Deep Multimodal Graph-Based Network for Survival Prediction from Highly Multiplexed Images and Patient Variables

Xiaohang Fu[1,*], Ellis Patrick[2,3,4,5], Jean Y. H. Yang[2,3,4,5], David Dagan Feng[1], and **Jinman Kim[1,4,5]**

[1]School of Computer Science, Faculty of Engineering, The University of Sydney, NSW 2006, Australia
[2]School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia
[3]Centre for Cancer Research, Westmead Institute of Medical Research, The University of Sydney, NSW 2145, Australia
[4]Sydney Precision Data Science Centre, The University of Sydney, NSW 2006, Australia
[5]Laboratory of Data Discovery for Health Limited ($D^2$4H), Science Park, Hong Kong SAR, China
*To whom correspondence should be addressed.
Contact: xiaohang.fu@sydney.edu.au; ellis.patrick@sydney.edu.au; jean.yang@sydney.edu.au; dagan.feng@sydney.edu.au; jinman.kim@sydney.edu.au

Abstract

The spatial architecture of the tumour microenvironment and phenotypic heterogeneity of tumour cells have been shown to be associated with cancer prognosis and clinical outcomes, including survival. Recent advances in highly multiplexed imaging, including imaging mass cytometry (IMC), capture spatially resolved, high-dimensional maps that quantify dozens of disease-relevant biomarkers at single-cell resolution, that contain potential to inform patient-specific prognosis. However, existing automated methods for predicting survival typically do not leverage spatial phenotype information captured at the single-cell level, and current methods tend to focus on a single modality, such as patient variables (PVs). There is no end-to-end method designed to leverage the rich information in whole IMC images and all marker channels, and aggregate this information with PVs in a complementary manner to predict survival with enhanced accuracy. We introduce a deep multimodal graph-based network (DMGN) that integrates entire IMC images and multiple PVs for end-to-end survival prediction of breast cancer. We propose a multimodal graph-based module that considers relationships between spatial phenotype information in all image regions and all PVs, and scales each region–PV pair based on its relevance to survival. We propose another module to automatically generate embeddings specialised for each PV to enhance multimodal aggregation. We show that our modules are consistently effective at improving survival prediction performance using two public datasets, and that DMGN can be applied to an independent validation dataset across the same antigens but different antibody clones. Our DMGN outperformed state-of-the-art methods at survival prediction.

Index Terms — Deep Learning; Graph Neural Network (GNN); Imaging Mass Cytometry (IMC); Multimodal; Survival Analysis

1. Introduction

The spatial organisation and structure of the tumour microenvironment surrounding tumour cells, and cellular phenotypic heterogeneity are associated with clinical outcomes and prognosis in cancer [1-3]. This information may be valuable for improving the accuracy of survival prediction models, which aim to predict the time to death events given an array of information regarding the patient. Recent techniques in highly multiplexed imaging mass cytometry (IMC) use isotope-tagged antibodies to capture maps of the spatial expression of over 30 biomarkers at the single-cell level [3-5]. These images contain detailed high-dimensional information and provide opportunities for further quantitative analysis, including survival prediction.

Traditional survival analysis typically relies on the Cox proportional hazards (CPH) model [6]. More recently, deep learning (DL) models have been successfully applied in a wide range of applications, including modelling the survival function. In contrast to conventional survival models, DL models do not require manual feature generation and selection, as they can automatically extract deep features from the inputs that are relevant to survival. As an example, DeepSurv [7] used a negative Cox partial likelihood loss to handle right-censored survival data. However, existing DL survival prediction methods typically focus on analysing a single input modality, especially patient variables (PVs), which include the patient's clinical information such as age or gene expression [7-11]. These methods do not consider spatial phenotype information at the cellular level that is relevant to disease prognosis.

Multiplexed images have recently been leveraged to predict oestrogen receptor (ER) status [12]. This demonstrated the potential of a graph neural network (GNN) [13], which models objects as nodes and capture the relationships between them as edges, to represent

cellular communities and cell-cell interactions. However, this approach is unimodal and does not consider PVs. It also relies on considerable image pre-processing to segment cells and create neighbourhoods, which adds extra labour and potential errors that would propagate to the survival analysis downstream.

1.1. Our contributions

- We introduce a deep multimodal graph-based network (DMGN) that integrates highly multiplexed IMC images and multiple PVs for survival prediction of cancer. To our knowledge, our method is the first to combine these two modalities for end-to-end survival prediction. The source code is available at https://github.com/xhelenfu/DMGN_Survival_Prediction.
- We propose a region-based multimodal relational module (RMRM) to combine features from both modalities. The RMRM graph considers relationships between all PVs and spatial phenotype information across the entire IMC image and all biomarkers. The relationship between each PV and spatial region is scaled based on its strength, which allows the RMRM to capture relationships in more detail and support explainability.
- We propose a patient variable embedding module (PVEM) to automatically generate embedded representations that are meaningful to each PV through auxiliary reconstruction tasks. These representations support the relational learning with image features in the RMRM and improve survival prediction.
- Our method does not require cell segmentation, marker selection, or any other input processing external to the basic DL data pipeline.

