, the already challenging circumstance of low sample counts, particularly for long-term survivors prevented us from introducing larger number of clusters. After excluding several patients with survival between three to five years, the number of samples with mRNA, miRNA, and DNA methylation modalities were reported to be 907, 633, and 656, along with feature dimensions of 18 276, 638, and 17 037, respectively. Since redundant features and noise may impact the classification performance, preprocessing and preselection of features were carried out on each omics data, separately. After min–max scaling, the features of mRNA and DNA methylation data with low variance (with threshold 0.02) were eliminated. Additionally, using ANOVA test on all modalities, with features having p-values > 0.05 were also excluded. This resulted in 2723, 582, and 129 features for mRNA, DNA methylation, and miRNA modalities, respectively. In order to train our proposed architecture, we randomly selected 80% of the data as training and 20% of the data as the hold-out (test set). The class distribution in the constructed test set was preserved as the original dataset. Table 1 shows the total number of samples for each class (long- vs. short-term survival) and each combination of modalities before the train-test split.

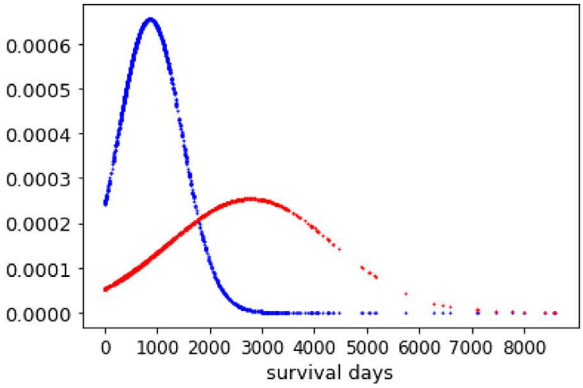


Fig. 3. Distribution of two classes (short- and long-term survivors) achieved through Expectation Maximization algorithm. Here the intersecting point between the two distributions (≈1800 days) is used as the cutoff for short- vs. long-term survivors.

3.6. Generalized method for multiple modalities

Considering having m distinct modalities $X^1 \in R^{n_1 \times d_1}, X^2 \in R^{n_2 \times d_2}$, $X^m \in R^{n_m \times d_m}$, the data can be partitioned into n different subsets: (i) samples having all modalities present with $X^{ic} \in R^{n_{ic} \times d_i}$ and $i \in \{1,2,3,..m\}$. (ii) samples that have only one modality present with $X^{iu} \in R^{n_{iu} \times d_i}$ and $i \in 1,2,3,\ldots,m$ (iii) samples having two modalities present $X^{A \setminus 2u} \in R^{n_{\{A \setminus 2u\}} \times d_f}$, where $A \setminus 2$ denotes two-level combination of subsets written as i-j with $i,j \in \{1,2,3,\ldots,m\}$ and $i \neq j$. Here, the subset i-j differs from the subset j-i and f denotes the first index of the subset i-j, …(n) and (iv) samples having m-1 modalities present with $X^{\{A \setminus m-1\}u} \in R^{n_{\{A \setminus m-1\}u} \times d_f}$ in which $A \setminus m-1$ representing the (m-1)-member subsets written as $i-j \ldots -k$ where $i \neq j \neq \ldots \neq k \in \{1,2,3,\ldots,m\}$ and f is the first index in the subset where the order of indexes matter. Hence, $X^{1-2-3u} \in R^{n_{1-2-3u} \times d_1}$ and $X^{2-1-3u} \in R^{n_{1-2-3u} \times d_2}$ are two separate representations.

For training the teachers in a hierarchical manner, we first train the models on each modality separately and get Te_i with $i \in \{1,2,3,\ldots,m\}$. Afterwards, we use these models to train the subsequent teachers with two modalities present and obtain Te_{i-j} with $i,j \in \{1,2,3,\ldots,m\}$. Next, we use the teachers Te_{i-j}, to obtain teachers trained on three common modalities and so on. As a result, we get the teacher models hierarchically. Note that the teachers trained with h modalities (i.e., h-level teachers) denoted by M_h which has the size of $(m..h)$. For instance, if $m = 1,2,3,4$ then $M_2 = 1-2, 1-3, 1-4, 2-3, 2-4, 3-4$, and $M_3 = 1-2-3, 1-2-4, 1-3-4, 2-3-4$. Note that, the h-level teachers use the data from all possible subset of indexes in M_h. For instance, the teacher model with index 1-3-4 is trained with each combination of modalities consisting of 1,3, and 4, i.e., $X^{(1-3-4u)}, X^{(3-1-4u)}$, and $X^{(4-1-3u)}$. Here, we assume $Te_{M_{hk}}$, with network parameters ϕ_{hk}, representing the model of the kth teacher from the h-level teachers. Note that M_{hk} refers to the kth element in set h. Hence, $M_{23} = 1-4$. The loss function to be minimized for the teacher $Te_{M_h k}$ is defined as:

$$\begin{aligned} Te_{M_{hk}}(\emptyset_{hk}) = min_{\phi_{hk}} \Sigma_i^{n_{M_{hk}}} l_{VAE}^{M_{hk}} + l_{clf}^{M_{hk}} + \Sigma_i^{n_{M_{hk}}} \Sigma_j^{|M_{h-1}|} a_j D_{KL} \\ \times [\sigma^t(SVAE_{M_{hk}}(X^{uM_{hk}} \\ , \phi_{hk}); t), \sigma^t(Te_{M_{(h-1)}}) \\ \times (X^{M_{(h-1)j}}), \phi_{(h-1)j}; t] \end{aligned} \quad (13)$$

Note that the above equation is the general form of Eq. (10). M_{h-1} represents the cardinality of set M_{h-1} and $n_{M_{hk}}$ is the sample size whose modalities are indexed by M_{hk}. Finally, we train the student model using all the obtained teachers. The number of teachers models to be trained with m modalities is 2^{m-1}. Hence, large number of modalities will result in high computational costs. A simple solution for reducing the computational complexity would be to prune the subset of teachers with suboptimal performances.

