Bioinformatics, 35, 2019, i446–i454
doi: 10.1093/bioinformatics/btz342
ISMB/ECCB 2019

OXFORD

Deep learning with multimodal representation for pancancer prognosis prediction

Anika Cheerla[1] **and Olivier Gevaert**[2,*]

[1]Monta Vista High School, Cupertino, CA 95014, USA and [2]Department of Medicine and Biomedical Data Science, Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305-5479, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Estimating the future course of patients with cancer lesions is invaluable to physicians; however, current clinical methods fail to effectively use the vast amount of multimodal data that is available for cancer patients. To tackle this problem, we constructed a multimodal neural network-based model to predict the survival of patients for 20 different cancer types using clinical data, mRNA expression data, microRNA expression data and histopathology whole slide images (WSIs). We developed an unsupervised encoder to compress these four data modalities into a single feature vector for each patient, handling missing data through a resilient, multimodal dropout method. Encoding methods were tailored to each data type—using deep highway networks to extract features from clinical and genomic data, and convolutional neural networks to extract features from WSIs.

Results: We used pancancer data to train these feature encodings and predict single cancer and pancancer overall survival, achieving a C-index of 0.78 overall. This work shows that it is possible to build a pancancer model for prognosis that also predicts prognosis in single cancer sites. Furthermore, our model handles multiple data modalities, efficiently analyzes WSIs and represents patient multimodal data flexibly into an unsupervised, informative representation. We thus present a powerful automated tool to accurately determine prognosis, a key step towards personalized treatment for cancer patients.

Availability and implementation: https://github.com/gevaertlab/MultimodalPrognosis

Contact: ogevaert@stanford.edu

1 Introduction

Estimating tumor progression or predicting prognosis can aid physicians significantly in making decisions about care and treatment of cancer patients. To determine the prognosis of these patients, physicians can leverage several types of data including clinical data, genomic profiling, histology slide images and radiographic images, depending on the tissue site. Yet, the high-dimensional nature of some of these data modalities makes it hard for physicians to manually interpret these multimodal biomedical data to determine treatment and estimate prognosis (Gevaert *et al.*, 2006, 2008). Next, the presence of inter-patient heterogeneity warrants that characterizing tumors individually is essential to improving the treatment process (Alizadeh *et al.*, 2015). Previous research has shown how molecular signatures such as gene expression patterns can be mined using machine learning and are predictive of treatment outcomes and prognosis. Similarly, recent work has shown that quantitative analysis of histopathology images using computer vision algorithms can provide additional information on top of what can be discerned by pathologists (Madabhushi and Lee, 2016). Thus, automated machine-learning systems, which can discern patterns among high-dimensional data may be the key to better estimate disease aggressiveness and patient outcomes. Another implication of inter-patient heterogeneity is that tumors of different cancer types may share underlying similarities. Thus, pancancer analysis of large-scale data across a broad range of cancers has the potential to improve disease modeling by exploiting these pancancer similarities. Multi-institutional projects such as The Cancer Genome Atlas (TCGA) (Campbell *et al.*, 2018; Malta *et al.*, 2018; Weinstein *et al.*, 2013), which collected standardized clinical, multiomic and imaging data for a wide array of cancers, are crucial to enable this kind of pancancer modeling.

Automated prognosis prediction, however, remains a difficult task mainly due to the heterogeneity and high dimensionality of the available data. For example each patient in the TCGA database has thousands of genomic features (e.g. microRNA or mRNA) and high resolution histopathology whole slide images (WSIs). Yet, based on

previous work, only a subset of the genomic image features are relevant for predicting prognosis. Thus, to successfully develop a multimodal model for prognosis prediction, an approach is required that can efficiently work with clinical, genomic and image data, in essence multimodal data. Here, we tackle this challenging problem by developing a pancancer deep learning architecture drawing from unsupervised and representation learning techniques, and developing a learning architecture that exploits large-scale genomic and image data to the fullest extent. The main goal of this contribution is to harness the vast amount of TCGA data available to develop a robust representation of tumor characteristics that can be used to cluster and compare patients across a variety of different metrics. Using unsupervised representation techniques, we develop pancancer survival models for cancer patients using multimodal data including clinical, genomic and WSI data.

2 Background

Prognosis prediction can be formulated as a censored survival analysis problem (Cox, 2018; Luck *et al.*, 2017), predicting both if and when an event (i.e. patient death) occurs within a given time period. Given the unique statistical distribution of survival times, they are canonically parameterized using the 'hazard function', such as in standard Cox regression.

In recent years, many different approaches have been attempted to predict cancer prognosis using genomic data. For example Zhang *et al.* (2017) used an augmented Cox regression on TCGA gene expression data to get a C-index of 0.725 in predicting glioblastoma. MicroRNA data in particular have shown high relevance as a measure for disease modeling and prognosis (Calin and Croce, 2006; Cheerla and Gevaert, 2017; Esquela-Kerscher and Slack, 2006; Liu *et al.*, 2017), with Christinat and Krek (2015), achieving a C-index of 0.77 on a subset of renal cancer data using random forest classifiers. However, despite the high performance of machine learning models based on molecular data alone, there is still scope for improvement; after all, the tumor environment is a complex, rapidly evolving milieu that is difficult to characterize through molecular profiling alone (Alizadeh *et al.*, 2015; de Bruin *et al.*, 2013; Lovly *et al.*, 2016).

