biology

MDPI

Article

Integration of Multimodal Data from Disparate Sources for Identifying Disease Subtypes

Kaiyue Zhou [1,2], Bhagya Shree Kottoori [1], Seeya Awadhut Munj [1], Zhewei Zhang [2], Sorin Draghici [1,3] and Suzan Arslanturk [1,*]

[1] Department of Computer Science, Wayne State University, Detroit, MI 48201, USA; kyzhou@wayne.edu (K.Z.); bhagyashree.k@wayne.edu (B.S.K.); seeya.munj@wayne.edu (S.A.M.); sorin@wayne.edu (S.D.)
[2] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; 15111059@bjtu.edu.cn
[3] Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48201, USA
* Correspondence: suzan.arslanturk@wayne.edu

Simple Summary: The diagnostic and treatment strategies of cancer remain generally suboptimal resulting in over-diagnosis or under-treatment. Though many attempts on optimizing treatment decisions by early prediction of disease progression have been undertaken, these efforts yielded only modest success so far due to the heterogeneity of cancer with multifactorial etiology. Here, we propose a deep-learning based data integration model capable of predicting disease progression by integrating collective information available through multiple studies with different cohorts and heterogeneous data types. The results have shown that the proposed data integration pipeline is able to identify disease progression with higher accuracy and robustness compared to using a single cohort, by offering a more complete picture of the specific disease on patients with brain, blood, and pancreatic cancers.

Abstract: Studies over the past decade have generated a wealth of molecular data that can be leveraged to better understand cancer risk, progression, and outcomes. However, understanding the progression risk and differentiating long- and short-term survivors cannot be achieved by analyzing data from a single modality due to the heterogeneity of disease. Using a scientifically developed and tested deep-learning approach that leverages aggregate information collected from multiple repositories with multiple modalities (e.g., mRNA, DNA Methylation, miRNA) could lead to a more accurate and robust prediction of disease progression. Here, we propose an autoencoder based multimodal data fusion system, in which a fusion encoder flexibly integrates collective information available through multiple studies with partially coupled data. Our results on a fully controlled simulation-based study have shown that inferring the missing data through the proposed data fusion pipeline allows a predictor that is superior to other baseline predictors with missing modalities. Results have further shown that short- and long-term survivors of glioblastoma multiforme, acute myeloid leukemia, and pancreatic adenocarcinoma can be successfully differentiated with an AUC of 0.94, 0.75, and 0.96, respectively.

Keywords: multimodal data fusion; imputation; deep learning; cancer progression

Citation: Zhou, K.; Kottoori, B.S.; Munj, S.A.; Zhang, Z.; Draghici, S.; Arslanturk, S. Integration of Multimodal Data from Disparate Sources for Identifying Disease Subtypes. *Biology* **2022**, *11*, 360. https://doi.org/10.3390/biology11030360

Academic Editors: Min Zhao, Ruifeng Hu and Dario Di Silvestre

Received: 28 January 2022
Accepted: 23 February 2022
Published: 24 February 2022

1. Introduction

Patients suffering from the same cancer disease may not only experience a high degree of symptomatic variability but also display significantly different responses to the same treatment. As a result, many cancers are over-diagnosed causing patients to receive unnecessary cancer treatments, while some patients do not receive the needed treatment. In addition, treatment responses vary significantly across different patients due to the complexity of cancer treatment. This can be greatly reduced by an early risk prediction model that can successfully differentiate between patients who are at higher-risk and need

the most aggressive treatments from those who will never progress, recur, or develop resistance to treatments.

Studies have shown that different modalities (including gene expression, DNA Methylation, miRNA, variant data, lifestyle, clinical data), all play an important role towards predicting the development of cancer [1,2]. Van et al., López-García et al., Zhou et al., and Lu et al. have successfully classified tumors subtypes of Acute Lymphoblastic Leukaemia (ALL) using miRNA expression [3–6]. Lauber et al. identified short- and long-term survivors of Acute Myeloid Leukemia (AML) through *DNMT3A*, *FLT3* and/or *NPM1* mutations [7]. Similarly, Jonckheere et al. investigated that membrane bound mucin is responsible for short-term survival in patients with pancreatic adenocarcinoma [8]. Although such findings provide a better understanding and can help individualize treatment decisions, they may not provide a complete picture of the disease.

The human body consists of a mass of interconnecting pathways, working together in symphony. Output of one process, or a pathway, is further used by another process, or a pathway, for proper functioning of the body. Hence, deriving results based on just one modality, (e.g., gene expression) may not provide sufficient information. Plotnikova et al. and Jonas et al. discussed miRNA playing an important role in regulating gene expression [9,10]. Aure et al. studied the interaction of methylation and miRNA and their effect on predicting breast cancer [11]. This further suggests that the investigation of clinically relevant disease subtypes cannot be achieved by analyzing data from a single source. The proposed simulated study aims to achieve a novel data integration methodology that fuses information from disparate sources with different cohorts.

Data fusion refers to the integration of multiple modalities or data sources (i) to obtain a more unified picture and comprehensive view of the relations, (ii) to achieve more robust results, (iii) to improve the accuracy and integrity, and (iv) to illuminate the complex interactions among data features [12–17]. Nguyen et al. [18] have used a perturbation based data integration method to identify disease subtypes. Though powerful, these data integration methods often fail to generate stable results when high dimensional data with limited samples are present, mainly due to their randomized nature. End-to-end adversarial-attention network for multimodal clustering (EAMC) [19] allows the integration of multiple modalities through its discriminator module, which guides its encoders to learn a latent representation distribution by assigning one modality as an anchor to others. In the supervised learning setup, Wang et al. [20] performed a multimodal fusion by channel exchanging (CEN), which dynamically exchanges the features between different modalities to build inter-modal fusion for better segmentation or translation. Though powerful, these methods lack the ability to integrate data from heterogeneous sources with multiple modalities.

One straightforward approach to address the heterogeneous data sources' integration is direct concatenation, which treats features extracted from different sources equally, by concatenating them into a feature vector. Treating different datasets equally by simply concatenating the features from disparate sources cannot achieve good performance due to different representation, distribution, scale, and density of data [21]. It also leads to challenges such as overfitting due to increased dimensionality of data after concatenation, as well as data redundancies and dependencies [22].

Multimodal data fusion [14] allows different datasets or different feature subsets of an object to be combined to describe the object comprehensively and accurately. The proposed research contributes to the young but growing field of multimodal data fusion with shared and unshared (modality-specific) data elements. Using heterogeneous datasets from disparate sources often lead to data blocks with partially shared features where observations (e.g., patients) from different sources differ in terms of the feature sets. Augmented multimodal setting allows an arbitrary collection of heterogeneous data sources to be partially coupled (i.e., one-to-one correspondence) through shared entities [23], which is illustrated in Figure 1. When multiple datasets (i.e., sources) with different modalities concerning an object exists, they cannot be simply merged into a single matrix for complementation of missing values due to each dataset having different distributions or feature dimensions.

Traditional methods, such as coupled matrix and tensor factorization [24–27], and context-aware tensor decomposition [28], are used for joint matrix factorization analysis of partially coupled data from multiple platforms by transferring the similarity between object pairs learned from one dataset to the other for more accurate complementation of missing values.