Table 1
Number of samples for each class and each combination of modalities.

	Modalities	#of patients with long-term survival	#of patients with short-term survival
Level 1 combinations	mRNA	252	655
	methylation	182	474
	miRNA	174	459
Level 2 combinations	mRNA-methylation	176	471
	methylation-miRNA	126	395
	mRNA-miRNA	168	457
Level 3 combinations	Complete data	130	393

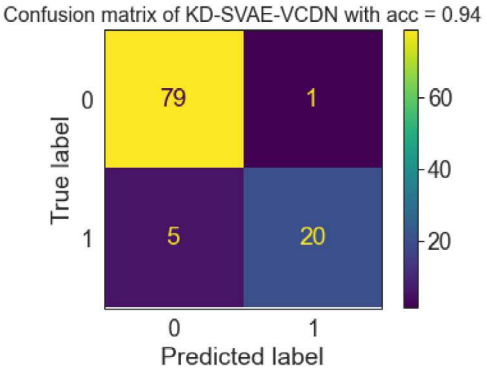


Fig. 4. Confusion Matrix of the KD_SVAE_VCDN model on test data for breast cancer progression prediction.

4. Results

4.1. Comparison of KD-SVAE-VCDN and its variations with existing multi-omics integration methods

For evaluation of the proposed KD-SVAE-VCDN, we compared the classification performance with some state-of-the-art multi-omics integration methods, including MOGONET [22] and DeepMO [13]. We also compared the results with some traditional machine learning approaches, including Xgboost and SVM using PCA analysis for feature dimension reduction (referred to as pca_Xgboost, and pca_svm, hereafter). In order to evaluate and compare the performances of each model, we employed the accuracy (ACC), balanced accuracy, F1 score (F1), area under the ROC curve (AUC), precision, and recall. We evaluated the performance of all models across 30 runs, with the mean and standard deviations of all performance measures reported in Table 2. Note that, the classification results for pca_Xgboost and pca_svm models are obtained by training on the directly concatenated preprocessed multi omics data as input.

Moreover, to test the effectiveness of SVAE and VCDN models separately, we compared the proposed method to its three different variations. (1) KD_AE_VCDN: Here, simple autoencoders (AE) were replaced by VAEs with the same number of layers and the same number of hidden layers. (2) KD_SVAE_NN: Here, a fully connected NN with the same number of layers as VCDN was used for integration. Instead of constructing a cross-discovery tensor, the latent representations of VAES for multi-omics data were concatenated and fed as input to the final NN. (3) SVAE_VCDN: Here, the common samples among multiple modalities were used as input to the student model, i.e., the knowledge distillation process and knowledge transfer via teacher models were not utilized. The KD_SVAE_VCDN outperformed its variations and alternative methods as shown in Table 2. The results have shown that the knowledge distillation has a significant contribution to the classification performance. Also, using a cross view tensor boosts the classification results slightly compared to KD_SVAE_NN due to the fact that it considers the intra-correlation among modalities during integration. Given that VCDN multiplies the hidden representations to build the cross-omics discovery tensor, accurate representations of the hidden layers will lead to exploiting the full potential of VCDNs. In fact, the noisy input can increase the prediction error; therefore, SVAEs contributed better as compared to traditional AEs. Note that, the obtained results in Table 2 indicate a statistically significant p-value (< 0.001).

Fig. 4 represents the confusion matrix on the test data with an accuracy of 94%. Moreover, for model performance comparison of KD_SVAE_VCDN with its variations, area under the ROC curves are depicted in Fig. 5a.

4.2. Performance of KD_SVAE_VCDN using different combinations of multi-omics data

Although we used three distinct modalities for the classification purpose, we evaluated the effectiveness of our proposed KD_SVAE_VCDN model under different combination of modalities to assess the necessity of integration of multi-omics data. The classification performance results of KD_SVAE_VCDN using three different omics data (combining mRNA expression, DNA methylation, and miRNA expression data denoted as mRNA + meth + miRNA), and using two different combination of modalities (combination of mRNA and miRNA expression data denoted as mRNA + miRNA, combination of DNA methylation and miRNA expression data denoted as meth + miRNA, combination of DNA methylation and mRNA expression data denoted as meth + mRNA), and using only single modalities are depicted in Fig. 5b. Also note that, during training using single modality, there was only one teacher and one student model involved.

4.3. Learning curve of distillation loss (D_{KL})

As an additional evaluation, the learning curve with different values of weights (a, b, c) for D_{KL} using the final student model is calculated on the test data which is presented in 5c. The parameters a, b, and c represent the weights of the distilled knowledge using the combination of modalities mRNA-methylation, mRNA-miRNA, and miRNA-methylation, respectively. As shown in Fig. 5c when there is no distilled knowledge, the loss value is not stable. However, when the D_{KL} is activated, the model is stabilized and the probability distributions along the class labels are well aligned.