2. Methods

2.1. Datasets

We used the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset [3, 4] as the primary dataset. IMC images and PVs were available for 456 patients (518 samples). The uncensored rate (observed death) was 45.2%, and the median survival time was 7.97 years. We used 9 variables as done by previous studies [7-10], comprising 5 clinical features (age at diagnosis, ER status, chemotherapy indicator (CT), hormone treatment indicator (HT), and radiotherapy indicator (RT)), and 4 gene indicators (MKI67, EGFR, PGR, and ERBB2). Invasive carcinoma cores of 0.6 mm were labelled with antibodies and quantified using a Hyperion Imaging Mass Cytometer (Fluidigm). The IMC images contained 50 channels with heights of 25–744 (average 464) pixels, and widths of 60–834 (average 478) pixels.

We used the Basel dataset [5] for further cross-validation and independent validation. IMC images and PVs were available for 285 patients (376 samples). The uncensored rate was 28.1%, and the median survival was 6.17 years. We used the same PVs as for the METABRIC dataset, except radiotherapy indicator (RT) was unavailable and was substituted by immune therapy (IT). Cores of 0.8 mm were analysed using a Hyperion Imaging System (Fluidigm). The IMC images contained 52 channels with heights of 223–883 (average 711) pixels, and widths of 178–1032 (average 746) pixels.

We used all censored and uncensored samples, and marker channels for both datasets in the model inputs. All images were resized to 176 × 176 pixels with bilinear interpolation. Details of the datasets are in Table 1.

2.2. Deep multimodal graph-based network (DMGN)

In our method, information from each modality is first extracted by two separate branches. IMC images are processed by a CNN backbone to output a grid of feature vectors that describe biomarker expressions in spatial regions (patches of pixels) across the image. PVs are processed by the PVEM to generate variable-specific embedding vectors. The feature vectors for both modalities are then integrated in the RMRM and used to predict survival. We present an overview of our method in Fig. 1.

We use the versatile ResNet-50 model [14] as the CNN backbone for visual feature extraction. We do not select for certain IMC markers; all channels in the IMC image are used. We insert a 1 × 1 convolution layer immediately after the 'conv5_x' block of the backbone to reduce the number of feature channels from 2,048 to 128. This minimises the risk of overfitting by reducing the number of parameters in the model. The output of the CNN backbone is a feature volume $\mathbf{h}^{\mathbf{IMC}} \in \mathbb{R}^{6 \times 6 \times 128}$ for each input image.

2.2.1. Patient variable embedding module (PVEM)

Prior to multimodal relational learning, each PV is transformed into feature embeddings by the PVEM. There are 9 units in the module, each specialised for a PV. The units consist of fully connected (FC) layers arranged to perform autoencoding of the input variable, i.e., they learn to reconstruct the input variable.

We define the units to output PV embeddings and reconstructions as:

$$\mathbf{h}_j^{\mathbf{PVEM}}(\mathbf{x}_j) = \mathbf{W}_{\mathbf{0},j}\mathbf{x}_j + \mathbf{b}_{\mathbf{0},j} \tag{1}$$

$$\mathbf{p}_j^{\mathbf{PVEM}}(\mathbf{h}_j^{\mathbf{PVEM}}) = \mathbf{W}_{\mathbf{1},j}\mathbf{ReLU}(\mathbf{h}_j^{\mathbf{PVEM}}) + \mathbf{b}_{\mathbf{1},j} \tag{2}$$

where $\mathbf{p}_j^{\mathbf{PVEM}} \in \mathbb{R}^{1 \times k}$ denotes the prediction for variable j. $\mathbf{p}_j^{\mathbf{PVEM}}$ is a single scalar (k = 1) for continuous PVs (age, MKI67, EGFR, PGR, and ERBB2), or class probabilities (k

= 2) for categorical variables (ER status, CT, HT, and RT/IT) with softmax activation. $\mathbf{h}_j^{\mathbf{PVEM}} \in \mathbb{R}^{128 \times 1}$ is the variable-specific feature vector embedding. The PVEM unit of a category is characterised by a set of trainable weights and biases of two FC layers, comprising $\mathbf{W}_{\mathbf{0},j} \in \mathbb{R}^{k \times 128}$, $\mathbf{W}_{\mathbf{1},j} \in \mathbb{R}^{128 \times k}$, $\mathbf{b}_{\mathbf{0},j} \in \mathbb{R}^{1 \times 128}$, and $\mathbf{b}_{\mathbf{1},j} \in \mathbb{R}^{1 \times k}$.

During training, $\mathbf{p}^{\mathbf{PVEM}}$ is used to compute auxiliary reconstruction losses L^{PVEM} that contribute to the total loss. L^{PVEM} is the sum of the mean squared errors of the continuous variables and cross-entropy losses of the categorical variables.

By minimising the auxiliary losses, PVEM ensures that $\mathbf{h}^{\mathbf{PVEM}}$ is a meaningful representation of each PV. This assists with multimodal relational processing in the RMRM. The auxiliary predictions are only used during model training and have no role at inference time.