Recently, the use of WSI data has been shown to improve the performance and generality of prognosis prediction. As WSIs are high resolution images of cellular architecture and environment with potentially only a fraction of the slide relevant to predicting prognosis, much of the literature focuses on hybrid approaches involving pathologist annotation of regions of interest (ROIs). For example Wang *et al.* (2014) match the performance of genomic models by using 500 × 500 pixel, physician-selected ROIs and handcrafted slide features to predict prognosis. More recently, deep learning provides a significant boost in predictive power. For example Yao *et al.* (2016) are able to significantly outperform all molecular profiling-based methods on two lung cancer datasets using only physician-selected ROIs and convolutional neural networks (CNNs). Other reports, including Beck *et al.* (2011) and Bejnordi *et al.* (2017), showing that histopathology image data contains important prognostic information that is complementary to molecular data. Yet, multimodal prognosis models are still highly underexplored (Momeni *et al.*, 2018a). To our knowledge, only one paper explores combining genomic and image data for prognosis showing that a lung-cancer genomic model (C-index 0.660) and WSI-based model with hand-annotated ROIs (C-index 0.613) can be combined to get a final classifier with C-index 0.691 (Zhu *et al.*, 2016).

Moreover, the WSI-based methods discussed above require a pathologist to hand-annotate ROIs, a tedious task. Arguably the most difficult part of automated, multimodal prognosis prediction is finding clinically relevant ROIs automatically. In the related field of tumor classification from WSIs, a 'decision-fusion' model that randomly samples patches and integrates them into a Gaussian mixture has yielded accurate predictions (Hou *et al.*, 2016). Moreover, more recent work has focused on using attention mechanisms to learn what patches are important (Momeni *et al.*, 2018b). However, in prognosis prediction, truly automated WSI-based systems have had limited success. One report uses a slide-based approach that relies on unsupervised learning—Zhu *et al.*'s (2017) recent paper uses K-means clustering to characterize and adaptively sample patches within slide images, achieving 0.708C-index on lung cancer data, a result that nearly rivals genomic-data approaches.

Previous research has focused mostly on single-cancer datasets, missing the opportunity to explore commonalities and relationships between tumors in different tissues. And although previous papers explore both genomic and imaging-based approaches, few models have been developed that integrate both data modalities. By exploiting multimodal data, as well as developing better methods to automate WSI scoring and extract useful information from slides, we have the potential to improve upon the state-of-the-art.

In recent years, CNNs have been used to significantly improve machine learning tasks (LeCun *et al.*, 2015) including missing value estimation in genomic data (Qiu *et al.*, 2018) and prediction of prognostic factors based on WSI (Momeni *et al.*, 2018b). A key component of the success of CNNs is their ability to deal with high-dimensional, unstructured data, in particular image data (Wang *et al.*, 2017). For example CNNs can accurately classify scenes from images by learning a set of flexible, hierarchical features (Zhou *et al.*, 2014). Even if the majority of pixel inputs are 'dropped out' completely for some samples, this model can still be trained to predict accurately and can handle the uncertainty (Wager *et al.*, 2013).

The prognosis prediction task is more unstructured than traditional deep learning tasks; instead of classifying from relatively small images (224 × 224 for ImageNet, e.g.), we must predict survival times from a combination of clinical, genomic and WSI images that are much higher resolution. Furthermore, patients span a wide variety of cancer types, and are often missing some form of imaging, clinical or genomic data, making it difficult to apply standard CNNs. Unsupervised learning has shown significant promise (Fan *et al.*, 2018). By learning unsupervised correlations among imaging features and genomic features, it may be possible to overcome the paucity of data labels. Similarly, representation learning techniques might allow us to exploit similarities and relationships between data modalities (Kaiser *et al.*, 2017). In prognosis prediction, it is crucial that the model maps similar patients to the same abstract representation in a way that is agnostic to data modality and availability. We propose to use unsupervised and representation learning to tackle many of the challenges that make prognosis prediction using multimodal data difficult.

3 Materials and methods

3.1 Datasets and tools

Our main source of data is preprocessed and batch corrected data from the PanCanAtlas TCGA project (Campbell *et al.*, 2018; Malta *et al.*, 2018; Weinstein *et al.*, 2013). This dataset contains data for

Table 1. Data distribution of TCGA data including missing data

Data type	Number of cases	Number of missing cases	Percentage missing (%)
Gene expression data	10 198	962	8.62
MicroRNA expression data	10 125	1035	9.27
WSI slide data	10 914	246	2.2
Clinical data	7512	3648	32.69
Survival target data (time of death)	11 121	39	0.35
Patients with complete data	6404	4756	42.62

Note: Survival data are available for the majority of patients, while microRNA and clinical data are missing in a subset of patients. Nearly 43% of patients have at least one type of missing data.

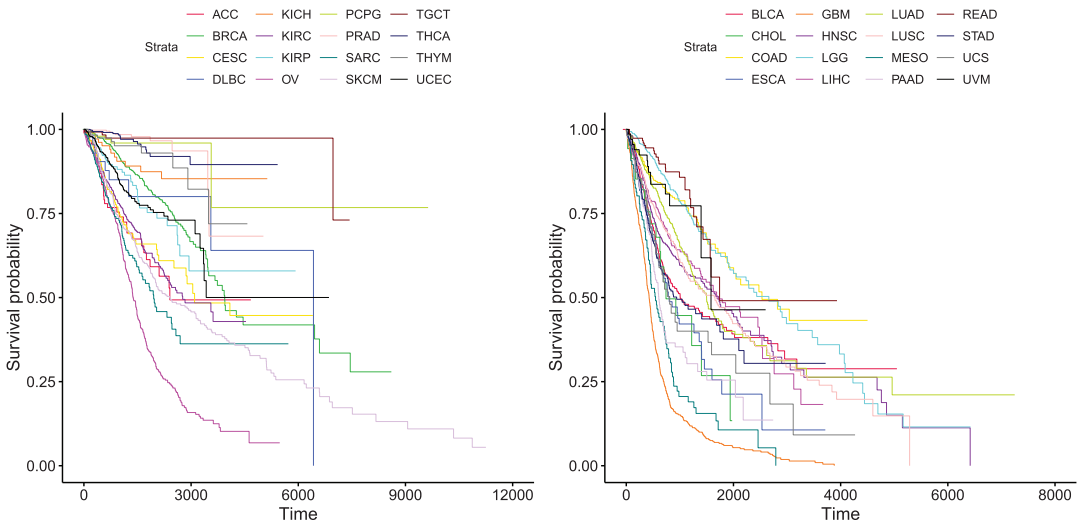

Fig. 1. Kaplan–Meier survival curves for all cancer sites in TCGA demonstrating that overall survival is tissue specific. The first graph contains the 10 cancers with the highest mean overall survival, the second graph contains the 10 cancers with the lowest mean overall survival