4.4. Comparison of results using different temperature values

In Fig. 5d, we have shown different temperature hyperparameters tuned in the student model's loss. Results show that when the temperature is 1.5, the performances of the KD_SVAE_VCDN in general outperforms all other temperature value performances.

Table 2
Comparison of classification model performances on TCGA breast carcinoma (BRCA) and the pan-kidney cohort (KIPAN) data using different variations of the proposed model, alternative deep learning models and traditional machine learning models. The top three performances are highlighted in green, yellow, and blue, respectively.

Method	ACC	Balanced accuracy	F1-Score	Precision	Recall	AUC
Breast carcinoma						
KD_SVAE_VCDN[a]	0.92 ± 0.01	0.90 ± 0.02	0.85 ± 0.02	0.85 ± 0.04	0.86 ± 0.04	0.93 ± 0.02
KD_AE_VCDN[a]	0.86 ± 0.02	0.82 ± 0.02	0.80 ± 0.02	0.77 ± 0.04	0.82 ± 0.03	0.88 ± 0.03
KD_SVAE_NN[a]	0.90 ± 0.02	0.87 ± 0.02	0.82 ± 0.03	0.83 ± 0.04	0.80 ± 0.05	0.91 ± 0.02
SVAE_VCDN[a]	0.78 ± 0.03	0.76 ± 0.02	0.72 ± 0.03	0.74 ± 0.02	0.69 ± 0.04	0.81 ± 0.03
MOGONET [22]	0.83 ± 0.5	0.77 ± 0.4	0.66 ± 0.2	0.65 ± 0.3	0.65 ± 0.3	0.86 ± 0.4
pca_Xgboost	0.86 ± 0.02	0.78 ± 0.04	0.69 ± 0.06	0.80 ± 0.07	0.61 ± 0.07	0.89 ± 0.02
pca_svm	0.85 ± 0.02	0.80 ± 0.03	0.70 ± 0.04	0.73 ± 0.06	0.69 ± 0.07	0.90 ± 0.02
DeepMO [13]	0.80 ± 0.03	0.78 ± 0.04	0.61 ± 0.04	0.52 ± 0.04	0.8 ± 0.03	0.86 ± 0.01
CustOmics [40]	0.72 ± 0.05	0.68 ± 0.05	0.73 ± 0.06	0.86 ± 0.06	0.72 ± 0.05	0.87 ± 0.04
SNF_SVM	0.70 ± 0.06	0.61 ± 0.06	0.41 ± 0.05	0.38 ± 0.06	0.44 ± 0.06	0.33 ± 0.05
tSNE_SVM	0.58 ± 0.05	0.56 ± 0.04	0.38 ± 0.04	0.29 ± 0.05	0.52 ± 0.05	0.61 ± 0.04
ConsensusClustering_SVM	0.68 ± 0.02	0.67 ± 0.03	0.50 ± 0.03	0.40 ± 0.03	0.67 ± 0.02	0.72 ± 0.02
Pan-kidney cohort						
KD_SVAE_VCDN[a]	0.90 ± 0.01	0.84 ± 0.02	0.81 ± 0.02	0.93 ± 0.04	0.71 ± 0.04	0.93 ± 0.02
KD_AE_VCDN[a]	0.87 ± 0.02	0.79 ± 0.04	0.73 ± 0.05	0.88 ± 0.06	0.82 ± 0.03	0.89 ± 0.03
KD_SVAE_NN[a]	0.88 ± 0.02	0.82 ± 0.04	0.78 ± 0.05	0.91 ± 0.04	0.69 ± 0.07	0.91 ± 0.02
SVAE_VCDN[a]	0.79 ± 0.02	0.73 ± 0.03	0.69 ± 0.04	0.82 ± 0.05	0.60 ± 0.04	0.81 ± 0.03
MOGONET [22]	0.77 ± 0.2	0.74 ± 0.3	0.60 ± 0.2	0.81 ± 0.3	0.78 ± 0.3	0.86 ± 0.2
pca_Xgboost	0.86 ± 0.04	0.73 ± 0.04	0.61 ± 0.06	0.88 ± 0.07	0.54 ± 0.05	0.80 ± 0.04
pca_svm	0.83 ± 0.05	0.72 ± 0.05	0.59 ± 0.06	0.63 ± 0.08	0.53 ± 0.05	0.76 ± 0.04
DeepMO [13]	0.85 ± 0.04	0.82 ± 0.03	0.72 ± 0.03	0.67 ± 0.04	0.79 ± 0.04	0.87 ± 0.02
CustOmics [40]	0.72 ± 0.04	0.68 ± 0.05	0.73 ± 0.06	0.86 ± 0.05	0.72 ± 0.05	0.87 ± 0.05
SNF_SVM	0.68 ± 0.03	0.57 ± 0.04	0.37 ± 0.03	0.42 ± 0.03	0.33 ± 0.04	0.63 ± 0.03
tSNE_SVM	0.66 ± 0.05	0.61 ± 0.05	0.46 ± 0.06	0.42 ± 0.05	0.50 ± 0.04	0.64 ± 0.04
ConsensusClustering_SVM	0.67 ± 0.03	0.66 ± 0.03	0.60 ± 0.03	0.52 ± 0.03	0.69 ± 0.03	0.72 ± 0.03

[a] The different variations of the proposed model.

4.5. Pathway analysis and biomarker discovery

Fig. 6 shows the top impacted pathways and their corresponding FDR adjusted p-values associated with aggressive breast carcinoma (<3 years of survival) using impact analysis [41]. Significant genes are presented in the diagram with their fold-changes color-coded. The results for KIPAN are listed in Figure S3.