2.2.2. Region-based multimodal relational module (RMRM)

The RMRM structure considers relationships between the spatial phenotype information in each IMC image region and all 9 PVs. This structure is designed to encode all pairs of relationships between each IMC region and PV, and allows each pair to be scaled based on its relevance to survival. We use two graph convolutional network (GCN) layers with the same graph topologies to integrate features from both modalities in the RMRM. We represent each spatial region as separate nodes in the graph, and each PV is further represented by individual nodes. The input feature vector for the node of IMC region *i* is $\mathbf{h}_i^{\mathbf{IMC}}$, and PV *j* is $\mathbf{h}_j^{\mathbf{PVEM}}$. There are 45 nodes in each RMRM layer (36 IMC regions and 9 PVs).

The encoded features of the nodes are transformed and passed between the two modalities via bidirectional edges and graph convolutions. In this way, IMC features are

integrated with PV features, and PV features are also integrated with IMC features. Graph convolutions allow each node to receive the aggregated feature vectors of its neighbouring (connected) nodes. Each IMC region receives the features of all PVs (through PV–IMC edges), and each PV receives the features of all IMC regions (through IMC–PV edges). Additionally, there is a self-edge for each node which preserves some features of each own node during a graph convolution. The adjacency matrix that describes the RMRM is presented in Fig. 2.

Each graph convolution in the RMRM linearly transforms the neighbouring feature vectors via learned weights, then sums the resulting vectors and applies an activation function to the output. In a standard GCN, the neighbouring nodes are weighted equally towards the central node. We used the graph attention network (GAT) [15] approach to allow the contributions of each neighbour node towards the central node to be adaptive, rather than equal and fixed. This allows the most important information to be prioritised. GAT operates on the graph layers of the GRM with topologies and connections as described in Fig. 2. The graph attentional layer computes normalised attention coefficients via a self-attention mechanism:

$$\alpha_{uv} = \frac{\exp\left(\mathrm{LeakyReLU}(\mathbf{a}^{\boldsymbol{T}}[\mathbf{W}\mathbf{h}_u \parallel \mathbf{W}\mathbf{h}_v])\right)}{\sum_{k\in\mathcal{N}_u}\exp\left(\mathrm{LeakyReLU}(\mathbf{a}^{\boldsymbol{T}}[\mathbf{W}\mathbf{h}_u \parallel \mathbf{W}\mathbf{h}_k])\right)} \tag{3}$$

where $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N\}$, $\mathbf{h}_n \in \mathbb{R}^F$ represents the input feature vectors to the GAT layer with $N = 45$, corresponding to the 9 PVs and 36 IMC regions (i.e., $\mathbf{h}^{\mathbf{PVEM}} \in \mathbb{R}^{128\times 9}$ and $\mathbf{h}^{\mathbf{IMC}} \in \mathbb{R}^{128\times 36}$), α_{uv} is the attention coefficient for node v to u, and $\mathcal{N}_u$ denotes the neighbourhood nodes of u. The feature vectors are linearly transformed via $\mathbf{W} \in \mathbb{R}^{F'\times F}$, the weight matrix shared for every node, then by $\mathbf{a} \in \mathbb{R}^{2F'}$, the weight vector of a single-layer feedforward neural network of the attention

mechanism, with || representing concatenation. LeakyReLU nonlinearity (with a slope of 0.2 for inputs < 0) followed by softmax normalisation are then applied.

The updated feature vectors, $\mathbf{h}' = \{\mathbf{h}'_1, \mathbf{h}'_2, \dots, \mathbf{h}'_N\}$, $\mathbf{h}'_n \in \mathbb{R}^{F'}$, can then be computed by proportionally aggregating over neighbouring nodes:

$$\mathbf{h}'_u = \sigma\left(\sum\nolimits_{v \in \mathcal{N}_u} \alpha_{uv} \mathbf{W}\mathbf{h}_v\right) \tag{4}$$

where σ is the exponential linear unit (ELU) nonlinearity [16].

Furthermore, we used the multi-head attention mechanism for better stability, computing multiple independent attention coefficients for each node that are aggregated via concatenation:

$$\mathbf{h}'_u = \|_{m=1}^{M} \sigma\left(\sum\nolimits_{v \in \mathcal{N}_u} \alpha_{uv}^{m} \mathbf{W}^m \mathbf{h}_v\right) \tag{5}$$

where, for the m-th attention head, α_{uv}^{m} are the normalised attention coefficients and $\mathbf{W}^m$ are the weights of the input linear transformation. In the first layer, we use 8 attention heads per node and F_0 = 8, thereby computing a total of 64 features, as done in the original paper [15]. For the second layer, we compute F_1 = 128 features with a single attention head.

2.2.3. Survival prediction and training

The RMRM outputs of the IMC feature volumes $\mathbf{h}^{\mathbf{IMC}''} \in \mathbb{R}^{128 \times 36}$ are subsequently vectorised by a global average pooling (GAP) layer [17], producing a vector of $\mathbf{g}^{\mathbf{IMC}} \in \mathbb{R}^{128 \times 1}$ per image. This vector is then concatenated with all RMRM PV output vectors, resulting in one vector with dimensionality 1,280 × 1 per input sample. The last group of layers comprises an FC layer, ReLU activation, and a final output FC layer with linear activation.

