Multi-view Modelling of Longitudinal Health Data for Improved Prognostication of Colorectal Cancer Recurrence

Danliang Ho HO.DANLIANG@U.NUS.EDU
NUS Graduate School, Integrative Sciences and Engineering Programme
National University of Singapore

Mehul Motani MOTANI@NUS.EDU.SG
Department of Electrical and Computer Engineering, N.1 Institute for Health,
Institute of Data Science, Institute for Digital Medicine (WisDM)
National University of Singapore

Abstract

Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide, with a high incidence of recurrence following surgical resection. Accurate prognostication of recurrence risk is essential to identify patients who may benefit from adjuvant therapies and improve their health outcomes. In our study, we propose a novel approach to CRC recurrence prognostication using multi-view deep learning. Our proposed approach, Fusion with Multi-view Attention (`FMA`), integrates static and longitudinal data from heterogeneous healthcare records, and learns complex interactions between data views to predict recurrence and time-to-recurrence. Our model achieves an AUROC score of 0.97, and precision, sensitivity and specificity scores of 0.80, 0.90 and 0.95 respectively, outperforming all-known published results based on the commonly-used CEA prognostic marker, as well as state-of-the-art CRC recurrence prognostication models. We show through a sensitivity analysis that incorporating multiple data views improves model performance significantly compared to using only a single view. We also show that our model accurately stratifies patients into risk groups that are associated with the actual 5-year recurrence-free survival, paving the way towards better identification of high-risk patients who may benefit from adjuvant therapies. Our proposed approach demonstrates the potential of multi-view modelling to push state-of-the-art in CRC recurrence prognostication and could contribute towards more personalised patient management and follow-up in the clinic.

1. Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide and the second most common cause of cancer deaths, killing over 1 million people per year (WHO). Among CRC survivors, post-operative recurrence presents a major health risk, with approximately 30% of patients eventually developing recurrent disease (Young et al., 2014; Alafchi et al., 2021). Cancer relapse is responsible for a vast majority of cancer-related deaths, with patients who experience recurrence within 2 years having a reported 5-year survival at only 34.7% (Ryuk et al., 2014). Since an accurate diagnosis could inform therapeutic options, improve symptoms and prolong survival (Mahmoudi et al., 2022; Safari et al., 2021), there is strong clinical impetus for early and accurate detection of recurrence.

Today, clinicians are limited in their ability to assess which patients will suffer a recurrence. Imaging scans, the gold standard for diagnosis, increases the costs of care and poses health risks for patients. On the other hand, patient follow-up through serum carcinoembryonic antigen (CEA) remains controversial, with systematic studies questioning its accuracy and prognostic value, due to its limited sensitivity, specificity, as well as inconclusive findings on its effect on reducing patient mortality (Sørensen et al., 2016; Shinkins et al., 2017).

Prediction of cancer relapse is highly complex and involves the interplay of a large number of associated risk factors, only some of which are known to the research and clinical community. Machine learning represents an attractive means to derive prediction models for this task, due to their ability to analyze heterogeneous datasets with a high number of variables. To this end, several studies have leveraged on machine learning tools (Ting et al., 2020; Achilonu et al., 2021; Xu et al., 2020; Skrede et al., 2020; Geessink et al., 2019; Jiang et al., 2020; Pai et al., 2022). While they show promising results (for example, Xu et al. (2020) reported an AUROC score of 0.761 for their best model), most of these studies consider either clinical data or histopathological data only for their analysis.

Multi-view modelling, a methodology that captures and integrates varied and complementary views of data to enhance learning, have shown increasing popularity in the machine learning domain (Yan et al., 2021). This concept is also compelling in medicine, as one can see it as an attempt to mimic the process of clinical expert decision-making, whereby clinicians are informed by a constellation of signs, symptoms, lab values and supportive imaging, when making clinical decisions (Kline et al., 2022). A few studies have adopted multi-view techniques for modelling tumour recurrence and survival (Castellanos et al., 2017; Wang et al., 2022; Ho et al., 2021b). However, most of these studies utilise only static data obtained from a single timepoint; as of our knowledge there is very limited work that employ multi-view modelling for longitudinal data analyses.

Longitudinal data is prevalent in many healthcare applications including our task of prognosticating CRC recurrence. While static data can be seen as a single slice of information in time, longitudinal data confers additional information about how different slices relate and interdepend upon each other. Furthermore the latter is more likely to be predictive of disease with a longer time-course to onset, as patterns within longitudinal data could act as a more accurate reflection of an individual's underlying physiological state at that current time. As such, it is important to develop multi-view methods that not only integrates data from multiple sources, but also incorporates the time dependency.