We also conducted upstream miRNA analysis and identified several miRNAs as potential biomarkers. The prediction of active miRNAs is based on enrichment of differentially downregulated target genes of the miRNAs. In general, miRNAs have an inhibitory effect on their targets. Therefore, for any given miRNA the method computes the ratio between the number of differentially downregulated targets, and all differentially expressed targets, and compares it to the ratio of all downwardly expressed targets to all targets. Overall, we calculate the probability of observing at least the number of differentially downregulated target genes for a given miRNA just by chance. This p-value is computed using the hypergeometric distribution. Figure S1 and S2 show the list of miRNAs with statistically significant ($p < 0.05$) FDR adjusted p-values along with the number of differentially expressed targets for BRCA and KIPAN.

5. Discussion

Omics technologies have rapidly advanced personalized medicine by using molecular-level data with unprecedented details. Thus, it has become increasingly important to leverage these omics data for supervised learning problems such as disease progression prediction. To this end, KD_SVAE_VCDN is proposed as a supervised multi-omics integration method for patients' survival prediction, where each omics data type is considered as one view of samples. The proposed KD_SVAE_VCDN model utilizes a knowledge distillation framework for incomplete multi-omics data integration in which SVAEs aims to learn omics specific features and VCDN tries to capture cross-omics correlations at the high-level latent representation space effectively. By conducting ablation studies, we have shown that using the KD structure and including incomplete data in the training phase for prediction of complete multi-modal data, we are able to achieve superior results compared to models using only common samples among modalities. Also, using SVAE as the main model in the KD_SVAE_VCDN architecture with VCDN for data integration was essential for successful multi-omics data integration and classification.

Furthermore, our model was trained on each individual type of omics data as well as various combinations of them, and the results were compared, as depicted in Fig. 5b. The performance of the model varied across different levels of combinations. Specifically, the results indicated that the combination of methylation, miRNA, and mRNA data outperformed other combinations. This suggests that the integration of these three modalities yields the most favorable outcomes in terms of model performance. Notably, DNA methylation was found to be the most effective in predicting breast cancer progression, followed by miRNA data. These results are consistent with findings in the literature. DNA methylation is a major epigenetic alteration that is commonly perturbed in breast cancers [13,28,39]. Note that, using only mRNA expression data in the model did not yield meaningful results, and hence its performance was not included in the plot for clarity.

Moreover, the learning curve plotted in Fig. 5c shows the effect of the amount of distilled knowledge transferred from the teachers which includes a combination of two different omics data on the student model. This observation reflects that when a = 1, b = 0.1, and c = 10 the loss, D_{KL}, has more stability. This further explains that the higher weight for the distilled knowledge from the integration of miRNA and methylation lead the final student loss to have less divergency. On the other hand, when a = 1, b = 10, and c = 0.1 (i.e., more distilled knowledge is transferred from the teachers to mRNA and miRNA data during training), there is a higher unstable loss in the student model. This shows that larger the weight for the distilled knowledge (such as the case for miRNA and methylation) the better the results in the student model.

In addition, the learning curve illustrates the impact of the amount of distilled knowledge from the previous teacher, which involves the integration of two distinct omics data, on the student model. This