1881 microRNAs, gene expression data for 60 383 genes, a wide range of clinical data, of which we used the race, age, gender and histological grade variables, and WSI data for over 11 000 patients. Table 1 describes the data distribution in more detail. Many patients do not have all data available, implying that classifiers and architectures that can deal with missing data are warranted. Each patient has a time of death recorded, right-censored up to a maximum of 11 000 days after diagnosis across all cancer sites. The 20 cancers we examine have significantly different survival patterns, as can be seen in Figure 1. We rely on the Python package openslide to efficiently read and parse WSIs and the PyTorch framework to enable the creation of neural network models. To train our models, we use an NVIDIA^{TM} GTX 1070 GPU.

The TCGA dataset of 11 160 patients was split into training and testing datasets in 85/15 ratio, stratifying by cancer type in order to ensure the same distribution of cancers in both the training and test sets.

3.2 Deep unsupervised representation learning

In order to train a pancancer model for prognosis prediction, we first attempt to compress multiple data modalities into a single feature vector that represents a patient. Previous work has found significant cross-correlations between different data types (e.g. gene expression, clinical, microRNA and image data) (Gevaert *et al.*, 2012; Momeni *et al.*, 2018a), and learning these relations in an unsupervised fashion could significantly improve the prognosis prediction process. Thus, we use a representation learning framework to guide our approach. Although approaches such as split-brain autoencoders induce convergence between different multimodal feature representations, they rely on reconstruction error, which may not be a good choice for heterogeneous data sources. Instead, we rely on a method inspired by Chopra *et al.* (2005), in which two different views of objects are passed through a Siamese network to create feature representations. For views from the same object, the cosine similarity between these feature representations is maximized, whereas for views from different objects, the cosine similarity is minimized. To ensure stability, a margin-based, hinge-loss formulation is used, such that different-object feature representations are only penalized if they fall within a margin M of the same-object representations. This forces different views of a single patient's information to have similar feature vectors, while avoiding mode collapse where all features predict exactly the same vector for all patients.

In this work, we use a similar formulation as (Chopra *et al.*, 2005), but with some modifications. Because of the different data modalities, instead of using a Siamese network, we use one deep neural network for each data type, with differing architectures described in Figure 2. We define the feature space to have a length of 512 based on empirical evidence (data not shown). Since we have

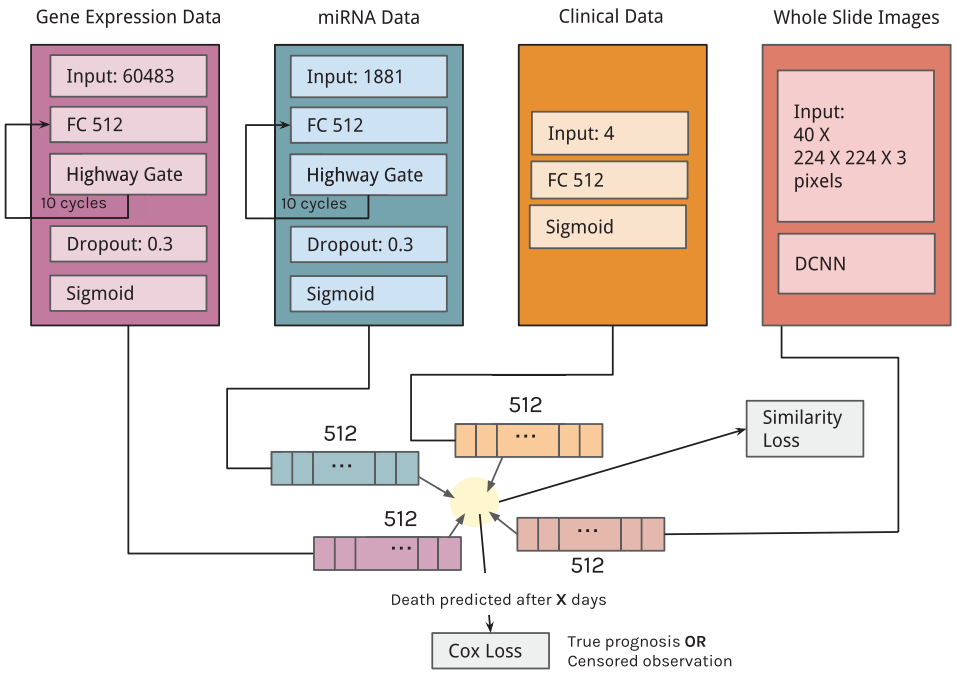

Fig. 2. Structure of the unsupervised model: the similarity loss can be visualized as projecting representations of different modalities in the same space. Each modality uses a different network architecture. For the clinical data, we use FC layers with sigmoid activations, for the genomic data we use deep highway networks (Srivastava *et al.*, 2015) and for the WSI images, we use the SqueezeNet architecture (Iandola *et al.*, 2016) (see main text for architecture details). These architectures generate feature vectors that are then aggregated into a single representation and used to predict overall survival