In our work, we make the following contributions:

- We propose a novel deep-learning based multi-view model that we term Fusion with Multi-view Attention (`FMA`), that is capable of ingesting and integrating both static and longitudinal data from multiple sources and formats (specifically clinical data, lab measurements across time, and longitudinal radiological scan reports), to prognosticate recurrence in CRC patients. Our model achieves state-of-the-art performance at 90% sensitivity, 94.5% specificity, and 97.2% AUROC.

- `FMA` extracts good representations using specialised feature extractors, then performs view-wise attention to fuse multiple views, dynamically focusing on the view with impor-

tant features. To the best of our knowledge, we are the first to propose an attention-based approach for multi-view learning in healthcare applications.

- By extending the linear Cox proportionate hazards modelling approach to our framework, we imbue our model the capability of modelling survival data. `FMA` achieves high C-index and integrated brier scores of 0.960 and 0.036 respectively. As far as we know, our model's performance on survival data surpasses that of all other models developed for the same purpose.

Generalizable Insights about Machine Learning in the Context of Healthcare

Healthcare data is highly heterogeneous, originating from different data sources and possessing different data characteristics and structures. Multi-view learning on this already-available data source may allow us to better utilise accumulated knowledge to model complex clinical problems such as cancer recurrence prognostication with greater accuracy, compared to traditional single-view modelling. In this study we describe a versatile framework that can be applied to any kind of static and longitudinal datasets with the use of appropriate feature extractors, and extended to incorporate additional data views. Furthermore it has no hard requirements on complete information across views and can work with variable number of patients per view. We show that our method benefits downstream tasks such as prediction and survival analysis, and that our full model possesses a performance gain over the single-view models. While we demonstrate our approach on the task of CRC prognostication, our method may also be applied to address other clinical problems for better predictions.

2. Related Work

Our work is positioned in the field of multi-view analysis, which is concerned with the exploitation of complementary information from distinct feature sets, or views, to learn more comprehensive representations as compared to single-view learning methods. The main advantage of doing so is to achieve better learning performance and improve generalization. Multi-view learning methods have achieved great practical success in machine learning and have been adopted in several domains such as computer vision and natural language processing (Yang et al., 2021; Shan et al., 2022; Zhang et al., 2022; Fu et al., 2010). Techniques in multi-view modelling can be broadly categorised into methods for representation alignment (of which Canonical correlation analysis and related extensions are dominant techniques (Hotelling, 1992; Horst, 1961; Akaho, 2007)), and methods for representation fusion. Our work falls under the latter category and uses neural network-based fusion methods to learn a joint representation of the different views.

Related work in the healthcare and biomedical domain has mainly focused on the use of large -omics datasets to develop predictive models for various tasks such as biomarker development and patient or disease subtyping. The approaches are based on either supervised classification or unsupervised clustering, and may or may not involve deep neural networks (Liu et al., 2016; Higdon et al., 2015; Planey and Gevaert, 2016). An example of a study that utilised neural networks for multi-view learning is that of Castellanos et al. (2017), where the authors trained an ensemble model that used genetic mutation, proteomics and

RNA expression information to predict recurrence in CRC patients. Another important avenue of work uses multi-view imagery as a data source, for example Wang et al. (2022) and Lu et al. (2018) used deep features learnt from radiology scans, as well as clinical data to predict recurrence in head and neck cancer patients and diagnose Alzheimer's disease, respectively. Our work differs from theirs in that none of these works consider longitudinal data sources.

Studies on multi-view modelling for longitudinal data in biomedical applications are much more limited. Among existing works, Lee et al. (2019) explored longitudinal data integration methods similar to ours, but for the purpose of predicting Alzheimer's disease (AD) progression. Our work differs from theirs in two aspects: 1) The complexity of the input formats and modelling approach: they model only numerical time-series using Gated Recurrent Units while our approach extends to longitudinal text which is a more complex data format. Also at the data integration step, we do not simply concatenate the learnt fixed-length feature vectors, but rather rely on the attention mechanism to dynamically focus on the view with most important features. 2) They perform binary classification only when predicting AD progression, while we extend this to the more pertinent task of modelling survival data.

3. Methods

3.1. Cohort Selection

Our study was conducted using medical data from a cohort of approximately 1000 patients diagnosed with Stage I to Stage III CRC, with no evidence of metastatic disease. All patients underwent surgical resection of the primary tumour and were referred to a local hospital for post-operative follow-up. Informed consent was obtained for all patients prior to study enrollment, and institutional ethics approval was obtained for this study. Deidentification was performed by assigning each patient a unique serial number upon study entry, and all personal identifiers were removed prior to data analysis.