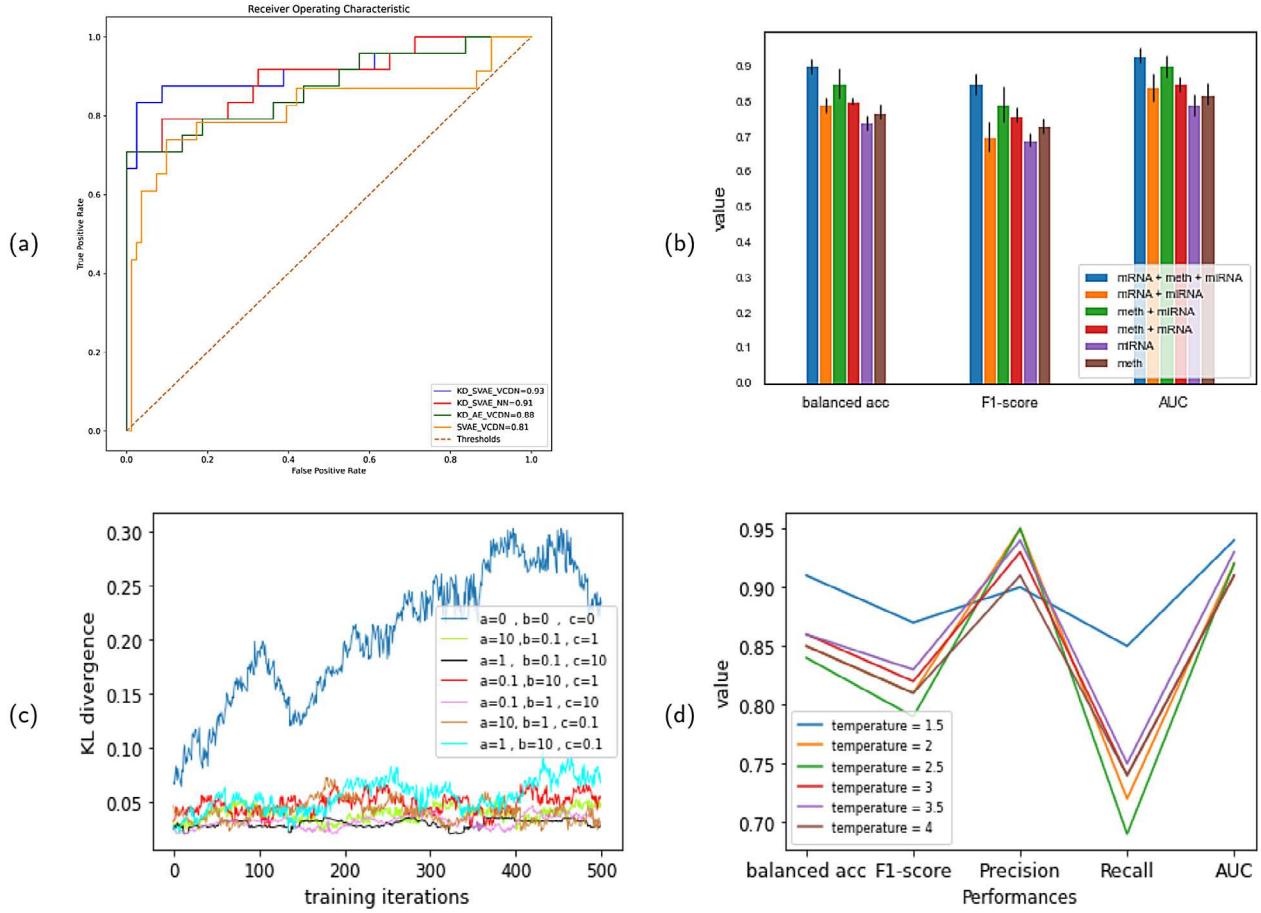


Fig. 5. (a) Performance comparison using different model variations for predicting the progression of breast cancer. (b) Performance comparison of multi-omics data integration on different combination of modalities using the proposed KD_SVAE_VCDN model. Results are generated on test data across n = 30 runs. (c) Comparison of learning curves for KL divergence at different weights. (d) Student model performance comparison for different temperature values in KL divergence loss.

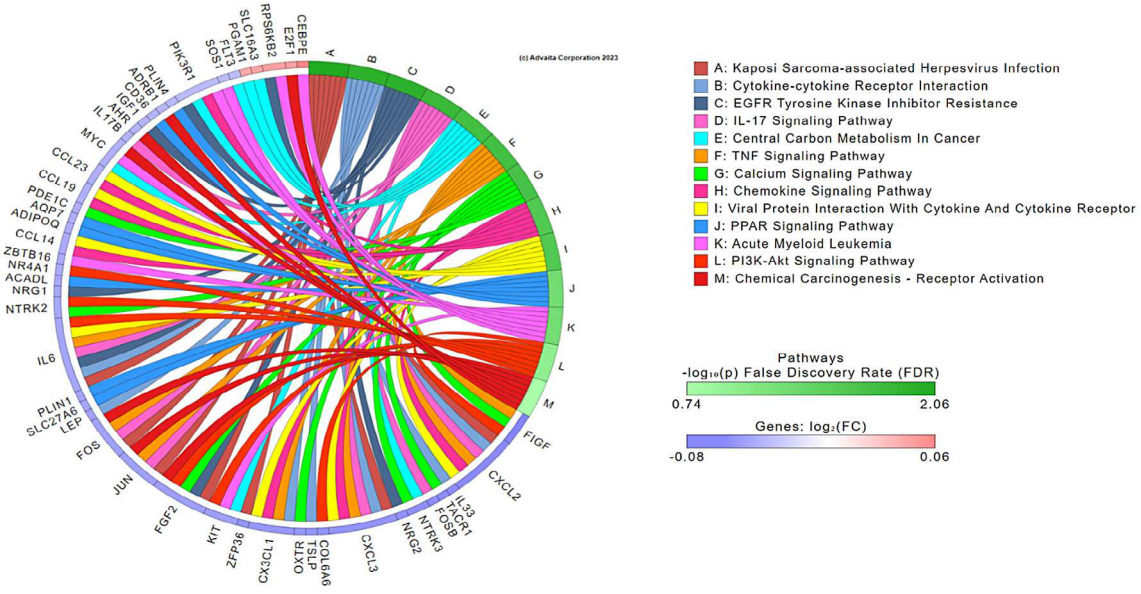


Fig. 6. Significantly impacted pathways for aggressive breast carcinoma along with the set of genes within each pathway. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observation reveals that when the hyperparameters are set to a = 1, b = 0.1, and c = 10, the loss D_{KL} exhibits greater stability. This suggests that assigning a higher weight to the distilled knowledge obtained from the integration of miRNA and methylation data results in a more consistent and less divergent student loss. Conversely, when the hyperparameters are set to a = 1, b = 10, and c = 0.1, implying

a higher weight for the distilled knowledge from mRNA and miRNA data, the student model exhibits a more unstable loss. This indicates that a larger weight for the distilled knowledge obtained from the combination of miRNA and methylation data yields superior results in the student model.