more than two different modalities, we sum over the similarity loss for each pair of modalities that are present. We can define the loss $l_{sim}(\theta)$ as in Equations (1)–(3):

$$sim_\theta(x,y) = \sum_{i,j \in \text{modalities}} \frac{\hat{h}_{\theta,i}(x_i) \cdot \hat{h}_{\theta,j}(y_j)}{|\hat{h}_{\theta,i}(x_i)||\hat{h}_{\theta,j}(y_j)|} \tag{1}$$

$$L_\theta(x,y) = \max(0, M - sim_\theta(x,y) + sim_\theta(x,x)) \tag{2}$$

$$l_{sim}(\theta) = \sum_{x,y} L_\theta(x,y) \tag{3}$$

where x_i is the data for modality i and $\hat{h}_{\theta,i}$ is the predictive model for modality i. Note that the parameter M controls the 'tightness' of the clustering. If M is high, feature vectors for a given patient are permitted to be relatively different, as long as they stay similar to a certain extent. If M is low, feature vectors for a patient are forced to be much closer together, which is usually more ideal, but can also cause mode collapse. We settled on $M = 0.1$ as the default value based on our observations that it is the smallest value of M that does not cause mode collapse. This loss is computed between every pair of patients in a batch. Thus, the unsupervised model must learn to recognize important, patient-distinguishing patterns in genomic and image data. Moreover, it must learn how patterns in one modality correspond to patterns in a different modality, so it can generate similar encodings for both. As a result, this method naturally generates compact patient representations that are resilient to missing data. The entire process is summarized in Figure 2.

3.3 Prognosis prediction

In addition to learning the feature representation, the model must also accurately predict prognosis. Because this is a survival data problem, we aim to maximize the concordance score or C-index. Previous research has defined the Cox loss function (Katzman *et al.*, 2016), which optimizes the Cox partial likelihood, as the best way to maximize concordance differentiably. Thus, we add a final prediction layer that maps the 512 feature vector to a survival prediction. We use the standard formulation of Cox loss to train the model. Cox loss is defined as

$$l_{cox}(\theta) := -\sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j:T_j > T_i} e^{\hat{h}_\theta(x_j)} \right) \tag{4}$$

where the values T_i, E_i and x_i are, respectively, the survival time, the censorship flag and the data for each patient, and $\hat{h}_\theta$ represents the neural network model trained to predict survival times. The loss is computed over all patients whose lack of survival was observed. Combining with the unsupervised model, the overall loss becomes

$$l(\theta) = l_{sim}(\theta) + l_{cox}(\theta) \tag{5}$$

3.4 Model architectures

We use a dedicated CNN architecture for each data type. For the clinical data, we use fully connected (FC) layers (Fig. 2) with sigmoid activations and dropout as encoders. For the gene and microRNA data, we use highway networks as the architecture (Srivastava *et al.*, 2015). Because of the complexity and scale of WSI images, we use the CNN architecture to encode the image data. These architectures are now described in more detail.

The genomic and microRNA patient data sources are represented by dense, large one-dimensional vectors and neural networks are not the traditional choice for such problems, e.g. support vector machines or random forests are more commonly used (Daemen *et al.*, 2008, 2009). However, in order to differentiably optimize the similarity and Cox loss, we must use CNNs to predict these features. Recent improvements to the state-of-the-art have made deep

learning approaches competitive with other approaches. Thus, we use deep highway networks to train 10-layer deep feature predictors without compromising gradient flow through a neural gating approach (Srivastava *et al.*, 2015). Highway networks use LSTM-style sigmoidal gating to control gradient flow between deep layers, combating the problem of 'vanishing' and 'exploding' gradient in very deep feed forward neural networks (Fig. 2).

In order to represent and encode WSIs, we need to develop machine learning methods that can effectively 'summarize' WSIs. However, the high resolution of WSIs makes learning from them in their entirety difficult. Thus, there must be an element of stochastic sampling and filtering involved. In this work, we use a relatively simple approach to sample ROIs. We sample 200 224 × 224 pixel patches at the highest resolution, then compute the 'color balance' of each patch; i.e. how far the average (R, G, B) color value deviates from the mean (R, G, B) value of the entire WSI using mean-squared error. Then, we select the top 20% of these 200 patches (or 40 patches) as ROIs; this ensures that 'non-representative' patches belonging to white-space and over-staining are ignored. These 40 ROIs represent, on average, 15% of the tissue region within the WSI. Next, we apply a SqueezeNet model (Iandola *et al.*, 2016) on these 40 ROIs, with the last layer being replaced by the length-512 feature encoding predictor. The architecture is detailed in Figure 3. This model is connected to the broader network as shown in Figure 2, and is trained using the similarity and Cox loss terms.

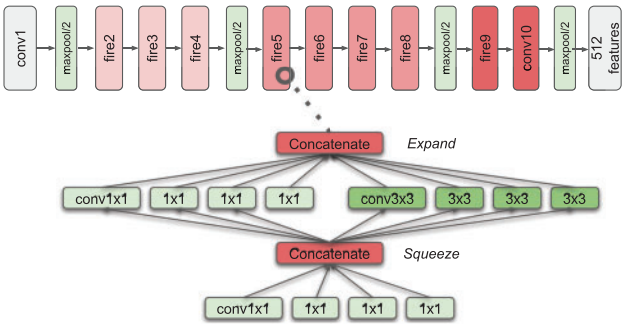

Fig. 3. The SqueezeNet model architecture. The SqueezeNet architecture consists of a set of fire modules interspersed with maxpool layers. Each fire module consists of a squeeze layer (with 1 × 1 convolution filters) and expand layer (with a mix of 1 × 1 and 3 × 3 convolution filters). This fire module architecture helps to reduce the parameter space for faster training. We replaced the final softmax layer of the original SqueezeNet model with the 512-length feature encoding predictor

Because the SqueezeNet model is designed to be computationally efficient, we can train on a large percentage of the WSI patches without sacrificing performance. We tuned the hyper parameters of these model architectures on a validation set to find the final model parameters (Figs 2 and 3). To evaluate the performance of our model, we use the concordance score (C-index) on the test dataset.