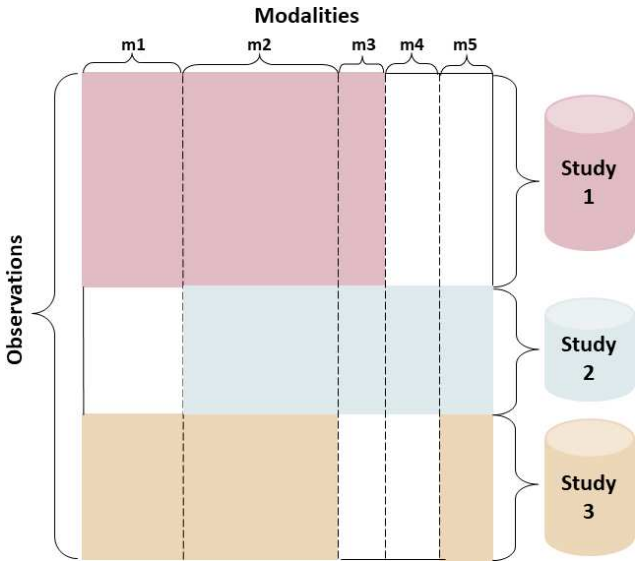


Figure 1. An illustration of data consisting of multiple sources with shared (i.e., m2) and unshared modalities (i.e., m1, m3, m4, and m5).

Several alternative methods proposed for the complementation of missing values are summarized below. Yang et al. [29] proposed a semi-supervised learning approach to vote the predictions made by each modality individually. The incomplete modalities were filled with zeros, which may lead to biased predictions when high amounts of missing data are present. A computational approach based on deep neural networks to predict methylation states in single cells (DeepCpG) [30] was proposed to predict methylation states in single cells. Similarly, Yu et al. [31] imputed the missing DNA methylation values using a mixture regression model. Zhou et al. [32] proposed an autoencoder-like architecture that imputes the missing mRNA values by a nonlinear mapping from DNA methylation to gene expression data. The network was trained on a large scale pan-cancer dataset and then specifically fine-tuned for a targeted cancer through transfer learning. Bischke et al. [33] proposed a generative adversarial network to synthesize information from multiple modalities through a segmentation network. Ma et al. [34] proposed a reconstruction network, referred to as multimodal learning with severely missing modality (SMIL), that outputs a posterior distribution from which the missing modality is reconstructed through sampling using modality priors. With the priors learned from the existing modalities, such meta-learning framework predicts an embedding for missing modalities that can then be used for subsequent classification tasks. The cascaded residual autoencoder [35] was explored to impute the missing data by learning the complex relationship among certain modalities. This design required the network to compute the loss between each pair of cascaded autoencoder blocks, which may significantly increase the trainable parameters. Learning to recommend with missing modalities (LRMM) [36] is a controlled simulated study that randomly removes several features during training followed by the reconstruction of missing modalities through a generative autoencoder. This approach is similar to our proposed model, which is able to generate robust results, even with sparse or entirely missing modalities using image and textual data. However, the data used in this study contain spatial and/or temporal information that is mostly lacking in genomic data. Though powerful, classical data fusion approaches fail to integrate information from multiple data sources with disparate populations consisting of unshared modalities that are completely missing in one source while preserving the patient level information for further prediction tasks.

To address these issues, we propose a deep learning based data integration technique able to perform joint analysis on disparate heterogeneous datasets by discovering the salient knowledge of missing modalities. This is achieved by learning the latent association between existing and missing modalities followed by subsequent reconstruction of missing modalities. Our contributions are summarized as follows:

1. To the best of our knowledge, our approach is the first study that aims to discover the salient genetic knowledge of a completely missing modality through a mapping function learned by the neural network. This neural network model is able to reconstruct a lower dimensional representation of the missing information based on the correlation between shared and unshared modalities across data sources. Such mapping provides the ability to produce more accurate and consistent identification of aggressive and indolent patients for lethal cancers;
2. We have discovered patient subgroups and disease subtypes that have significantly different survival patterns through an unsupervised learning approach combined with manual adjustments which was then used for labeling the samples;
3. We quantitatively demonstrate that our work outperforms other baselines with partially available modalities.

In this paper, due to the small sample size, we adopt an autoencoder-like framework [37,38] for the compression and reconstruction of each modality. The lower dimensional latent representation will be used for the classification tasks. Similar to SMIL, we train our network to learn the approximated priors of the missing modality with respect to existing modalities. In our approach, a modality will be completely missing during the inference phase. Another substantial difference from previous studies is that our design only contains dense layers as our data do not have any spatial or temporal information.

2. Method

In this study, we aim to integrate the knowledge from disparate sources with shared and unshared modalities for a multimodal classification task. As seen in Figure 1, the data can consist of multiple sources (i.e., studies) with heterogeneous populations. In this paper, we conduct a simulated study by defining two heterogeneous populations, namely the reference data and the target data. Both the reference and the target data may contain shared and unshared modalities. The reference data are considered as the model's training set for learning associations between its shared and unshared modalities, and the target data are considered as the testing set where the representation of an unshared modality that is completely missing (within the test set) can be inferred through the trained model. We formally define the reference and target data as $M^r = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^{d_i}, i = 1, 2, 3, \ldots\}$ and $M^t = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^{d_i}, i = 1, 2, \ldots\}$, respectively. In our simulated study, all modalities in M^r namely mRNA, DNA methylation, and miRNA are present, and one modality in M^t namely DNA methylation is absent. We assume that the ground truth labels (e.g., aggressive vs. indolent tumor) for the reference and target data, Y^r and Y^t, are known.

In case when all modalities are available in the target data, a simple autoencoder-like network [38,39] can be trained for multimodal classification. Particularly, Zhou et al. [38] showed that such an autoencoder-based classifier can handle data with high-dimensionality and limited sample size. However, in practice, certain modalities can be completely missing preventing the construction of a robust classifier. Therefore, our objective is to learn a mapping between the shared and unshared modalities using other sources (i.e., our training data/reference), and predict a lower dimensional representation of the missing modalities within the testing set. As such, we propose the architectures in Section 2.1.

2.1. Learning Associations between Shared and Unshared Modalities

For the aforementioned purpose, we have designed three independent network architectures for different scenarios. Different from commonly seen studies, all layers in our proposed architectures are dense, as we only consider vector-like data. Moreover, the

elements in a single vector do not have any spatial or temporal information. The detail of the fundamental baseline architecture is provided in Table 1, which will be discussed in Section 2.1.1.