Our results, as illustrated on Fig. 6 have shown that HHV-8, also known as Kaposi's sarcoma-associated herpesvirus (KSHV) which causes Kaposi's sarcoma, a cancer commonly occurring in patients with AIDS, is identified as a significantly impacted pathway on patients with aggressive breast cancer. Several studies suggested that HHV-8 was related to breast cancer by immuno-serological testing, PCR and southern hybridization [42,43] We identified the chemokine signaling pathway as another significantly impacted pathway. The chemokine receptors have been reported as prognostic markers in breast cancer metastasis [44] which confirms our findings. Inflammation has emerged as a pivotal factor in various stages of tumor development, encompassing initiation, promotion, angiogenesis, and metastasis. Notably, cytokines occupy a significant role in driving these processes [45,46] The findings presented within this dataset underscore the substantial influence of the family of cytokine pathways, implying their integral involvement in governing both the initiation and protection mechanisms associated with breast cancer. The significance of central carbon metabolism in the advancement of mammary carcinoma has also been underscored [47] Furthermore, the role of EGRF Tyrosine Kinase Inhibitor has been highlighted in HER2-enriched breast cancer [48], and the stimulation of breast cancer growth through the up-regulation of the oncoprotein hepatitis B X-interacting protein has been associated with TNF-α all confirming our findings [49]

MicroRNA-124 is reported to suppress the invasion and proliferation of breast cancer cells in several studies [50,51] Dong et al. reported the decreased expression of miR-124 as a cause of tumor progression and poor prognosis in patients with breast cancer [40] Further studies are required to investigate the role of miR-124 in patients with aggressive breast cancer. Note that, these data were analyzed in the context of pathways obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Release 100.0+/11-12, Nov 21) and miRNAs from the miRBase (MIRBASE Version:22.1,10/18).

Regarding the limitations of our study, we acknowledge the challenges posed by the scarcity of common samples among multi-omics datasets and the limited availability of data for each modality. This constraint prompted us to carefully select a subset of cancer types to ensure meaningful integration and robust results. Additionally, the necessity to establish a reliable and accurate labeling strategy compelled us to exclude samples with survival days falling between short- and long- survival intervals.

We identified some directions for future research including (i) Evaluation of generalizability: This involves evaluating the generality of the proposed approaches on diverse datasets and biological systems. This could include benchmarking the performance of the KD_SVAE_VCDN on different datasets with varying characteristics, such as different data sizes, and data distributions to assess its robustness and generalizability across different contexts. (ii) Extending to multiple sources: While the experiments in the current study focused on integrating data from three heterogeneous modalities, there is potential to further extend the model by integrating data from additional modalities from different sources. This could involve incorporating additional types of omics data, such as transcriptomics, proteomics, or epigenomics, to capture a more comprehensive view of the biological system. Empirical evaluation of these approaches could help assess the model's generality and performance in handling diverse data types. (iii) Extension to imaging data: While the current study focused on integrating multi-omics data, future work could involve extending the proposed approaches to incorporate imaging data. Integration of histopathological images with multi-omics data can enrich the information which could lead to more comprehensive view and better prediction results. (iv) Interpretability and explainability: While KD_SVAE_VCDN can learn complex representations of data, the learned representations may not always be easily interpretable or explainable. Developing methods or techniques to improve the interpretability and explainability of the learned representations could facilitate the adoption of the approach in real-world applications and aid in generating biologically meaningful insights.

6. Conclusion

In this paper, we presented a novel framework, referred to as Knowledge Distillation and Supervised Variational AutoEncoders utilizing View Correlation Discovery Network (KD_SVAE_VCDN), for integrative analysis of multi-omics data with limited common samples. By leveraging variational autoencoders and knowledge distillation techniques, our framework unlocks the full information potential among multiple modalities, allowing for a more accurate and robust understanding of disease progression. Our approach addresses the challenges associated with integrating multi-omics data, such as high dimensionality, data heterogeneity, inconsistent data distributions, and small number of common samples, which are commonly faced in traditional approaches.

CRediT authorship contribution statement

Sima Ranjbari: Conceived & designed the project, Performed the experiments, Analyzed the data & results, Wrote the paper. **Suzan Arslanturk:** Conceived & designed the project, Analyzed the data & results, Wrote the paper.

Declaration of competing interest

None Declared.

Data availability

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. The source code will be publicly available at https://github.com/Sima-Ranjbari upon publication.

Acknowledgments

This research was funded by the National Science Foundation (NSF: #1948338), and the Department of Defense (DoD: #W81XWH-21-1-0570). Both authors have read and agreed to the published version of the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104512.

References

[1] Eugene Lin, Hsien-Yuan Lane, Machine learning and systems genomics approaches for multi-omics data, Biomarker Res. 5 (2017) 1–6.
[2] Sijia Huang, Kumardeep Chaudhary, Lana X. Garmire, More is better: recent progress in multi-omics data integration methods, Front. Genet. 8 (2017) 84.
[3] Bingbing Xie, Zifeng Yuan, Yadong Yang, Zhidan Sun, Shuigeng Zhou, Xiangdong Fang, MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine, Breast Cancer Res. Treat. 169 (2018) 625–632.
[4] Kaiyue Zhou, Bhagya Shree Kottori, Seeya Awadhut Munj, Zhewei Zhang, Sorin Draghici, Suzan Arslanturk, Integration of multimodal data from disparate sources for identifying disease subtypes, Biology 11 (3) (2022).
[5] Bin Baek, Hyunju Lee, Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data, Sci. Rep. 10 (1) (2020) 18951.
[6] So Yeon Kim, Hyun-Hwan Jeong, Jaesik Kim, Jeong-Hyeon Moon, Kyung-Ah Sohn, Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies, Biol. Direct 14 (1) (2019) 1–13.