3.5 Multimodal dropout

Dropout is a commonly used regularization technique in deep neural network architectures in which some randomly selected neurons are dropped out during the training, forcing other neurons to step in to make predictions for missing neurons. This technique results in less overfitting and more generalization (Srivastava *et al.*, 2014). We developed a variation of dropout, multimodal dropout, to improve the network's ability to deal with missing data. In this method, instead of dropping neurons, we drop entire feature vectors corresponding to each modality, and scale up the weights of the other modalities correspondingly similar to our previous work (Momeni *et al.*, 2018a). This is applied to each data sample during training with probability P for each modality, to force the network to create representations that are robust to missing data modalities. We experimented with a number of different values for P before settling on 25% as optimal.

3.6 Visualization

T-distributed stochastic neighbor embedding, or T-SNE, is a commonly used visualization technique that maps points in high-dimensional vector spaces into lower-dimensions (Maaten and Hinton, 2008). Unlike other dimensionality reduction techniques like Principal Component Analysis (PCA), T-SNE produces more visually interpretable results by converting vector similarities into joint probabilities, generating visually distinct clusters that represent patterns in the data. Here, we use T-SNE to cluster and show the relationships between our length-512 feature vectors representing patients. Because T-SNE is computationally intensive, we first used PCA to project these vectors into a 50-dimensional space, then apply T-SNE to map them into 2D space.

4 Results and discussion

4.1 Unsupervised learning representations

We first evaluated the unsupervised representation learning of our model architecture by visualizing the encodings of the pancancer patient cohort (Fig. 4). Clusters of patients with similar feature representations tend to have the same traits (race, sex and cancer type),

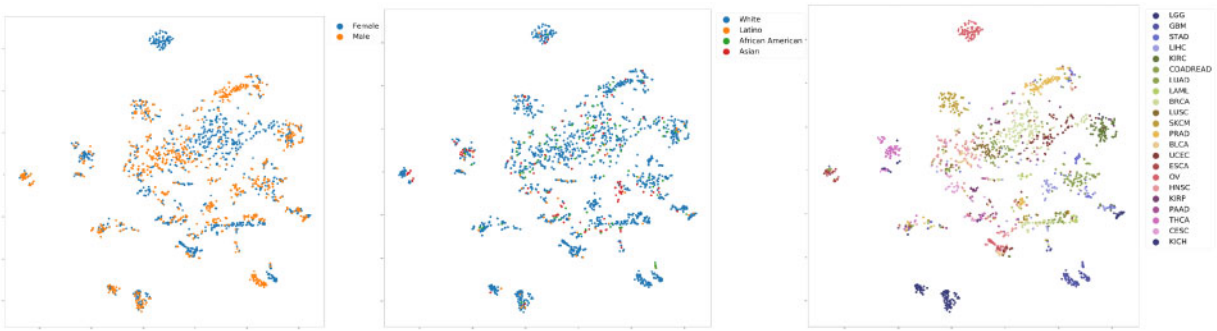

Fig. 4. T-SNE-mapped representations of feature vectors T-SNE-mapped representations of feature vectors for 500 patients within the testing set. The 512-length feature vectors were compressed using PCA (50 features) and T-SNE into the 2D space. These representations manage to capture relationships between patients; e.g. patients with the same sex were generally clustered together (left image), and to a lesser extent, patients of the same race and same cancer type tended to be clustered as well (center and right), even when those clinical features were not provided to the model

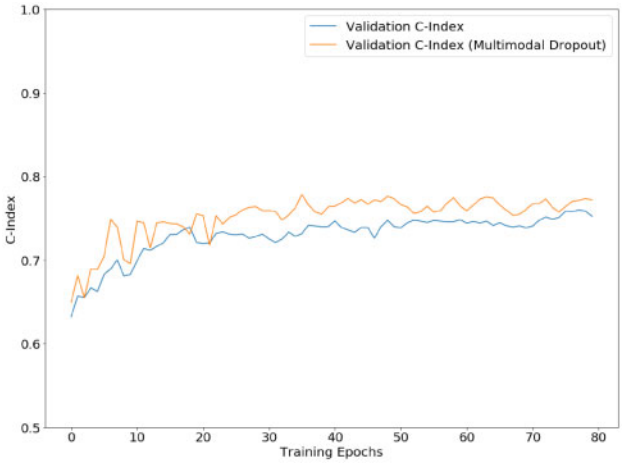

Fig. 5. Evaluation of multimodal dropout: learning rate in terms of C-index of the model on the validation dataset for predicting prognosis across 20 cancer sites combining multimodal data. The model converges after 40 epochs and shows that multimodal dropout improves the validation performance

even though the model was not explicitly trained on these variables. The CNN model thus learned, in an unsupervised fashion, relationships between factors such as sex, race and cancer type across different modalities. These results suggest that the unsupervised model can effectively summarize information from multimodal data and our proposed unsupervised encoding could act as a pancancer 'patient profile'.

4.2 Evaluation of multimodal dropout

Next, we evaluated the use of the multimodal dropout when integrating multimodal clinical, gene expression, microRNA and WSIs across 20 cancer sites to predict the survival of patients. We train the models for 80 epochs and we see model convergence within that span (Fig. 5). This analysis also showed that the validation C-index improves when using multimodal dropout during training (Fig. 5), indicating that randomly dropping-out feature vectors during training improves the network's ability to build accurate representations from missing multimodal data.