The output of the model is a scalar $\hat{\mathrm{h}}_\theta(x)$ that estimates the log-risk function in the Cox model [7]. The model is trained by minimising the sum of L^{PVEM} and mean negative Cox partial likelihood loss L^{Surv}, which is suitable for right-censored survival data:

$$L^{Surv} = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{\mathrm{h}}_\theta(x_i) - \log \sum_{j \in \mathcal{R}(T_i)} e^{\hat{\mathrm{h}}_\theta(x_j)} \right) \tag{6}$$

$$L^{total} = \lambda L^{PVEM} + L^{Surv} + \gamma \|\theta\|_2^2 \tag{7}$$

where x_i denotes the input IMC image and PVs of sample i, θ represents the model parameters, E indicates an observed event ($N_{E=1}$ is the number of patients with recorded death), T is the overall survival time (for E = 1) or censored time (for E = 0), and $\mathcal{R}(T_i)$ are the patients with $T \geq T_i$. λ is a hyperparameter to scale L^{PVEM}, and γ is the L_2 regularisation weight.

2.3. Implementation details

The following implementation and hyperparameter choices were kept consistent for all experiments to ensure a fair comparison. The networks were trained end-to-end from scratch for 300 epochs with a batch size of 8. We employed the Adam optimiser [18] to minimise the training loss (Eq. 7) at a fixed learning rate of 0.0001, with a first moment estimate of 0.9 and second moment estimate of 0.999. We set λ to 4, as determined empirically (evaluation in Supplementary Table S1), and γ to 0.0001. Weights of the convolutional layers were initialised using He et al.'s method [19], while biases were initialised to zero. Glorot (Xavier) initialisation [20] was used for FC layers. Dropout was not used in any experiment.

The channels of the images were mean-subtracted and normalised to unit variance using the training set means and standard deviations. We employed standard online (on-

the-fly) image data augmentation by randomly applying a flip (horizontal or vertical), or rotation (of 90, 180 or 270 degrees) to the input images. The order of training examples was shuffled for every epoch. All networks were implemented using the PyTorch framework [21]. Both training and testing were performed with a 12GB NVIDIA GTX Titan V GPU.

2.4. Evaluation metrics

Survival prediction performance of the methods was evaluated using Harrell's concordance index (C-index) [22], which is one of the standard evaluation metrics in survival analysis. The C-index ranges from 0 to 1.0, where 1.0 indicates perfect ranking of predicted death times, and 0.5 indicates random predictions.

The log-rank test was used to assess the difference between two Kaplan-Meier survival curves. A log-rank *p*-value below 0.05 is considered statistically significant.

2.5. Ablation study and independent validation

We performed an ablation study to validate each component of our DMGN by assessing their effects on survival prediction performance. We first determined the performance using only a single modality—either IMC images or PVs. The model for image-only inputs consisted of a ResNet-50 backbone with the same survival prediction output layers as in the DMGN. The model for the PVs consisted of the PVEM (but without auxiliary reconstruction losses) and the same survival prediction output layers.

We then benchmarked conventional approaches to combine multimodal information, including concatenation of multimodal feature vectors, and attention-weighted

aggregation. Attention is a common approach in many multimodal applications that use weights to select and combine relevant features from both modalities [23-25].

We assessed the performance of our DMGN without auxiliary PVEM reconstruction losses. Furthermore, we investigated a modified RMRM that operates on $\mathbf{h}^{\mathbf{IMC}''} \in \mathbb{R}^{128 \times 1}$ instead of $\mathbf{h}^{\mathbf{IMC}''} \in \mathbb{R}^{128 \times 36}$, by moving GAP to just before the RMRM instead of just after. This effectively removes all spatial information in the RMRM by compressing the image features to a single feature vector.

We performed 5-fold cross-validation on both datasets to show the effectiveness of our method and that it outperforms previous methods using in-silico measurements in terms of higher C-index. We randomly assigned 80% of the patients for training and the remaining 20% for testing in each fold. We also carried out independent validation where the model was trained using METABRIC data but validated using all available Basel data. We used markers common to both datasets and removed RT from METABRIC data and IT from Basel data. There was no additional input pre-processing or model finetuning for this independent validation.

2.6. Comparison to existing methods

We benchmarked our proposed DMGN to existing survival prediction methods using both datasets. State-of-the-art methods for survival prediction on the METABRIC dataset included DeepHit [11], DeepSurv [7], BroadSurv [9], DAGSurv [10], and AttentionSurv [8]. We report the C-index for the 3-year predicted survival for DAGSurv as it is a discrete-time survival model, following previous survival analysis studies [26, 27]. All these methods were designed for single modality inputs and use the same 9 PVs as our DMGN (age, ER status, CT, HT, RT, MKI67, EGFR, PGR, and ERBB2), except for

DeepHit, which used 21 variables. We also report results for the CPH model [6] and random survival forest (RSF) [28].

We used the CPH model [6] with various input modality combinations and regularisation as the benchmark for the Basel dataset, due to a lack of recent methods. We used either L_1, L_2, or L_1+L_2 regularisation strategies. The inputs were either cell type proportions, PVs, or both cell type proportions and PVs. Cell type proportions of each IMC image were computed using metacluster cell type counts [5], divided by the total cell count for the image.