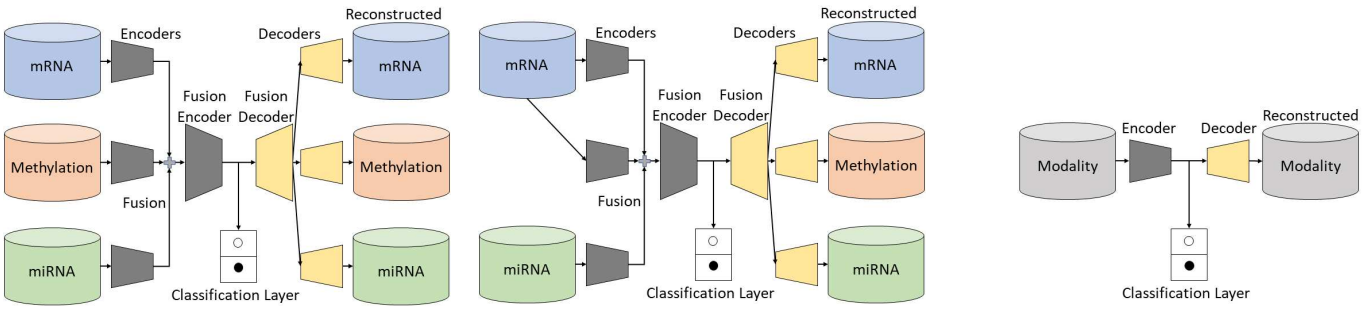
Table 1. The detailed number of units in the backbone architecture excluding the first and last layers.

Module	Neurons in Layer 1	Neurons in Layer 2
Encoder 1	1024	256
Encoder 2	1024	256
Encoder 3	64	-
Fusion Encoder	576	36
Fusion Decoder	512	-

Instead of imputing severely missing values as seen in several other studies (up to, e.g., 90%), we propose a method that can learn the embedding of a completely missing modality specific to a single source (i.e., DNA methylation in our study) through the knowledge available in the other existing data source(s), which will be discussed in Section 2.1.2.

2.1.1. Complete Fusion Autoencoder

When all shared and unshared modalities are present, the complete fusion autoencoder (CFA) in Figure 2a is used to fuse the different inputs into a binding latent representation, which will then be decoded to reconstruct the corresponding modalities. The latent representation of the fused modalities is used for the classification task. For simplicity, we denote the encoders and decoders as single layers. The fused feature $Z = fuse(\{\mathbf{z}_i | \mathbf{z}_i \in \mathbb{R}^{h_i}, i = 1,2,3\})$ highly compresses the prior knowledge of $\{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^{d_i}, i = 1,2,3\}$ and is fed to another classification layer, where the $fuse$ function is the concatenation or averaging operation, and h_i denotes the hidden dimension of modality i. More specifically, the mapping f of $\mathbf{x}_i \rightarrow \mathbf{z}_i \rightarrow \hat{\mathbf{x}}_i$ can be denoted as $\mathbf{z}_i = act(W_i^1 \mathbf{x}_i + b^1)$ and $\hat{\mathbf{x}}_i = act(W_i^2 \mathbf{z}_i + b^2)$, where $W_i^1 \in \mathbb{R}^{d_i \times h_i}, b^1 \in \mathbb{R}^{h_i}, W_i^2 \in \mathbb{R}^{h_i \times d_i}, b^2 \in \mathbb{R}^{d_i}$, and act is the non-linear activation function (rectified linear unit (ReLU)). We apply 1D batch normalization on each block (consisting of dense and ReLU layers) in the encoders to alleviate internal covariate shift [40].



(**a**) Complete fusion autoencoder. (**b**) Incomplete fusion autoencoder. (**c**) Single-modal autoencoder.

Figure 2. The overview of our proposed architectures. As the backbone, the complete fusion autoencoder (**a**) has three separate encoders for the modalities, i.e., encoder 1, encoder 2, and encoder 3. The fusion encoder then merges the learned representations from each thread for latter usage. Without losing generality, encoder 2 in (**b**) takes the first modality as input, to learn the correlation between mRNA and methylation. (**c**) simply takes every individual modality as input, respectively.

The optimizer of CFA utilizes the L2 loss as the reconstruction loss between each pair of $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$:

$$L_{rec} = \sum_{i=1}^{M} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2, \quad (1)$$

where M is the number of possible modalities (three in our study).

For each modality, let p and q be the prior and posterior distributions, $\exists \mathbf{x}_i \, s.t. \, p(\mathbf{z}_i) \approx p(\mathbf{x}_i)$, where $p(\mathbf{x}_i)$ is hard to measure. Given the available modalities, $p(\mathbf{z}_i)$ can be approximated by $p_\phi(\mathbf{z}_i)$ through Equation (1), where ϕ denotes the trainable parameters in the encoders and decoders.

2.1.2. Incomplete Fusion Autoencoder

In the event that an unshared modality (e.g., $\mathbf{x}_2$) is completely missing in the testing set, we train the network with the reference data to learn a mapping function f' of $\mathbf{x}_1 \rightarrow \mathbf{z}_{2|1} \rightarrow \hat{\mathbf{x}}_2$, where $\mathbf{z}_{2|1}$ denotes the hidden feature representation of $\hat{\mathbf{x}}_2$ given $\mathbf{x}_1$. The reconstructed output $\hat{\mathbf{x}}_2$ is optimized using the L2 loss according to $\mathbf{x}_2$. As there are still reconstructions for the existing modalities ($\mathbf{x}_1 \rightarrow \hat{\mathbf{x}}_1$ and $\mathbf{x}_3 \rightarrow \hat{\mathbf{x}}_3$), we name this network as incomplete fusion autoencoder (IFA) for better illustration. The framework of IFA is shown in Figure 2b.

In CFA, all $p(\mathbf{x}_i)$ can be approximated by f. However in IFA, $\nexists \mathbf{x}_2 \, s.t. \, p(\mathbf{z}_2) \approx p(\mathbf{x}_2)$. Therefore, we use an encoder to learn the approximation $p(\mathbf{x}_2|\mathbf{x}_1)$ through f', such that:

$$\begin{aligned} p(\mathbf{x}_2) &\approx p(\mathbf{x}_2|\mathbf{x}_1) \approx p(\mathbf{z}_2|\mathbf{x}_1) \approx p(\mathbf{z}_2) \\ &\approx p_\phi(\mathbf{z}_{2|1}|\mathbf{x}_{2|1}) \approx p_\phi(\mathbf{z}_{2|1}|\mathbf{x}_1), \end{aligned} \tag{2}$$

where the true posterior $p(\mathbf{z}_2|\mathbf{x}_1)$ is estimated by the distribution of $p_\phi(\mathbf{z}_{2|1}|\mathbf{x}_1)$ which is learned by the network, representing the prior distribution of the missing modality, i.e., $p(\hat{\mathbf{x}}_2)$. We assume that Equation (2) is satisfied if $\mathbf{x}_{2|1}$ highly correlates with $\mathbf{x}_1$.

The procedure for calculating a mapping function $f' = \mathbf{x}_1 \rightarrow \mathbf{z}_{1|2} \rightarrow \hat{\mathbf{x}}_2$ is described as $\mathbf{z}_{2|1} = act(W_2^1 \mathbf{x}_1 + b^1)$, $\hat{\mathbf{x}}_2 = act(W_2^2 \mathbf{z}_{2|1} + b^2)$, where $W_2^1 \in \mathbb{R}^{d_1 \times h_2}, b^1 \in \mathbb{R}^{h_2}, W_2^2 \in \mathbb{R}^{h_2 \times d_2}$. Since $\hat{\mathbf{x}}_2$ is still obtained in this thread, the reconstruction loss is same as Equation (1). The other threads for modality 1 and 3 remain the same as in Section 2.1.1.