4.3 Pancancer prognosis prediction

Next, we used our model on the test dataset to predict prognosis in single cancer and pancancer experiments. We compared different combinations of modalities, always including clinical data, and we evaluated the use of multimodal dropout. We observed that only for the integration of clinical and mRNA, multimodal dropout did not improve the results. For the model that is trained with all modalities, many of the cancer types (15 out of 20) have a higher C-index compared to the training without multimodal dropout with an average an improvement of 2.8%. Similar results are observed for integrating less data modalities (Table 2). In addition, the pancancer model integrating clinical, mRNA, miRNA and WSI achieves an overall C-index of 0.78 on all cancers with multimodal dropout versus 0.75 without dropout. Also for the other pancancer models integrating two or three data modalities, an improvement in multimodal dropout was observed except for the integration of clinical and mRNA data (Table 2).

4.4 Essential data modalities

Next, we investigated using different combinations of modalities together with clinical data, to examine if the genomic and image modalities are crucial for prognosis prediction. We observed that miRNA is the most informative modality while mRNA is the least informative in a pancancer setting when integrating all modalities (C-index of 0.75 versus 0.60 for the baseline pancancer model, Table 2). For single cancers, different combinations of modalities are important. For eight cancer sites, the integration of all four modalities is the best with the most striking example KICH (C-index 0.95). Next, for six cancer sites, integration of clinical, miRNA and WSI gives the best or equal performance to the model integrating all four modalities, suggesting that mRNA is also not essential in these single cancer models for prognosis prediction (Table 2). For example, the best model for KIRP, OV and LUAD results from integrating clinical, miRNA and WSI with C-index of 0.86, 0.69 and 0.77, respectively, suggesting that these three data modalities are sufficient and necessary for these cancer sites prognosis determination.

4.5 Pancancer pretraining evaluation

Next, we tested if training on pancancer data actually improved the prediction of survival across each individual cancer site. To test this, we compared the multimodal pancancer results with the results of models trained on each cancer site using an 85–15 train–test split, separately for the multimodal dropout model using all data modalities (i.e. clin + miRNA + mRNA + WSI), and compared the performance for survival prediction using exactly the same test cases for each cancer site. This showed that for all cancer sites pancancer training improves the results except for KIRC where a drop of 6% was observed (Table 3).

4.6 Comparison with previous work

All previous work on prognosis prediction using genomic and WSI data has focused on specific cancer types and data modalities. For example, Christinat and Krek (2015) achieved the highest C-index (0.77) thus far, on renal cancer data (TCGA-KIRC). As can be seen from our results, our method performed slightly worse (0.740) on the same type of data. However, our method outperforms a multimodality classifier on lung adenocarcinoma by Zhu *et al.* (2016) (0.726 versus 0.691C-index). In general there is no 'fair comparison' that can be made between this method and the previous state-of-the-art, especially because most previous papers discard patients with missing data modalities, while our proposed model is able to train and predict with missing data included. Moreover, our methods achieve comparable or better results from previous research by resiliently handling incomplete data and predicting across 20 different cancer types.

5 Conclusion

In this paper, we demonstrate a multimodal approach for predicting prognosis using clinical, genomic and WSI data. First, we developed an unsupervised method to encode multimodal patient data into a common feature representation that is independent of data type or modality. We then illustrated that these unsupervised patient encodings are associated with clinical features, and that patients with similar characteristics tend to cluster together in 'representation-space'. These feature representations act as an integrated multimodal patient profile, enabling machine learning models to compare and contrast patients in a systematic fashion. Thus, these encodings could be useful in a number of contexts, ranging from prognosis prediction to treatment recommendation.

Table 2. Model performance using C-index on the 20 studied cancer types, using different combinations of data modalities