3. Results

3.1. Ablation study and validation

The prediction performance of our DMGN was consistently superior compared to the various DL baselines for the two datasets (Table 2). The single modality (image or PVs only) results indicate that PVs contain more useful information for survival prediction compared to images. The higher scores of all multimodal approaches compared to single modality approaches demonstrate the importance in using both modalities.

In both datasets the concatenation approach improved performance slightly compared to using only PVs. Concatenation is a suboptimal approach to combine multimodal information as it does not extract any further useful relationships from the two modalities that would be more informative compared to single modality information. The improvements from attention-weighted aggregation were more substantial, as this

approach prioritises the multimodal features according to their importance for survival prediction.

All DMGN variants were superior to concatenation and attention-based aggregation. Performance of the DMGN was lower without auxiliary reconstruction losses. This finding supports the motivation and design of reconstruction losses, i.e., they allow the PVEM output features to be more descriptive of each PV, thereby allowing more meaningful relational learning with image features and ultimately aiding survival prediction.

DMGN performance was also inferior without spatial image information during relational processing in the RMRM. Our proposed topology considers different relationships between each PV to various regions of the image. This allows multimodal relationships to be extracted in more detail. The ablation study shows that the spatial simplification of IMC images has a deleterious effect in the present multimodal fusion task.

To further illustrate the utility of our model, the Kaplan-Meier survival curves for one testing fold from the METABRIC (left) and Basel (right) datasets are shown in Fig. 3. Our DMGN successfully stratified the patients into two significantly different high and low risk groups (log-rank $p = 3\times10^{-21}$ for METABRIC, $p = 2\times10^{-9}$ for Basel). The curves for the other folds displayed similarly significant separation between the two groups (all log-rank $p < 10^{-5}$).

The independent validation performance for our DMGN that was trained on METABRIC data but tested on Basel data was lower than our model that was both trained and tested on Basel data (Table 3). This finding is not surprising for DL, since learned weights adapt to the characteristics of the training set and typically do not generalise to

new data at the same accuracy [29]. The reduction in independent validation performance would also be compounded by heterogeneous characteristics between the datasets, and differences in panel design and imaging protocols [3-5].

3.2. Comparison to existing methods

Our DMGN had superior survival prediction performance compared to all existing methods for both datasets (Table 4 and Table 5). Performance of CPH with different input modalities on the Basel dataset (Table 5) followed the same pattern as the DL ablations experiments (Table 2), where image-only inputs resulted in the lowest performance. Performance using PV-only inputs was higher but was surpassed by multimodal inputs.

3.3. Weights of patient variables

We analysed the weights of the RMRM connections between PVs and image regions (PV–IMC intermodal edges, cyan in Fig. 2). The average weights of the test samples in all cross-validation folds and their rankings for both datasets are presented in Table 6. The weights revealed different patterns and characteristics of the two cohorts. The rankings for the two cohorts were different, e.g., the top two PVs were ‘age’ and ‘MKI67’ for METABRIC but were ‘ERBB2’ and ‘HT’ for Basel. The differences in the weights of the PVs were smaller for Basel compared to METABRIC.

4. Discussion

4.1. Ablation study and validation

Our modules provide advantages that are not feasible in conventional multimodal fusion approaches such as with concatenation and attention. The RMRM captures multimodal relationships across different spatial regions of the image, and scales the contributions from each feature to prioritise the most important relationships. Multimodal aggregation is also carried out with meaningful representations of the PVs due to the PVEM. The ablation study (Table 2) validated the benefits of using both modalities (IMC images and PVs), and the effectiveness of our DMGN architecture with PVEM and RMRM for survival prediction.

Our DMGN has demonstrated competence in discriminating high- and low-risk groups (Fig. 3), which is practically beneficial as this knowledge may be used to guide personalised patient management. However, the performance of DMGN may be limited when directly applied to an independent cohort with different data characteristics (Table 3). While it was encouraging that our DMGN was robust to a reordering and omission of markers when trained and tested on IMC experiments with similar but ultimately different panel designs, standardised imaging protocols or pre-processing would clearly help to ensure comparable image intensity distributions and alleviate performance reductions [30].

4.2. Comparison to existing methods

In addition to the contributions provided by our PVEM and RMRM, our DMGN is distinct from existing methods in our use of IMC images in addition to PVs. These images

spatially quantify multiple clinically established breast cancer biomarkers such as MKI67 and HER2, and characterise spatial cellular heterogeneity, composition, and organisation [3-5], which are potential indicators of cancer prognosis [1, 2]. The use of a single modality in survival analysis (i.e., PVs) as done by previous studies is therefore limited, and our results show that the value in multiplexed images should be exploited. Our model can leverage both modalities, and aggregate PVs and spatial features across the image to consistently outperform the state-of-the-art (Table 4 and Table 5).

4.3. Weights of patient variables

The PVEM allows a level of explainability as the relative strengths between the PVs and image regions can be assessed according to the weights of each connection in the graph. This ability of the PVEM may be used to understand the contributing factors towards survival predictions and signify samples for further analysis. For example, image regions with particularly high or low PVEM weights to certain PVs may be quantitatively analysed at the cellular level.