2.1.3. Single-Modal Autoencoder

A single-modal autoencoder (SMA) is a standard autoencoder. We use such a network to perform baseline studies with respect to each single modality, as shown in Figure 2c. The number of units in each network may slightly differ for different modalities due to different data dimensions.

2.1.4. Classification Layer

All three architectures described above have a classification layer for the prediction task, which consists of a dense layer and a sigmoid layer. We adopt the commonly used cross entropy loss for this classification task:

$$L_{ce} = -\sum_{j=1}^{C} Y_j log(Y_j'), \tag{3}$$

where Y_j denotes the ground truth label, Y_j' denotes the predicted probability of the j'th class, and C denotes the number of classes. Particularly for IFA, let g be a function in this classification layer, such that $Y' = g[f(\mathbf{x}_1), f'(\mathbf{x}_1), f(\mathbf{x}_3)]$, where f' maps $\mathbf{x}_1$ to the posterior distribution of the missing modality $\mathbf{x}_2$ based on the assumption made in Equation (2).

2.2. Joint Loss Optimization

As the weights in the autoencoder and classification layer are updated simultaneously, the objective is to minimize the joint loss:

$$L_{joint} = \alpha L_{rec} + \beta L_{ce}, \tag{4}$$

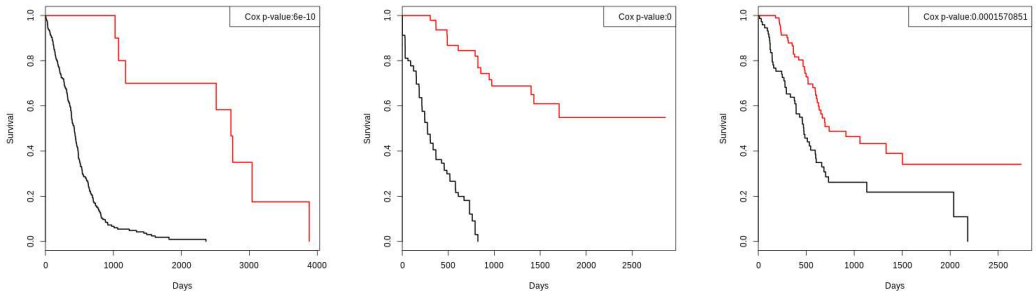
where α and β are the ratios for each loss. We empirically set them both to 1.

3. Experiments

In order to evaluate the effectiveness of the proposed IFA, we conduct several experiments using glioblastoma multiforme (GBM), acute myeloid leukemia (LAML), and pancreatic adenocarcinoma (PAAD) datasets by predicting patients' disease progression status (short- vs. long-term survivors) and report the balanced accuracy and ROC AUCs.

3.1. Data Preparation

Here, we use the preprocessed GBM, LAML, and PAAD data (the three datasets or cancers hereafter) from The Cancer Genome Atlas (TCGA) as fully controlled simulated studies. There are a total of 273, 143, and 175 patients in the three datasets, respectively, each containing three separate modalities (mRNA, DNA Methylation, and miRNA). The data are normalized along the feature dimensions in all our analyses. In order to mimic the situation that one of the modalities is completely missing, for each of these cancers, we randomly select 36% of the samples as our testing set and remove their associated m_2 (i.e., DNA methylation) modality, keeping only m_1 (i.e., gene expression) and m_3 (i.e., miRNA). The remainder of samples along with their three modalities (m_1, m_2, and m_3) are reserved for training. Due to the limited sample sizes, we only simulate circumstances of two data sources in this study. The dimensions of m_1, m_2, and m_3 are shown in Table 2.



(**a**) (**b**) (**c**)

Figure 3. The Kaplan–Meier survival curves of the three cancers generated using our labeling strategy, indicating the reliability of our ground-truth labels. Here the black curves are representing the short-term survivors and the red curves are representing the longer-term survivors. (**a**) GBM. (**b**) LAML. (**c**) PAAD.

Our goal is to understand disease progression by classifying patients as short- and long-term survivors at the time of diagnosis and the approach discussed below is used to define labels for each sample in the train and validation sets for model building. In order to annotate the samples as short- and long-term survivors (or aggressive vs. indolent), we rely on an unsupervised learning algorithm, referred to as perturbation clustering for data integration and disease subtyping (PINS) [18], that is shown to be effective in subtype discovery using molecular data. PINS utilizes a consensus-like voting mechanism to select the best number of clusters among all modalities based on the k-means algorithm. The agreement between modalities is calculated to identify sub-populations through a hierarchical clustering approach. Next, we make some manual corrections on several observations that are classified incorrectly (outliers). For instance, if a GBM patient has lived only for 203 days after diagnosis, and is clustered as a long-term survivor by PINS, we manually correct that subject's class label. After such minor corrections, we achieve two patient subgroups for each cancer annotated as short- and long-term survivors. The subtypes are validated using Kaplan–Meier analysis, and their statistical significance is assessed using Cox regression. The survival curves in Figure 3 show a clear and statistically significant separation between two groups of patients after the manual setup of labels generated through PINS, indicating the reliability of our ground truth labeling. After annotation, we end up having two groups, i.e., aggressive and indolent samples for different cancers as shown in Table 2.

3.2. Implementation Details

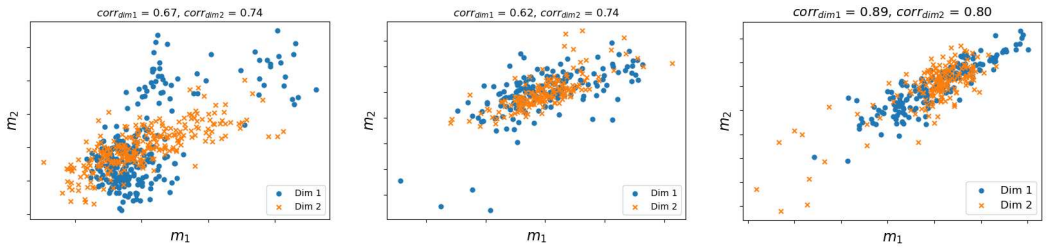
We implement our networks in PyTorch 1.4.0 with NVIDIA Titan RTX GPU. The Adam optimizer with a learning rate of 0.01 is used for training. All proposed models are trained for 100 epochs. The simple yet effective design of our network resulted in a run time of less than 2 min for a 5-fold cross validation model.

Table 2. Dimensions of each modality (m_1, m_2, m_3) and the number of samples in each subtype for different cancers.

Cancer Type	Gene Expression (m_1)	DNA Methylation (m_2)	miRNA (m_3)	Short-Term Survival	Long-Term Survival
GBM	12,042	22,833	534	253	20
LAML	16,818	22,288	552	91	52
PAAD	14,105	20,006	257	75	100

3.3. Correlation between Shared and Unshared Modalities

In Section 2.1.2, we discussed that the prior approximation of a missing modality can improve the final prediction if there exists a high correlation between shared and unshared modalities. For this reason, we apply the partial least squares (PLS) algorithm [41] to first obtain the maximized cross-covariance matrix between shared and unshared modalities (i.e., m_1 and m_2), from which a Pearson correlation coefficient is calculated to indicate how well the two modalities are correlated. Figure 4 visualizes the PLS canonical correlation between the two modalities of the three cancers with reduced dimensionalities. The correlation scores are 0.67 and 0.74, 0.62 and 0.74, 0.89 and 0.80, respectively, for each dimension. The high correlations indicate that it is indeed possible to learn the knowledge of an unshared missing modality in the target dataset through a mapping function learnt from the associations between shared and unshared modalities within a reference dataset. Correlation between multiple modalities has been reported by several other studies [42,43].