	Clin+miRNA+mRNA+WSI			Clin+miRNA			Clin+mRNA			Clin+miRNA+mRNA			Clin+miRNA+WSI		
Cancer site	Baseline	Multimodal dropout	Delta (%)	Baseline	Multimodal dropout	Delta (%)	Baseline	Multimodal dropout	Delta (%)	Baseline	Multimodal dropout	Delta (%)	Baseline	Multimodal dropout	Delta (%)
BLCA	0.65	0.73	12.6	0.66	0.69	4.4	0.60	0.58	−4.4	0.65	0.62	−5.1	0.65	0.68	4.3
BRCA	0.77	0.79	3.0	0.80	0.80	−0.1	0.57	0.56	−1.9	0.73	0.73	0.3	0.77	0.77	0.0
CESC	0.73	0.76	4.6	0.77	0.76	−1.2	0.67	0.62	−6.9	0.74	0.74	0.4	0.78	0.76	−2.5
COADREAD	0.72	0.74	3.8	0.78	0.75	−4.8	0.72	0.58	−20.0	0.77	0.64	−16.9	0.70	0.74	4.5
HNSC	0.61	0.67	10.4	0.64	0.64	0.7	0.58	0.55	−5.4	0.63	0.66	4.6	0.61	0.65	6.6
KICH	0.95	0.93	−2.0	0.82	0.85	3.0	0.80	0.84	5.5	0.73	0.77	5.9	0.81	0.88	9.7
KIRC	0.73	0.73	−0.3	0.70	0.72	3.1	0.61	0.65	5.9	0.65	0.66	2.7	0.68	0.61	−11.1
KIRP	0.84	0.79	−6.0	0.76	0.79	4.1	0.65	0.64	−1.0	0.61	0.70	14.5	0.79	0.86	9.2
LAML	0.66	0.67	1.8	0.69	0.79	14.9	0.57	0.61	7.4	0.66	0.57	−12.8	0.61	0.59	−2.8
LGG	0.83	0.85	3.4	0.79	0.81	2.0	0.63	0.67	6.3	0.77	0.78	1.4	0.76	0.82	8.2
LIHC	0.72	0.77	7.6	0.73	0.74	2.7	0.64	0.69	7.7	0.68	0.67	−1.8	0.70	0.77	11.2
LUAD	0.72	0.73	1.3	0.72	0.72	−0.9	0.63	0.58	−8.9	0.73	0.69	−5.1	0.69	0.77	10.5
LUSC	0.67	0.66	−0.9	0.72	0.67	−6.5	0.50	0.51	2.1	0.62	0.60	−2.9	0.67	0.68	0.5
OV	0.63	0.67	6.4	0.65	0.63	−2.2	0.47	0.52	11.5	0.59	0.61	3.5	0.62	0.69	10.4
PAAD	0.71	0.74	3.5	0.68	0.71	3.8	0.57	0.61	7.6	0.59	0.64	8.9	0.69	0.69	0.3
PRAD	0.77	0.81	0.0	0.64	0.64	−0.3	0.60	0.58	−3.5	0.59	0.78	32.8	0.53	0.60	13.4
SKCM	0.68	0.72	5.2	0.68	0.68	−0.1	0.56	0.55	−0.1	0.58	0.72	24.3	0.67	0.72	6.8
STAD	0.76	0.78	2.6	0.75	0.76	1.5	0.63	0.54	−13.9	0.80	0.69	−14.1	0.72	0.74	2.6
THCA	0.95	0.90	−4.8	0.97	0.95	−2.6	0.82	0.54	−34.2	0.70	0.83	18.7	0.93	0.94	1.4
UCEC	0.85	0.85	0.6	0.81	0.85	4.3	0.63	0.63	0.0	0.66	0.78	18.2	0.77	0.80	3.0
Average improvement	2.8%			1.3%			−2.3%			3.9%			4.3%		
Pancancer	0.75	0.78	4.5	0.74	0.78	4.3	0.60	0.60	−1.2	0.75	0.78	3.6	0.76	0.78	3.2

Note: Cancer sites are defined according to TCGA cancer codes. For each cancer, the best result is bold faced. Delta refers to the relative performance improvement of the multimodal dropout model compared to the baseline.

Clin, clinical data; miRNA, microRNA expression data; mRNA, mRNA expression data; WSI, whole slide images.

Table 3. Comparison of pancancer training with single cancer training using the C-index showing that in the case of integrating clinical, miRNA, mRNA and WSI using multimodal dropout, for all but one cancer site (KIRC), pancancer training performs equally or outperforms training on each cancer individually

Cancer site	Single cancer	Pancancer	Difference (%)
BLCA	0.60	0.73	22
BRCA	0.62	0.79	28
CESC	0.52	0.76	48
COADREAD	0.58	0.74	28
HNSC	0.64	0.67	6
KICH	0.69	0.93	34
KIRC	0.78	0.73	−6
KIRP	0.51	0.79	56
LAML	0.65	0.67	4
LGG	0.73	0.85	18
LIHC	0.78	0.77	0
LUAD	0.72	0.73	1
LUSC	0.63	0.66	5
OV	0.54	0.67	24
PAAD	0.57	0.74	30
PRAD	0.76	0.81	7
SKCM	0.54	0.72	33
STAD	0.60	0.78	29
THCA	0.53	0.90	69
UCEC	0.67	0.85	28

We then used these feature representations to predict single cancer and pancancer prognosis. On 20 TCGA cancer sites, our methods achieve the overall C-index of 0.784. Furthermore, on cancer types that have few samples (e.g. KICH), our prognostic prediction model is able to estimate prognosis with relatively high accuracy, leveraging unsupervised features and information from other cancer types to overcome data scarcity.

Our work distinguishes itself in a number of ways, we demonstrate how to build a pancancer model of prognosis. Next, we show the use of multimodal data, novel representation learning techniques and methods such as multimodal dropout to create models that can generalize well and predict also in the absence of one or more data modalities. More specifically, while learning unsupervised relationships between clinical, genomic and image data, our proposed CNN is forced to develop a unique, consistent representation for each patient. Finally, we propose an efficient automated WSI analysis by sampling ROIs per patient representing on average 15% of patient's lesions.

6 Future work

Although we have created an algorithm to select patches from WSI images, our work for modeling WSI can be further improved. Refining the CNN architecture used for encoding the biopsy slides is crucial to further improve the performance. Future research, likely should focus on learning which image patches are important, rather than randomly sampling patches. Furthermore, we can use more advanced, deeper architectures and advanced data augmentation. Another intriguing possibility is using transfer learning on models designed to detect low-level cellular activity like mitoses (Zagoruyko and Komodakis, 2016). Because of the well-established connection between mitotic proliferation and cancer, this could help focus the CNN on important cellular features. Next, integrating more diverse sources of data is another key goal. In this research, resource constraints prevented us from exploring other data genomic modalities in TCGA, such as DNA methylation (Gevaert, 2015; Litovkin *et al.*, 2014) and DNA copy number data (Gevaert *et al.*, 2013; Gevaert and Plevritis, 2013), all of which have potentially untapped, prognostically relevant information.

Funding

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award R01EB020527, the National Institute of Dental and Craniofacial Research (NIDCR) under award U01DE025188, and the National Cancer Institute (NCI) under awards U01CA199241 and U01CA217851. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Alizadeh,A.A. *et al.* (2015) Toward understanding and exploiting tumor heterogeneity. *Nat. Med.*, **21**, 846–853.