4.4. Future work

While our DMGN was specifically optimised and validated on IMC data, there is potential for it to be extended and applied to other multiplexed imaging assays [31]. This could include other proteomic assays such as co-detection by indexing (CODEX) [32] and cyclic immunofluorescence (CycIF) [33] which can measure a similar number of protein markers to IMC but can measure larger regions of tissue. It could also include transcriptomic assays such as sequential Fluorescence in situ Hybridization (seqFISH+)

[34] and high-definition spatial transcriptomics [35] that can measure an order of magnitude more features.

PV embeddings may be integrated with image features at different stages in the CNN. Such a design may be more effective at leveraging image features at different scales. Future iterations of the DMGN can include components designed to increase the explainability and interpretability of the architecture. Cell segmentation masks, or cell segmentation as a secondary objective may be incorporated into the architecture to delineate relationships between IMC markers and PVs that are inside or outside cell areas. This may be beneficial for performance or explainability.

5. Conclusion

We introduced an end-to-end DL model for survival prediction that integrates multiple PVs with biomarker expressions from highly multiplexed images. Our RMRM aggregates features from all regions and channels of IMC images with PVs, and scales each intermodal relationship to improve survival prediction performance. Our PVEM can automatically produce embedded representations for each PV to support intermodal learning and survival prediction. Our method showed potential to be directly applied to an independent dataset without further training, and consistently surpassed existing methods for predicting the survival of breast cancer patients.

Acknowledgements

The authors thank their colleagues at The University of Sydney and the Sydney Precision Data Science Centre for their support and intellectual engagement.

Funding

This work has been partly supported by Australian Research Council grants (DP200103748 and DE200100944), and AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong.

Conflict of Interest: none declared.

Data availability

All data used in this study are publicly available. The METABRIC dataset [3, 4] may be obtained from https://idr.openmicroscopy.org/ (accession code idr0076). The Basel dataset [5] may be obtained from https://doi.org/10.5281/zenodo.3518284.

References

[1] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature Reviews Clinical Oncology,* vol. 15, no. 2, pp. 81-94, 2018/02/01 2018, doi: 10.1038/nrclinonc.2017.166.

[2] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu, "Tumour heterogeneity in the clinic," *Nature,* vol. 501, no. 7467, pp. 355-364, 2013/09/01 2013, doi: 10.1038/nature12627.

[3] H. R. Ali *et al.*, "Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer," *Nature Cancer,* vol. 1, no. 2, pp. 163-175, 2020/02/01 2020, doi: 10.1038/s43018-020-0026-6.

[4] C. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature,* vol. 486, no. 7403, pp. 346-352, 2012/06/01 2012, doi: 10.1038/nature10983.

[5] H. W. Jackson *et al.*, "The single-cell pathology landscape of breast cancer," *Nature,* vol. 578, no. 7796, pp. 615-620, 2020/02/01 2020, doi: 10.1038/s41586-019-1876-x.

[6] D. R. Cox, "Regression Models and Life-Tables," in *Breakthroughs in Statistics: Methodology and Distribution,* S. Kotz and N. L. Johnson Eds. New York, NY: Springer New York, 1992, pp. 527-541.

[7] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology,* vol. 18, no. 1, p. 24, 2018/02/26 2018, doi: 10.1186/s12874-018-0482-1.

[8] Z. Sun, W. Dong, J. Shi, K. He, and Z. Huang, "Attention-Based Deep Recurrent Model for Survival Prediction," *ACM Trans. Comput. Healthcare,* vol. 2, no. 4, p. Article 35, 2021, doi: 10.1145/3466782.

[9] G. Wu, J. Duan, J. Wang, L. Wang, C. Dong, and C. Lv, "BroadSurv: A Novel Broad Learning System-based Approach for Survival Analysis," in *2021 8th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, 10-12 Dec. 2021 2021, pp. 26-31, doi: 10.1109/ICCSS53909.2021.9721940.

[10] A. K. Sharma, R. Kukreja, R. Prasad, and S. Rao, "DAGSurv: Directed Ayclic Graph Based Survival Analysis Using Deep Neural Networks," presented at the Proceedings of The 13th Asian Conference on Machine Learning, Proceedings of Machine Learning Research, 2021.

[11] C. Lee, W. Zame, J. Yoon, and M. van der Schaar, "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32, no. 1, 04/26 2018, doi: 10.1609/aaai.v32i1.11842.

[12] P. Martin-Gonzalez, M. Crispin-Ortuzar, and F. Markowetz, "Predictive modelling of highly multiplexed tumour tissue images by graph neural networks," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*: Springer, 2021, pp. 98-107.

[13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems,* vol. 32, no. 1, pp. 4-24, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, 2015.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903,* 2017.

[16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," 2016.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations (ICLR)*, 2014.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 7-13 Dec. 2015 2015, pp. 1026-1034, doi: 10.1109/ICCV.2015.123.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[21] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026-8037.

[22] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction," *Statistics in medicine*, vol. 3, no. 2, pp. 143-152, 1984.

[23] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated Multimodal Units for Information Fusion," 2017.

[24] R. Jain and C. Wigington, "Multimodal Document Image Classification," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 20-25 Sept. 2019 2019, pp. 71-77, doi: 10.1109/ICDAR.2019.00021.