3.2. Dataset description and preprocessing

3.2.1. Clinical data

The data consisted of 65 clinical variables potentially prognostic for recurrence, including demographic data, tumour characteristics, molecular profiling results and treatment parameters. The clinical data is presented in a tabular format. Data cleaning was performed by removing errors due to misspellings, duplications, letter case, extra white space and semantically similar categories. The problem of missing data was handled by considering whether the data is likely to be Missing-Not-At-Random (MNAR) using domain knowledge, in which case we denoted them as 'not_available', otherwise imputation was performed with multiple rounds of MICE (van Buuren and Groothuis-Oudshoorn, 2011). We also mined information from unstructured text fields via rule-based text extraction and added them into this tabular dataset. All categorical data was transformed with one-hot encoding while numerical data was min-max scaled.

3.2.2. Lab monitoring data

Longitudinal lab measurement data on post-operative carcinoembryonic antigen (CEA) levels were collected at multiple timepoints between data of surgery and date of recurrence or most recent follow-up, whichever was earlier. CEA is a blood-based tumour marker that is commonly used in the clinic for population screening of CRC as well as post-operative monitoring of CRC patients, albeit having limited sensitivity and specificity (Sørensen et al., 2016; Shinkins et al., 2017). The median length of follow-up was 40 months at a frequency of between 1-3 months on average. The median data points per patient was 14.

Data cleaning was performed following the method described in Section 3.2.1. We imputed missing longitudinal data through linear interpolation for timeseries using pandas package. We resampled the data monthly to create evenly-spaced intervals, then zero-padded it to the maximum length of the time-series. Our data is numeric and right-skewed; hence we first performed log transformation to remove skewness followed by min-max scaling, prior to model input.

3.2.3. Text reports for radiological scans

We obtained the accompanying text reports for radiological scans (CT, PET and MRI imaging) for approximately 97% of the patients who were undergoing post-operative follow-up. The data was longitudinal, with each patient possessing multiple reports collected between date of surgery and date of recurrence or most recent follow-up, whichever was earlier. The text reports consisted of interpretations of the scan image by a certified radiologist. The median length of followup was 54 months, at a frequency of around 8 months, and the median number of datapoints per patient is 6.

We concatenated all reports for each patient, with the most recent report at the beginning. The problem of variable sequences lengths was handled by chunking long documents into shorter ones within the 512 token limit, or padding for documents that fall short of the maximum length. The reports were tokenized using a pre-trained tokenizer, before input into a pre-trained `ClinicalBERT` model (Huang et al., 2020) (see Section 4.4).

3.2.4. Augmentation of text reports

Since our dataset is imbalanced with 3 times more non-recurrent patients compared to recurrent ones, thus we aimed to improve the performance of our feature extractor model through data augmentation of the minority class. We increased the amount of data available to the model via two strategies:

- Replacement of random words using the nearest word embeddings
- Shuffling sentences within each report

Word embedding replacement We obtained and processed a corpus of radiological reports obtained from MIMIC-III CXR Database (Johnson et al., 2016; Goldberger et al., 2000). For each word token in the corpus, we obtained the corresponding word embedding from a pre-trained `ClinicalBERT` model, and represented it within a K-dimensional tree (KDTree) using Sklearn. We indexed the tree with the original word token. Next, we tokenised each report and replaced random tokens at a sampling rate of 0.1. For each

identified token, we queried it against our KDTree to output the nearest word embedding. The corresponding token was used to replace our original token.

Sentence shuffling Our intuition for this approach was based on the observation that our medical reports consisted largely of bullet points, and there is less dependency between sentences. Thus shuffling the sentences does not do much to change the meaning of the report. We shuffle sentences only for the standalone report for each timepoint. This was done by the use of the SPACY package to identify sentence boundaries, followed by custom code to perform a random shuffle of identified sentences.

4. Model building

4.1. Full model

`FMA` is a neural network-based architecture that processes and integrates data from multiple views. Figure 1 shows a schematic outline of the full model. It comprises a two meta-layer architecture:

1. The first meta-layer processes incoming data (tabular, time-series or text) using specialised networks adapted to extract high quality features from each view.
2. The second meta-layer is an overall network that combines outputs from the first meta-layer to learn an integrated feature representation for the prediction task.

Figure 2 shows a schematic representation of each feature extractor. We also describe the model components below.