Beck,A.H. *et al.* (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.*, **3**, 108ra113.

Bejnordi,B.E. *et al.* (2017) Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In: *IEEE 14th International Symposium on Biomedical Imaging 2017 (ISBI 2017)*, pp. 929–932. IEEE, Melbourne, Australia.

Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857.

Campbell,J.D. *et al.* (2018) Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell Rep.*, **23**, 194.

Cheerla,N. and Gevaert,O. (2017) Microrna based pan-cancer diagnosis and treatment recommendation. *BMC Bioinform.*, **18**, 32.

Chopra,S. *et al.* (2005) Learning a similarity metric discriminatively, with application to face verification. In: *Proceedings 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, Vol. **I**, IEEE Computer Society, Los Alamitas, CA, pp. 539–546.

Christinat,Y. and Krek,W. (2015) Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. *Oncotarget*, **6**, 10521.

Cox,D.R. (2018) *Analysis of Survival Data*. Routledge, New York.

Daemen,A. *et al.* (2008) Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. In: *Pacific Symposium on Biocomputing 2008*, pp. 166–177. World Scientific, Singapore.

Daemen,A. *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support. *Genome Med.*, **1**, 39.

de Bruin,E.C. *et al.* (2013) Intra-tumor heterogeneity: lessons from microbial evolution and clinical implications. *Genome Med.*, **5**, 101.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Fan,H. *et al.* (2018) Unsupervised person re-identification: clustering and fine-tuning. *ACM Trans. Multimedia Comput. Commun. Appl.*, **14**, 83.

Gevaert,O. (2015) Methylmix: an r package for identifying DNA methylation-driven genes. *Bioinformatics*, **31**, 1839–1841.

Gevaert,O. and Plevritis,S. (2013) Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features. In: *Pacific Symposium on Biocomputing 2013*, pp. 123–134. World Scientific, Singapore.

Gevaert,O. *et al.* (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, **22**, e184–e190.

Gevaert,O. *et al.* (2008) Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients. In: *Pacific Symposium on Biocomputing 2008*, pp. 279–290. World Scientific, Singapore.

Gevaert,O. *et al*. (2012) Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results. *Radiology*, **264**, 387–396.

Gevaert,O. *et al*. (2013) Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface Focus*, **3**, 20130013.

Hou,L. *et al*. (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Los Alamitas, CA, pp. 2424–2433.

Iandola,F.N. *et al*. (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. arXiv:1602.07360.

Kaiser,L. *et al*. (2017) One model to learn them all. *Int. J. Comput. Vision*.

Katzman,J. *et al*. (2016) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, **1606**, 1–15.

LeCun,Y. *et al*. (2015) Deep learning. *Nature*, **521**, 436.

Litovkin,K. *et al*. (2014) Methylation of PITX2, HOXD3, RASSF1 and TDRD1 predicts biochemical recurrence in high-risk prostate cancer. *J. Cancer Res. Clin. Oncol*., **140**, 1849–1861.

Liu,Y. *et al*. (2017) MiRNAs predict the prognosis of patients with triple negative breast cancer: a meta-analysis. *PLoS One*, **12**, e0170088.

Lovly,C.M. *et al*. (2016) Tumor heterogeneity and therapeutic resistance. *Am, Soc. Clin. Oncol. Educ. Book*, **36**, e585–e593.

Luck,M. *et al*. (2017) Deep learning for patient-specific kidney graft survival analysis. arXiv:1705.10245.

Maaten,L. v d. and Hinton,G. (2008) Visualizing data using t-sne. *J. Mach. Learn. Res*., **9**, 2579–2605.

Madabhushi,A. and Lee,G. (2016) Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal*., **33**, 170–175.

Malta,T.M. *et al*. (2018) Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, **173**, 338–354.

Momeni,A. *et al*. (2018a) Deep recurrent attention models for histopathological image analysis, bioRxiv, 438341. Springer Nature Switzerland, Cham, Switzerland.

Momeni,A. *et al*. (2018b) Dropout-enabled ensemble learning for multi-scale biomedical data. In: *International MICCAI Brainlesion Workshop*, pp. 407–415. Springer.

Qiu,Y.L. *et al*. (2018) *A deep learning framework for imputing missing values in genomic data*, bioRxiv, 406066.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Srivastava,R.K. *et al*. (2015) Highway networks. arXiv:1505.00387.

Wager,S. *et al*. (2013) Dropout training as adaptive regularization. In: *Advances in Neural Information Processing Systems*, Curran Associates, Red Hook, NY, pp. 351–359.

Wang,H. *et al*. (2014) Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinform*., **15**, 310.

Wang,S. *et al*. (2017) Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. *Med. Image Anal*., **40**, 172–183.

Weinstein,J.N. *et al*. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet*., **45**, 1113.

Yao,J. *et al*. (2016) Imaging biomarker discovery for lung cancer survival prediction. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 649–657. Springer Nature Switzerland, Cham, Switzerland.

Zagoruyko,S. and Komodakis,N. (2016) Wide residual networks. arXiv: 1605.07146.

Zhang,X. *et al*. (2017) Pathway-structured predictive model for cancer survival prediction: a two-stage approach. *Genetics*, **205**, 89–100.

Zhou,B. *et al*. (2014) Learning deep features for scene recognition using places database. In: *Advances in Neural Information Processing Systems*, Curran Associates, Red Hook, NY, pp. 487–495.

Zhu,X. *et al*. (2016) Imaging-genetic data mapping for clinical outcome prediction via supervised conditional Gaussian graphical model. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 455–459. IEEE, Danvers, MA.

Zhu,X. *et al*. (2017) WSISA: making survival prediction from whole slide histopathological images. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitas, CA, pp. 7234–7242.