[25] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478-493, 2020, doi: 10.1109/JSTSP.2020.2987728.

[26] Y. Zhang, E. M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M. A. Haider, and F. Khalvati, "CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging," *BMC medical imaging*, vol. 20, no. 1, pp. 1-8, 2020.

[27] H. Kim, J. M. Goo, K. H. Lee, Y. T. Kim, and C. M. Park, "Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas," *Radiology*, vol. 296, no. 1, pp. 216-224, 2020.

[28] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, vol. 2, no. 3, pp. 841-860, 2008.

[29] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[30] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning," (in eng), *Radiology*, vol. 295, no. 1, pp. 4-15, 2020, doi: 10.1148/radiol.2020192224.

[31] S. M. Lewis *et al.*, "Spatial omics and multiplexed imaging to explore cancer biology," *Nature Methods*, vol. 18, no. 9, pp. 997-1012, 2021/09/01 2021, doi: 10.1038/s41592-021-01203-6.

[32] S. Black *et al.*, "CODEX multiplexed tissue imaging with DNA-conjugated antibodies," *Nature Protocols*, vol. 16, no. 8, pp. 3802-3835, 2021/08/01 2021, doi: 10.1038/s41596-021-00556-8.

[33] J.-R. Lin, M. Fallahi-Sichani, and P. K. Sorger, "Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method," *Nature Communications*, vol. 6, no. 1, p. 8390, 2015/09/24 2015, doi: 10.1038/ncomms9390.

[34] C.-H. L. Eng *et al.*, "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+," *Nature*, vol. 568, no. 7751, pp. 235-239, 2019/04/01 2019, doi: 10.1038/s41586-019-1049-y.

[35] S. Vickovic *et al.*, "High-definition spatial transcriptomics for in situ tissue profiling," *Nature Methods*, vol. 16, no. 10, pp. 987-990, 2019/10/01 2019, doi: 10.1038/s41592-019-0548-y.

Figures and Tables

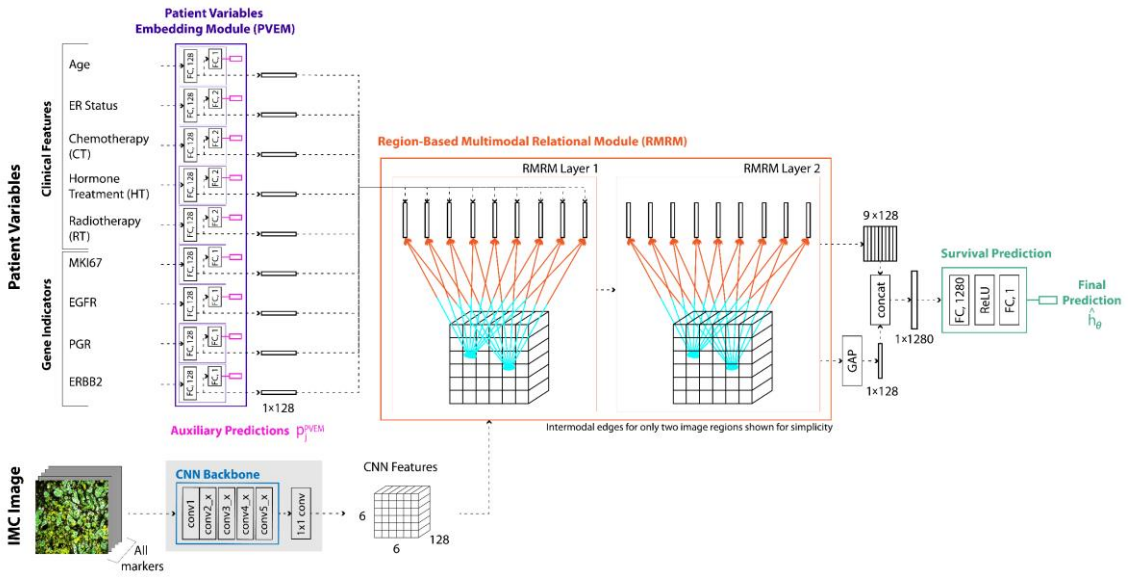

Fig. 1. Proposed DMGN model for end-to-end survival prediction from PVs and IMC images. PVs are processed by the PVEM to produce embeddings. The IMC image is processed by a CNN backbone to create deep image features. Both modalities are combined in the RMRM and then used to predict survival.

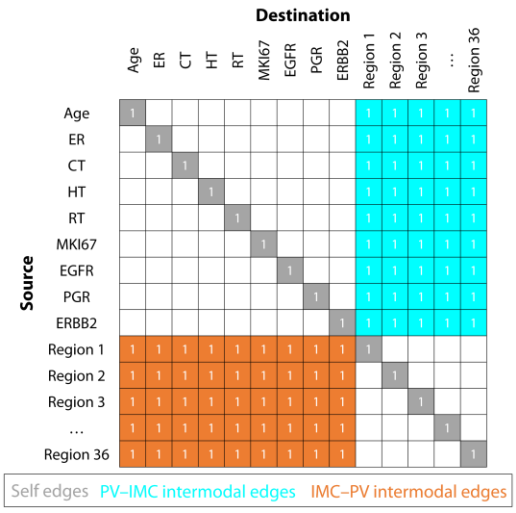

Fig. 2. RMRM adjacency matrix. Filled squares are edges between source and destination nodes. Blank squares represent the absence of an edge (zeros are omitted for simplicity). All 36 image regions are connected to the 9 PVs by bidirectional edges.