(**a**) (**b**) (**c**)

Figure 4. Visualization of the correlation between m_1 and m_2 with dimension reduction. Different markers represent the first and second dimensions. The two modalities are highly correlated as the points lie around the first diagonal. (**a**) GBM. (**b**) LAML. (**c**) PAAD.

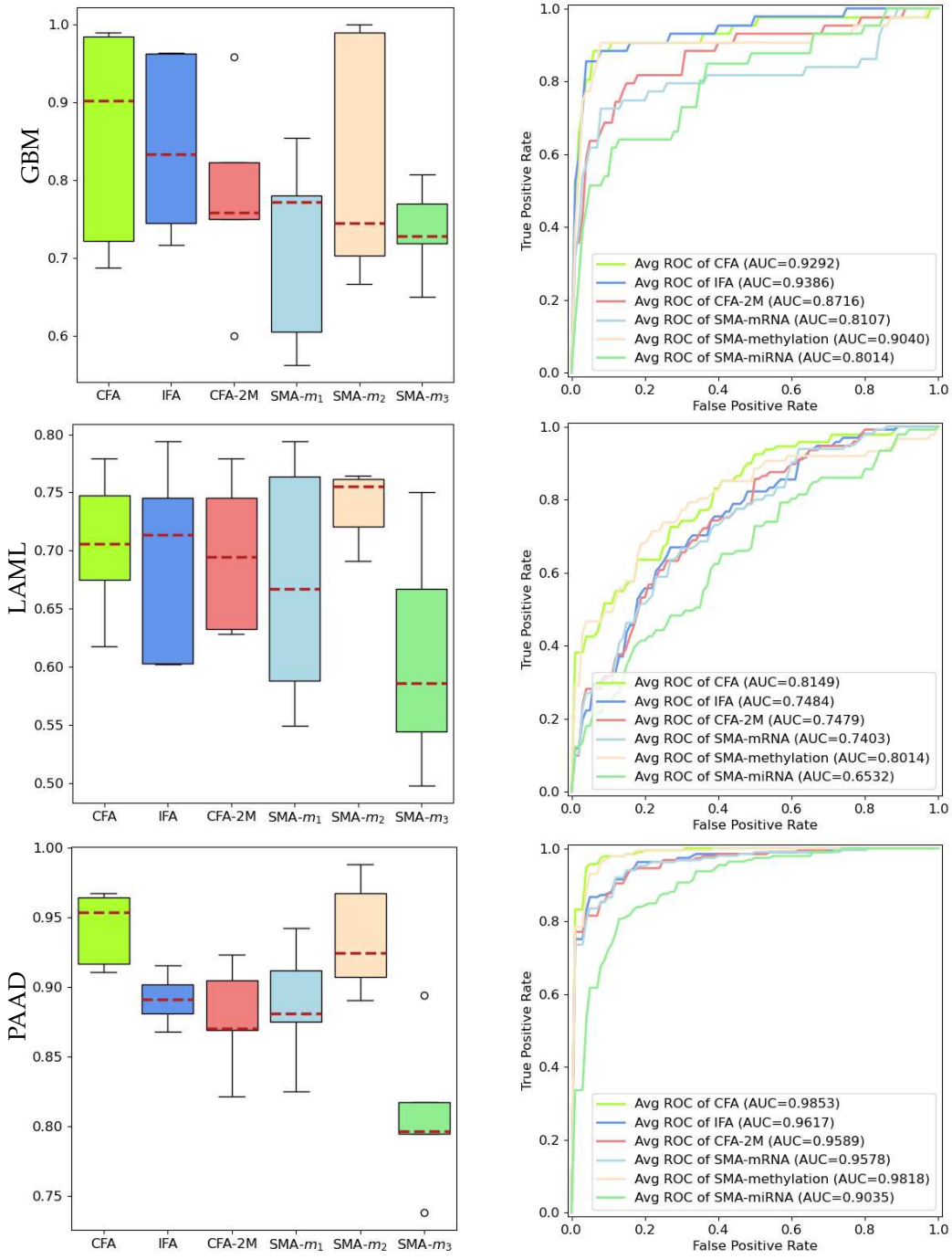
3.4. Evaluation Metrics

We adopt the commonly used evaluation metrics in our experiments including balanced accuracy and ROC AUC.

3.5. Prediction Performances

We conduct a 5-fold cross validation to compare the predictive performances of the several proposed network architectures with other baselines. Regardless of the small sample sizes, all baselines produce desirable performances due to the effective feature compression fulfilled by Equation (1). The testing samples are selected according to different random seeds. First, we report the prediction performance of all actual modalities integrated (m_1, m_2, and m_3) using the complete data (i.e., the ground truth) through our CFA architecture. Next, the m_2 modality is removed, and the IFA architecture is used to learn a fused representation of the actual m_1 and m_3, combined with the predicted m_2. The fused representation is then

used to predict the disease progression and the prediction performances are reported. Next, the prediction performance of the two modalities, i.e., m_1 and m_3 using the CFA architecture is reported, which is denoted as CFA-2M. Moreover, we directly report the performance of every single modality separately using our baseline model SMA. We do not compare our CFA (with all modalities) with previous deep learning based multimodal classification methods due to the lack of spatial information. CEN [20] also supports vector-like data by setting the parameters of height and width to 1. However, we do not compare our results with CEN as complete multimodal classification is not our primary focus.



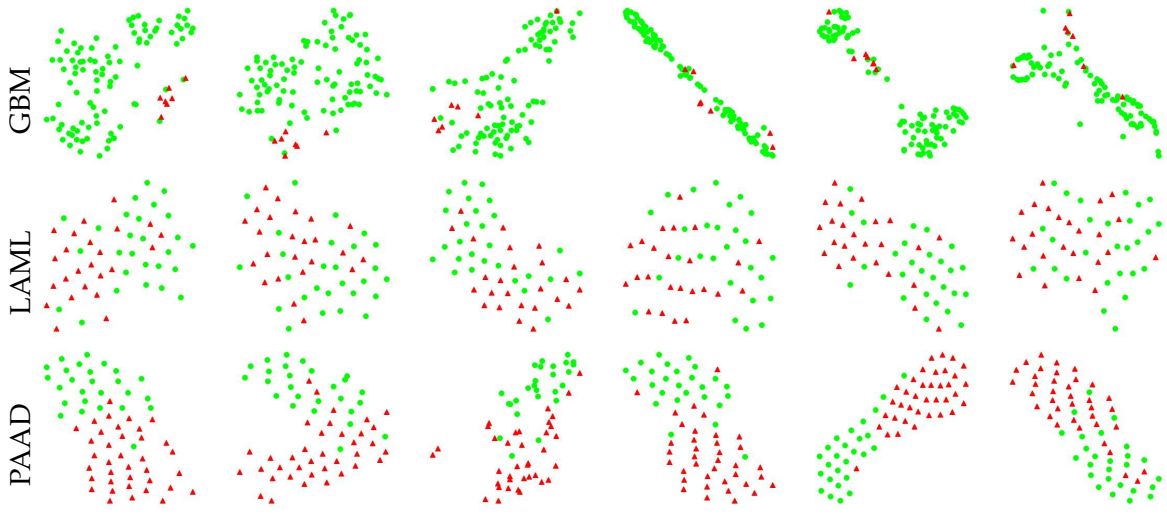
(**a**) Balanced accuracy. (**b**) ROCs.