[7] Zhaoxiang Cai, Rebecca C. Poulos, Jia Liu, Qing Zhong, Machine learning for multi-omics data integration in cancer, Iscience (2022) 103798.
[8] Hyun Jae Cho, Mia Shu, Stefan Bekiranov, Chongzhi Zang, Aidong Zhang, Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment, Bioinformatics 39 (4) (2023) btad113.
[9] Min Yang, Huandong Yang, Lei Ji, Xuan Hu, Geng Tian, Bing Wang, Jialiang Yang, A multi-omics machine learning framework in predicting the survival of colorectal cancer patients, Comput. Biol. Med. 146 (2022) 105516.
[10] Yong Jin Heo, Chanwoong Hwa, Gang-Hee Lee, Jae-Min Park, Joon-Yong An, Integrative multi-omics approaches in cancer research: From biological networks to clinical subtypes, Mol. Cells 44 (7) (2021) 433–443.
[11] zhuohui Wei, Yue Zhang, Wanlin Weng, Jiazhou Chen, Hongmin Cai, Survey and comparative assessments of computational multi-omics integrative methods with multiple regulatory networks identifying distinct tumor compositions across pan-cancer data sets, Brief. Bioinform. 22 (3) (2021).
[12] Xiujing He, Xiaowei Liu, Fengli Zuo, Hubing Shi, Jing Jing, Artificial intelligence-based multi-omics analysis fuels cancer precision medicine, Semin. Cancer Biol. 88 (2023) 187–200.
[13] Yuqi Lin, Wen Zhang, Huanshen Cao, Gaoyang Li, Wei Du, Classifying breast cancer subtypes using deep neural networks based on multi-omics data, Genes 11 (8) (2020) 888.
[14] Li Zhang, Chenkai Lv, Yaqiong Jin, Ganqi Cheng, Yibao Fu, Dongsheng Yuan, Yiran Tao, Yongli Guo, Xin Ni, Tieliu Shi, Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma, Front. Genet. 9 (2018) 477.
[15] Zhiwei Rong, Zhilin Liu, Jiali Song, Lei Cao, Yipe Yu, Mantang Qiu, Yan Hou, MCluster-VAEs: An end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data, Comput. Biol. Med. 150 (2022) 106085.
[16] Sushmita Paul, et al., Capturing the latent space of an autoencoder for multi-omics integration and cancer subtyping, Comput. Biol. Med. 148 (2022) 105832.
[17] Hu Song, Chengwei Ruan, Yixin Xu, Teng Xu, Ruizhi Fan, Tao Jiang, Meng Cao, Jun Song, Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model, Exp. Biol. Med. 247 (11) (2022) 898–909.
[18] Jiudi Lv, Junjie Wang, Xiujuan Shang, Fangfang Liu, Shixun Guo, Survival prediction in patients with colon adenocarcinoma via multiomics data integration using a deep learning algorithm, Biosci. Rep. 40 (12) (2020).
[19] Hua Chai, Xiang Zhou, Zhongyue Zhang, Jiahua Rao, Huiying Zhao, Yuedong Yang, Integrating multi-omics data through deep learning for accurate cancer prognosis prediction, Comput. Biol. Med. 134 (2021) 104481.
[20] Amina Lemsara, Salima Ouadfel, Holger Fröhlich, Pathme: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data, BMC Bioinform. 21 (2020) 1–20.
[21] Tzu-Hao Wang, Cheng-Yang Lee, Tzong-Yi Lee, Hsien-Da Huang, Justin Bo-Kai Hsu, Tzu-Hao Chang, Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion, Cancers 13 (11) (2021) 2528.
[22] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, Kun Huang, MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nat. Commun. 12 (1) (2021) 3445.
[23] Nima Zafari, Parsa Bathaei, Mahla Velayati, Fatemeh Khojasteh-Leylakoohi, Majid Khazaei, Hamid Fiuji, Mohammadreza Nassiri, Seyed Mahdi Hassanian, Gordon A Ferns, Elham Nazari, et al., Integrated analysis of multi-omics data for the discovery of biomarkers and therapeutic targets for colorectal cancer, Comput. Biol. Med. (2023) 106639.
[24] Kaiyue Zhou, Suzan Arslanturk, Douglas B. Craig, Elisabeth Heath, Sorin Draghici, Discovery of primary prostate cancer biomarkers using cross cancer learning, Sci. Rep. 11 (1) (2021).
[25] Tao Zhou, Mingxia Liu, Kim-Han Thung, Dinggang Shen, Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data, IEEE Trans. Med. Imaging 38 (10) (2019) 2411–2422.
[26] Federico Taverna, Jermaine Goveia, Tobias K Karakach, Shawez Khan, Katerina Rohlenova, Lucas Treps, Abhishek Subramanian, Luc Schoonjans, Mieke Dewerchin, Guy Eelen, et al., BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization, Nucl. Acids Res. 48 (W1) (2020) W385–W394.
[27] Xuehuan He, Rupture Risk Assessment for Ascending Thoracic Aortic Aneurysms: Macroscopic Rupture Pattern and Microstructural Connection (Ph.D. thesis), The University of Iowa, 2022.
[28] Nikola Simidjievski, Cristian Bodnar, Ifrah Tariq, Paul Scherer, Helena Andres Terre, Zohreh Shams, Mateja Jamnik, Pietro Liò, Variational autoencoders for cancer data integration: design principles and computational practice, Front. Genet. 10 (2019) 1205.
[29] Hui Shen, Juliann Shih, Daniel P Hollern, Linghua Wang, Reanne Bowlby, Satish K Tickoo, Vésteinn Thorsson, Andrew J Mungall, Yulia Newton, Apurva M Hegde, et al., Integrated molecular characterization of testicular germ cell tumors, Cell Rep. 23 (11) (2018) 3392–3406.