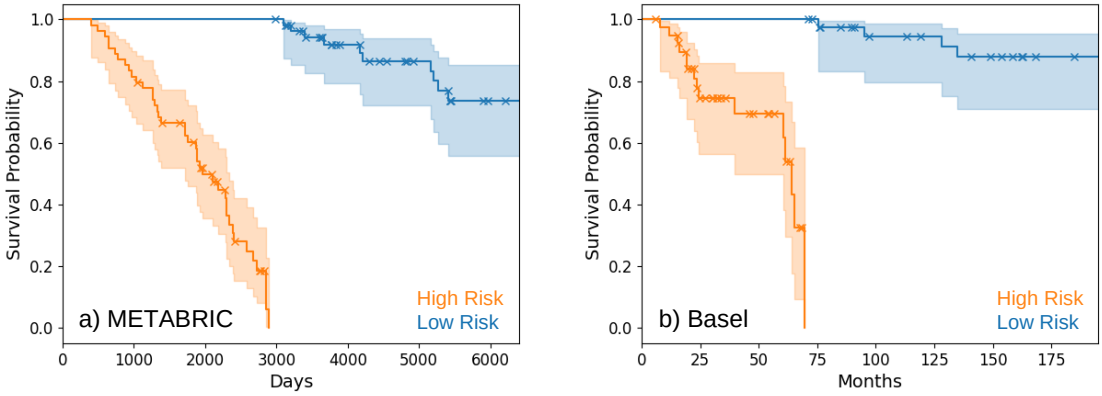

Fig. 3. Kaplan-Meier survival curves. DMGN predictions for high risk (higher than or equal to median) and low risk (lower than median) groups for a) METABRIC and b) Basel datasets for one testing fold. '×' indicates a censored patient. Predicted KM curves are significantly different for high and low risk groups (log-rank $p = 3\times10^{-21}$ for METABRIC, $p = 2\times10^{-9}$ for Basel).

Table 1. Details of the two public datasets used in this study.

	METABRIC	Basel
Number of patient variables	9	9
Median censored time	7.97 years	6.17 years
Uncensored rate	45.2%	28.1%
Total number of patients	456	285
Total number of samples	518	376
Number of IMC channels	50	52

Table 2. Prediction performance (C-index) of the proposed DMGN compared to DL baselines with different inputs and fusion methods. Bold indicates the best performance for each dataset.

Method	METABRIC	Basel
IMC image only	0.6581	0.6769
PVs only	0.7147	0.7078
Multimodal, concatenation	0.7174	0.7131
Multimodal, attention	0.7327	0.7242
Multimodal, DMGN (no aux losses)	0.7398	0.7354
Multimodal, DMGN (no spatial)	0.7399	0.7359
Multimodal, DMGN (proposed)	**0.7484**	**0.7479**

Table 3. Validation of the proposed DMGN on an independent test set.

Method	C-index
Test METABRIC (trained METABRIC)	0.7380
Test Basel (trained Basel)	0.7146
Independent test Basel (trained METABRIC)	0.6312

Table 4. Comparison of the proposed DMGN to existing methods on the METABRIC dataset.

Method	C-index
CPH	0.632
RSF	0.620
DeepSurv [7]	0.654
DeepHit [11]	0.691
BroadSurv [9]	0.643
DAGSurv [10]	0.689
AttentionSurv [8]	0.708
DMGN (proposed)	**0.748**

Table 5. Comparison of the proposed DMGN to existing methods on the Basel dataset.

Method	C-index
CPH (cell type proportions, L_2)	0.584
CPH (PVs, L_2)	0.686
CPH (cell type proportions + PVs, L_1)	0.696
CPH (cell type proportions + PVs, L_2)	0.698
CPH (cell type proportions + PVs, L_1 + L_2)	0.697
DMGN (proposed)	**0.748**

Table 6. Average RMRM weights of the PVs to image regions and their rank in descending order.

Patient Variable	METABRIC		Basel	
	Weight	Rank	Weight	Rank
Age	0.1374	1	0.0982	4
MKI67	0.1324	2	0.0967	6
EGFR	0.0935	7	0.0911	9
PGR	0.1000	4	0.0932	7
ERBB2	0.0906	8	0.1021	1
CT	0.1140	3	0.0986	3
RT (METABRIC) / IT (Basel)	0.0948	5	0.0971	5
HT	0.0894	9	0.1006	2
ER	0.0939	6	0.0925	8

Supplementary Materials

Table S1. Prediction performance (C-index) of the proposed DMGN with various λ.

λ value	METABRIC	Basel
0 (no auxiliary loss)	0.7398	0.7354
0.1	0.7407	0.7219
1.0	0.7392	0.7314
2.0	0.7462	0.7411
4.0	**0.7484**	**0.7479**
5.0	0.7283	0.7241