Figure 5. Comparison of balanced accuracy and ROC performances among the proposed IFA and all baseline predictors.

Figure 5 shows that inferring the highly predictive missing modality through the proposed data fusion model allows the construction of a predictor (IFA) that achieves comparable performances to CFA using all actual modalities, and is better than the baseline models that use only the two available modalities (CFA-2M) or a single modality (SMA-m_1, SMA-m_3). Since we have constructed a fully controlled simulated study, we are able to demonstrate the performance of SMA-m_2 (and, hence, CFA) in Figure 5a, which in practice would be absent. Results have further shown that certain modalities (i.e., DNA methylation, also referred to as SMA-m_2) carry more predictive information than others on specific cancers as can be seen in Figure 5a (for LAML and PAAD). The proposed study, therefore, is able to learn the latent representation of a modality with strong predictive capability that is completely missing in one source through another data source that carries shared entities. Specifically, the area under the ROC curve (AUC) reported in Figure 5b is calculated based on the averaged TPR values (across 5-fold runs) and linearly interpolated FPR values. Both balanced accuracy and AUC performances of our proposed models confirm the IFA's ability to successfully fuse the knowledge learned from the missing modality. A t-test is applied to compare the performances of IFA with CFA-2M, SMA-m_1, SMA-m_3 across 100 runs on GBM and the results further confirmed that the distributions of balanced accuracies are significantly different with p-values of 0.04, <0.0001, and <0.0001, respectively. Similar significance levels are observed when the IFA models are compared with other baselines on LAML and PAAD except for LAML in which no significant difference between IFA and CFA-2M is noted.

3.6. Effective Compression

We show the t-SNE plots in Figure 6 for the compressed (and fused for further multimodal prediction) features in our baseline models. The same testing sets from one of the 5-fold cross-validation analyses are used for fair comparison across models. The plots show that the two patient subgroups in the latent space are well separated for both CFA (Figure 6a) and IFA (Figure 6b) models, while CFA-2M (Figure 6c) performs less optimal separation when compared with the other two models, indicating the proposed IFA's ability to identify the short- and long-term survivors as accurately as CFA. For SMA using single modalities (Figure 6d–f), only the most predictive modality (i.e., DNA methylation) shows relatively well separated results as shown in Figure 6e. Figure 6 also indicates that our proposed models can accurately and robustly compress the multimodal and high-dimensional data with extremely limited sample size through conventional deep learning methods.

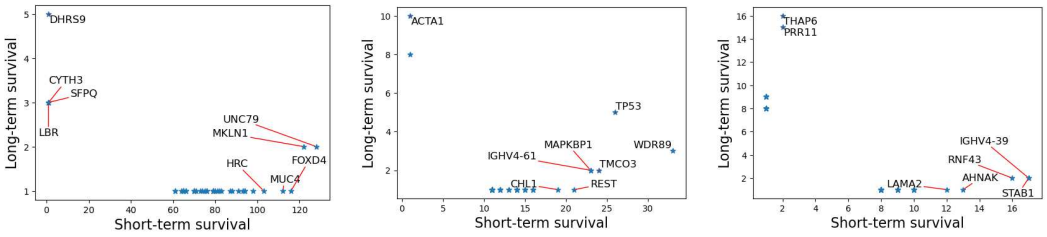


(**a**) CFA. (**b**) IFA. (**c**) CFA-2M. (**d**) SMA-m_1. (**e**) SMA-m_2. (**f**) SMA-m_3.

Figure 6. The t-SNE plots for compressed latent representations of proposed IFA and other baseline methods.

3.7. Functional Analysis

We have further identified mutations that are highly abundant in the short-term survival group, but not in the long-term survival groups for the three cancers. The results are reported in Figure 7. Here, each point represents a gene, and the coordinates are representing the number of patients having at least a variant in that gene for long- vs. short-term survivors. In principle, we are mostly interested in genes that are highly mutated in one group and not in the other which corresponds to the top-left and the bottom-right corners of the graph.



(**a**) (**b**) (**c**)

Figure 7. Number of patients in each group for each mutated gene in the three cancers. The *x*-axes represent the count for short-term survivors, and the *y*-axes represent the count for long-term survivors. Interesting genes appear in the lower right or upper left corners. (**a**) GBM. (**b**) LAML. (**c**) PAAD.

3.8. Ablation Study

An ablation study can be conducted without using a mapping function, i.e., by only using existing modalities. Such results are already demonstrated as CFA-2M in Figure 5.

4. Discussion

Data collection has been the principle bottleneck for advancement in the life sciences, particularly in genomics, engineering, and healthcare, it is not always possible to have access to different modalities collected from different populations. For instance, the disease progression prediction performance of pancreatic adenocarcinoma patients will significantly reduce when DNA methylation data are absent (as can be seen from Figure 5—performance of CFA-2M). The importance of DNA methylation on PAAD prediction is further supported by several studies, including Mishra et al. and Tan et al. [44,45]. The proposed integrative subtyping system, however, will circumvent the many challenges associated with missing modalities and the need to collect additional data by exploiting the current availability of vast genomic, epidemiologic, and clinical data.

Although deep learning models reach impressive prediction accuracies, their nested non-linear structure makes them highly non-transparent, i.e., it is not clear what information from the input data makes them actually arrive at their decisions. For clinicians, these models appear as "black boxes" and, hence, hamper their confidence in using them for clinical decision making, mainly because they are unable to compare to and integrate their expert opinion with the predictions. This, however, can be greatly reduced by explainable AI techniques that aims to understand how the model arrives at the decisions [46–48]. In this study, although we believe that the use of an autoencoder may lead to limitations in explainability, alternative models that exclude the encoding (i.e., predicting the actual values of a missing modality instead of a latent representation) would lead to curse of dimensionality issues given the high dimensional nature of genomic data with limited samples. Additionally, the use of autoencoders results in a lower dimensional latent representation of the data which prevents overfitting issues in subsequent prediction tasks.

In an effort to increase the usage in clinical practice, we examined the detailed mechanisms captured by our proposed classification model, for the three TCGA datasets, in terms of clinical variables, pathways, gene ontology (GO), and functional analysis. The more aggressive GBM subtype appears to affect mostly males with a 60% male dominance. The

median age for the long- and short-term survivors are 35 and 60 years, respectively. GO analysis using iPathwayGuide (Advaita) suggests that the aggressive group has a stronger regulation of glial and astrocyte differentiation (p-values of 0.018 and 0.020) when compared to the less aggressive group. Chinnaiyan et al. [49] reported a similar phenomenon in aggressive glioma. Contrary to long-term survivors, the pathway analysis using iPathwayGuide (Advaita) has shown that the Phagosome pathway is significantly impacted (FDR corrected p-value: 0.037) on the short-term survivors group. Associations between Phagosome and glioblastoma has also been reported by Cammarata et al. [50].