[30] Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, Yike Guo, Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 765–769.
[31] Abedalrhman Alkhateeb, Li Zhou, Ashraf Abou Tabl, Luis Rueda, Deep learning approach for breast cancer inclust 5 prediction based on multiomics data integration, in: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020, pp. 1–6.
[32] Abedalrhman Alkhateeb, Ashraf Abou Tabl, Luis Rueda, Deep learning in multi-omics data integration in cancer diagnostic, in: Deep Learning for Biomedical Data Analysis: Techniques, Approaches, and Applications, Springer, 2021, pp. 255–271.
[33] Suzan Arslanturk, Sorin Draghici, Tin Nguyen, Integrated cancer subtyping using heterogeneous genome-scale molecular datasets, in: Pacific Symposium on Biocomputing 2020, World Scientific, 2019, pp. 551–562.
[34] Tianle Ma, Aidong Zhang, Integrate multi-omics data with biological interaction networks using multi-view factorization AutoEncoder (MAE), BMC Genom. 20 (2019) 1–11.
[35] Jonathan Mitchel, Kevin Chatlin, Li Tong, May D. Wang, A translational pipeline for overall survival prediction of breast cancer patients by decision-level integration of multi-omics data, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 1573–1580.
[36] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, Riccardo Bellazzi, Integrated multi-omics analyses in oncology: a review of machine learning methods and tools, Front. Oncol. 10 (2020) 1030.
[37] Dongdong Lin, Jigang Zhang, Jingyao Li, Chao Xu, Hong-Wen Deng, Yu-Ping Wang, An integrative imputation method based on multi-omics datasets, BMC Bioinform. 17 (1) (2016) 1–12.
[38] Sarmistha Das, Indranil Mukhopadhyay, TiMEG: an integrative statistical method for partially missing multi-omics data, Sci. Rep. 11 (1) (2021) 1–16.
[39] Qi Wang, Liang Zhan, Paul Thompson, Jiayu Zhou, Multimodal learning with incomplete modalities by knowledge distillation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1828–1838.
[40] Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède, CustOmics: A versatile deep-learning based strategy for multi-omics integration, PLoS Comput. Biol. 19 (3) (2023) e1010921.
[41] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, Roberto Romero, A novel signaling pathway impact analysis, Bioinformatics 25 (1) (2009) 75–82.
[42] Robert Newton, John Ziegler, Dimitra Bourboulia, Delphine Casabonne, Valerie Beral, Edward Mbidde, Lucy Carpenter, Gillian Reeves, D Maxwell Parkin, Henry Wabinga, et al., The sero-epidemiology of Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) in adults with cancer in Uganda, Int. J. Cancer 103 (2) (2003) 226–232.
[43] Chun-Ru Hsu, Tsong-Ming Lu, Lengsu William Chin, Chi-Chiang Yang, Possible DNA viral factors of human breast cancer, Cancers 2 (2) (2010) 498–512.
[44] Debarati Mukherjee, Jihe Zhao, The role of chemokine receptor CXCR4 in breast cancer metastasis, Am. J. Cancer Res. 3 (1) (2013) 46.
[45] Marcela Esquivel-Velázquez, Pedro Ostoa-Saloma, Margarita Isabel Palacios-Arreola, Karen E Nava-Castro, Julieta Ivonne Castro, Jorge Morales-Montor, The role of cytokines in breast cancer development and progression, J. Interferon Cytokine Res. 35 (1) (2015) 1–16.
[46] Joseph Antoine Salvator Fabre, Jérôme Giustiniani, Christian Garbar, Yacine Merrouche, Frank Antonicelli, Armand Bensussan, The interleukin-17 family of cytokines in breast cancer, Int. J. Mol. Sci. 19 (12) (2018) 3880.
[47] Adam D Richardson, Chen Yang, Andrei Osterman, Jeffrey W Smith, Central carbon metabolism in the progression of mammary carcinoma, Breast Cancer Res. Treat. 110 (2008) 297–307.
[48] George Iancu, Dragos Serban, Cristinel Dumitru Badiu, Ciprian Tanasescu, Mihai Silviu Tudosie, Corneliu Tudor, Daniel Ovidiu Costea, Anca Zgura, Raluca Iancu, Danut Vasile, Tyrosine kinase inhibitors in breast cancer, Exp. Ther. Med. 23 (2) (2022) 1–10.
[49] Xiaoli Cai, Can Cao, Jiong Li, Fuquan Chen, Shuqin Zhang, Bowen Liu, Weiying Zhang, Xiaodong Zhang, Lihong Ye, Inflammatory factor TNF-α promotes the growth of breast cancer via the positive feedback loop of TNFR1/NF-κB (and/or p38)/p-STAT3/HBXIP/TNFR1, Oncotarget 8 (35) (2017) 58338.
[50] Nier Cha, Baoqing Jia, Yinzai He, Wei Luan, Wenhua Bao, Xiuhua Han, Weishi Gao, Yanwei Gao, MicroRNA-124 suppresses the invasion and proliferation of breast cancer cells by targeting TFAP4, Oncol. Lett. 21 (4) (2021) 1.
[51] Wei-Luo Cai, Wen-Ding Huang, Bo Li, Tian-Rui Chen, Zhen-Xi Li, Cheng-Long Zhao, Heng-Yu Li, Yan-Mei Wu, Wang-Jun Yan, Jian-Ru Xiao, microRNA-124 inhibits bone metastasis of breast cancer by repressing interleukin-11, Mol. Cancer 17 (1) (2018) 1–14.