Our PAAD and LAML classification results have shown that the classes are highly influenced by methylation profiles. For LAML, the median age for the long and short term survivors are 50 and 61 years, respectively. However, there is no dominance of age or gender in one group over the other in PAAD.

As seen in Figure 7, the short-term survival group is significantly rich in *MUC4*, *FOXD4*, and *HRC* mutations. *MUC4* is a transmembrane mucin that plays an important role in epithelial renewal and differentiation [51]. Several studies have identified associations between *MUC4* and GBM progression [52–54].

Similarly, we identified rich mutation counts in patients with aggressive LAML in genes including *TP53*, *TMCO3*, and *WDR89* and *REST*. According to Barbosa et al. although the tumor suppressor gene, TP53 has lower mutation frequencies in patients with LAML, such mutations are associated with high risk of relapse and resistance to treatment, which supports our findings [55].

Lastly, our findings has shown that several genes including *RNF43* and *STAB1* are reported to be associated with poor PAAD survival. *RNF43*, an E3 ubiquitin-protein ligase that acts as a negative regulator of the Wnt signaling pathway, is reported to be associated with various cancer types including PAAD [56]. On the contrary, *PRR11* is identified as a variant associated with long-term survival.

5. Conclusions

In this paper, we have presented a deep fusion model that is able to integrate knowledge of multiple studies with partially coupled data through shared entities. The proposed model is able to learn the knowledge of an entirely missing modality within one source through a mapping between the shared and unshared modalities within different sources. The results suggested that all modalities are functioning to the disease prediction, and are dependent on each other. Therefore, studying them together instead of separating them as independent sources of disease predictors, will provide more insights into the aggressiveness of the disease. We conducted several experiments using simulated data from the TCGA glioblastoma multiforme, acute myeloid leukemia, and pancreatic adenocarcinoma datasets. The proposed method enables a more robust and accurate prediction of the three cancers' progression through integration, which is critical for making optimal treatment decisions.

Our models could be extended to other diseases if there exists correlation between the shared and unshared modalities. As a step towards overcoming the domain shift challenge, our approach has the potential to learn the complete knowledge of an unseen data source with missing modalities to improve the classification performance. Generally, our approach could be extended to more than two sources, as long as the additional testing set has the shared modality with the training set. As a future work, the integration of more than two separate data sources can be studied whenever more data are available.

The results of the proposed model 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. The developed framework is expected to be a valuable precision medicine resource for the wider scientific community on other diseases.

Author Contributions: K.Z. and S.A. conceived and designed the project. K.Z. and S.A. performed the experiments. K.Z., S.A., B.S.K., S.A.M., Z.Z. and S.D. analyzed the data and the results. K.Z. and S.A. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation (NSF: #1948338), the Department of Defense (DoD: #W81XWH-21-1-0570), and the National Institutes of Health (NIH: #2P50CA186786-06).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: 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, accessed on 1 September 2021. The source code and data are publicly available at https://github.com/ky-zhou/MMFDA, uploaded on 22 February 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dieterich, M.; Stubert, J.; Reimer, T.; Erickson, N.; Berling, A. Influence of lifestyle factors on breast cancer risk. *Breast Care* **2014**, *9*, 407–414. [CrossRef]
2. Leitzmann, M.F.; Rohrmann, S. Risk factors for the onset of prostatic cancer: Age, location, and behavioral correlates. *Clin. Epidemiol.* **2012**, *4*, 1. [CrossRef] [PubMed]
3. van IJzendoorn, D.G.; Szuhai, K.; Briaire-de Bruijn, I.H.; Kostine, M.; Kuijjer, M.L.; Bovée, J.V. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput. Biol.* **2019**, *15*, e1006826. [CrossRef]
4. López-García, G.; Jerez, J.M.; Franco, L.; Veredas, F.J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS ONE* **2020**, *15*, e0230536. [CrossRef] [PubMed]
5. Zhou, L.; Guo, Z.; Wang, B.; Wu, Y.; Li, Z.; Yao, H.; Fang, R.; Yang, H.; Cao, H.; Cui, Y. Risk Prediction in Patients with Heart Failure with Preserved Ejection Fraction Using Gene Expression Data and Machine Learning. *Front. Genet.* **2021**, *12*, 412. [CrossRef] [PubMed]
6. Lu, J.; Getz, G.; Miska, E.A.; Alvarez-Saavedra, E.; Lamb, J.; Peck, D.; Sweet-Cordero, A.; Ebert, B.L.; Mak, R.H.; Ferrando, A.A.; et al. MicroRNA expression profiles classify human cancers. *Nature* **2005**, *435*, 834–838. [CrossRef]
7. Lauber, C.; Correia, N.; Trumpp, A.; Rieger, M.A.; Dolnik, A.; Bullinger, L.; Roeder, I.; Seifert, M. Survival differences and associated molecular signatures of DNMT3A-mutant acute myeloid leukemia patients. *Sci. Rep.* **2020**, *10*, 12761. [CrossRef]
8. Jonckheere, N.; Auwercx, J.; Hadj Bachir, E.; Coppin, L.; Boukrout, N.; Vincent, A.; Neve, B.; Gautier, M.; Treviño, V.; Van Seuningen, I. Unsupervised hierarchical clustering of pancreatic adenocarcinoma dataset from TCGA defines a mucin expression profile that impacts overall survival. *Cancers* **2020**, *12*, 3309. [CrossRef]
9. Plotnikova, O.; Baranova, A.; Skoblov, M. Comprehensive analysis of human microRNA–mRNA interactome. *Front. Genet.* **2019**, *10*, 933. [CrossRef]
10. Jonas, S.; Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* **2015**, *16*, 421–433. [CrossRef]
11. Aure, M.R.; Fleischer, T.; Bjørklund, S.; Ankill, J.; Castro-Mondragon, J.A.; Børresen-Dale, A.L.; Tost, J.; Sahlberg, K.K.; Mathelier, A.; Tekpli, X.; et al. Crosstalk between microRNA expression and DNA methylation drives the hormone-dependent phenotype of breast cancer. *Genome Med.* **2021**, *13*, 72. [CrossRef] [PubMed]
12. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
13. LeCun, Y.; Ranzato, M. Deep learning tutorial. In *Tutorials in International Conference on Machine Learning (ICML 2013)*; Citeseer: Atlanta, GA, USA, 2013; pp. 1–29.
14. Xu, C.; Tao, D.; Xu, C. A survey on multi-view learning. *arXiv* **2013**, arXiv:1304.5634.
15. Zheng, V.W.; Zheng, Y.; Xie, X.; Yang, Q. Collaborative location and activity recommendations with gps history data. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 1029–1038.
16. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
17. Arslanturk, S.; Draghici, S.; Nguyen, T. Integrated Cancer Subtyping using Heterogeneous Genome-Scale Molecular Datasets. *Pac. Symp. Biocomput.* **2020**, *25*, 551–562.
18. Nguyen, T.; Tagett, R.; Diaz, D.; Draghici, S. A novel approach for data integration and disease subtyping. *Genome Res.* **2017**, *27*, 2025–2039. [CrossRef]
19. Zhou, R.; Shen, Y.D. End-to-end adversarial-attention network for multi-modal clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 14619–14628.
20. Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; Huang, J. Deep multimodal fusion by channel exchanging. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4835–4845.

21. Zheng, Y. Methodologies for cross-domain data fusion: An overview. *IEEE Trans. Big Data* **2015**, *1*, 16–34. [CrossRef]
22. Lahat, D.; Adali, T.; Jutten, C. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE* **2015**, *103*, 1449–1477. [CrossRef]
23. Mariappan, R.; Rajan, V. Deep collective matrix factorization for augmented multi-view learning. *Mach. Learn.* **2019**, *108*, 1395–1420. [CrossRef]
24. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef] [PubMed]
25. Acar, E.; Kolda, T.G.; Dunlavy, D.M. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv* **2011**, arXiv:1105.3422.
26. Beutel, A.; Talukdar, P.P.; Kumar, A.; Faloutsos, C.; Papalexakis, E.E.; Xing, E.P. Flexifact: Scalable flexible factorization of coupled tensors on hadoop. In Proceedings of the 2014 SIAM International Conference on Data Mining. SIAM, Philadelphia, PA, USA, 24–26 April 2014; pp. 109–117.
27. Papalexakis, E.E.; Faloutsos, C.; Sidiropoulos, N.D. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Trans. Intell. Syst. Technol. (TIST)* **2016**, *8*, 1–44. [CrossRef]
28. Ray, P.; Zheng, L.; Lucas, J.; Carin, L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* **2014**, *30*, 1370–1376. [CrossRef] [PubMed]
29. Yang, Y.; Zhan, D.C.; Sheng, X.R.; Jiang, Y. Semi-Supervised Multi-Modal Learning with Incomplete Modalities. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 2998–3004.
30. Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67. [CrossRef]
31. Yu, F.; Xu, C.; Deng, H.W.; Shen, H. A novel computational strategy for DNA methylation imputation using mixture regression model (MRM). *BMC Bioinform.* **2020**, *21*, 552. [CrossRef]
32. Zhou, X.; Chai, H.; Zhao, H.; Luo, C.H.; Yang, Y. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network. *GigaScience* **2020**, *9*, giaa076. [CrossRef]
33. Bischke, B.; Helber, P.; Koenig, F.; Borth, D.; Dengel, A. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In Proceedings of the 2018 IEEE International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, France, 4–6 September 2018; pp. 1–6.
34. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2302–2310.
35. Tran, L.; Liu, X.; Zhou, J.; Jin, R. Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1405–1414.
36. Wang, C.; Niepert, M.; Li, H. LRMM: Learning to recommend with missing modalities. *arXiv* **2018**, arXiv:1808.06791.
37. Azarkhalili, B.; Saberi, A.; Chitsaz, H.; Sharifi-Zarchi, A. DeePathology: Deep multi-task learning for inferring molecular pathology from cancer transcriptome. *Sci. Rep.* **2019**, *9*, 16526. [CrossRef]
38. Zhou, K.; Arslanturk, S.; Craig, D.B.; Heath, E.; Draghici, S. Discovery of primary prostate cancer biomarkers using cross cancer learning. *Sci. Rep.* **2021**, *11*, 10433. [CrossRef]
39. Cadena, C.; Dick, A.R.; Reid, I.D. Multi-modal Auto-Encoders as Joint Estimators for Robotics Scene Understanding. *Robot. Sci. Syst.* **2016**, *5*, 1.
40. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 448–456.
41. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
42. Xu, W.; Xu, M.; Wang, L.; Zhou, W.; Xiang, R.; Shi, Y.; Zhang, Y.; Piao, Y. Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal Transduct. Target. Ther.* **2019**, *4*, 1–11. [CrossRef] [PubMed]
43. Anastasiadi, D.; Esteve-Codina, A.; Piferrer, F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics Chromatin* **2018**, *11*, 37. [CrossRef]
44. Mishra, N.K.; Guda, C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* **2017**, *8*, 28990. [CrossRef]
45. Tan, A.C.; Jimeno, A.; Lin, S.H.; Wheelhouse, J.; Chan, F.; Solomon, A.; Rajeshkumar, N.; Rubio-Viqueira, B.; Hidalgo, M. Characterizing DNA methylation patterns in pancreatic cancer genome. *Mol. Oncol.* **2009**, *3*, 425–438. [CrossRef]
46. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4765–4774.
47. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
48. Liu, K.; Fu, Y.; Wang, P.; Wu, L.; Bo, R.; Li, X. Automating Feature Subspace Exploration via Multi-Agent Reinforcement Learning. In Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 207–215.
49. Chinnaiyan, P.; Kensicki, E.; Bloom, G.; Prabhu, A.; Sarcar, B.; Kahali, S.; Eschrich, S.; Qu, X.; Forsyth, P.; Gillies, R. The metabolomic signature of malignant glioma reflects accelerated anabolic metabolism. *Cancer Res.* **2012**, *72*, 5878–5888. [CrossRef]

50. Cammarata, F.P.; Torrisi, F.; Forte, G.I.; Minafra, L.; Bravatà, V.; Pisciotta, P.; Savoca, G.; Calvaruso, M.; Petringa, G.; Cirrone, G.A.; et al. Proton therapy and src family kinase inhibitor combined treatments on U87 human glioblastoma multiforme cell line. *Int. J. Mol. Sci.* **2019**, *20*, 4745. [CrossRef]
51. Gao, X.P.; Dong, J.J.; Xie, T.; Guan, X. Integrative Analysis of MUC4 to Prognosis and Immune Infiltration in Pan-Cancer: Friend or Foe? *Front. Cell Dev. Biol.* **2021**, *9*, 695544. [CrossRef]
52. Li, W.; Wu, C.; Yao, Y.; Dong, B.; Wei, Z.; Lv, X.; Zhang, J.; Xu, Y. MUC4 modulates human glioblastoma cell proliferation and invasion by upregulating EGFR expression. *Neurosci. Lett.* **2014**, *566*, 82–87. [CrossRef] [PubMed]
53. King, R.J.; Yu, F.; Singh, P.K. Genomic alterations in mucins across cancers. *Oncotarget* **2017**, *8*, 67152. [CrossRef] [PubMed]
54. Seifert, M.; Schackert, G.; Temme, A.; Schröck, E.; Deutsch, A.; Klink, B. Molecular characterization of astrocytoma progression towards secondary glioblastomas utilizing patient-matched tumor pairs. *Cancers* **2020**, *12*, 1696. [CrossRef] [PubMed]
55. Barbosa, K.; Li, S.; Adams, P.D.; Deshpande, A.J. The role of TP53 in acute myeloid leukemia: Challenges and opportunities. *Genes Chromosomes Cancer* **2019**, *58*, 875–888. [CrossRef]
56. Tu, J.; Park, S.; Yu, W.; Zhang, S.; Wu, L.; Carmon, K.; Liu, Q.J. The most common RNF43 mutant G659Vfs* 41 is fully functional in inhibiting Wnt signaling and unlikely to play a role in tumorigenesis. *Sci. Rep.* **2019**, *9*, 18557. [CrossRef]