4.2. Tabular feature extractor

It is known that deep learning underperforms on tabular data, and one reason is the lack of prior knowledge about the dataset structure that could be utilised by models with the appropriate inductive bias (Borisov et al., 2022; Shwartz-Ziv and Armon, 2021). We adopt the approach in Ho and Motani (2022) and create a deep learning model designed for tabular datasets. Our model injects spatial structure in tabular representations, that can be leveraged upon by a downstream CNN model. It first projects tabular inputs into a high-dimensional space using a fully-connected layer, followed by grouping the features as "images". Hence we may consider these high-dimensional features as different aspects of the original features, which can then be combined non-linearly using a CNN model. Our model learns the correct spatial order of these feature aspects as the FCN learns weights that determine how to project features in a manner that will allow the CNN to extract local patterns from each "image". We name this model `tab-cnn` and depict it in Figure 2(a).

4.3. Time-series feature extractor

We adopt Ho et al. (2021a)'s approach and implement a Transformer model designed for time-series data. We name this model `ts-transformer` and depict it in Figure 2(b). The model uses the Transformer backbone described by (Vaswani et al., 2017), but learns to focus on local context in time-series data, through two modifications to the architecture.

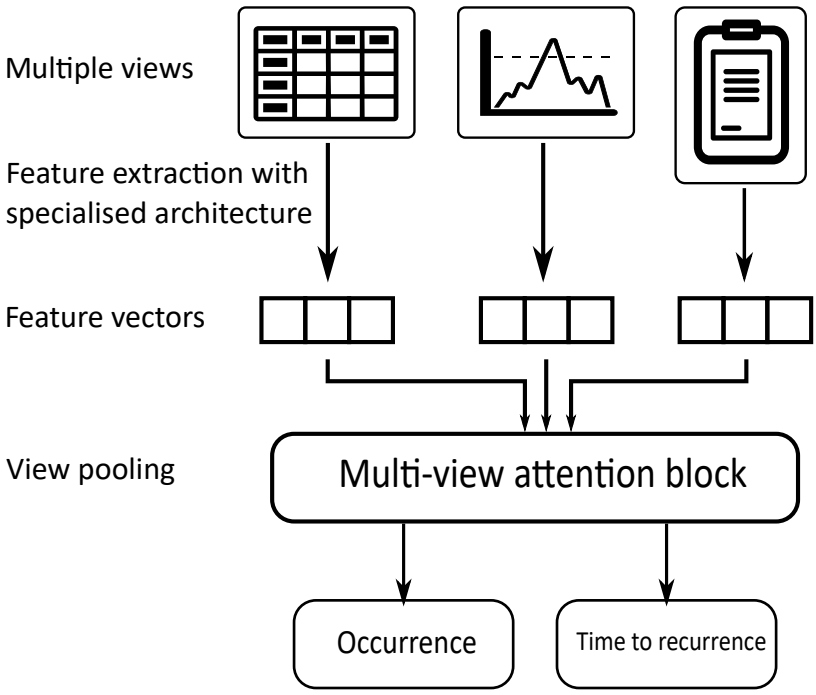

Figure 1: Schematic view of our proposed framework for multi-view modelling.

Convolutional self-attention (ConvSA) When learning the query, key and value vectors important for the self-attention calculation, instead of linear projections via fully connected networks, we employ causal 1D-CNNs with kernel size 3 and stride 1, convolving across the temporal dimension. This forces the query, key and value vectors to incorporate local context information in the resultant attention calculations.

Time-masking for localised attention (LA) We apply a time-mask in the decoder that limits the amount of backward attention, thereby restricting the decoder to only focus on short-term patterns within the immediate locality. We prevent backward attention of context past a fixed-size m by applying a lower triangular mask that sets elements below the mth-diagonal to $-\inf$ before the softmax calculation, such that the attention scores goes to zero and does not feed into the subsequent calculations.

4.4. Text feature extractor

We train a model for longitudinal text using a two-step approach. As described in Section 3.2.3, each patient has multiple text chunks due to constraints in the length of input data, up to 512 tokens. Hence, we first develop a chunk feature extractor that is not-specific to patients, and we use the pre-trained `ClinicalBERT` as our model. Each text chunk inherits the same recurrence label as that for the original patient, and we tokenise all chunks prior to feeding into the model. We finetune `ClinicalBERT` on these chunked data and we extract individual chunk features from the last embedding layer.

In the next step, we order the chunk features temporally for same patients, and create an attention model to combine these features based on the time dimension. The attention

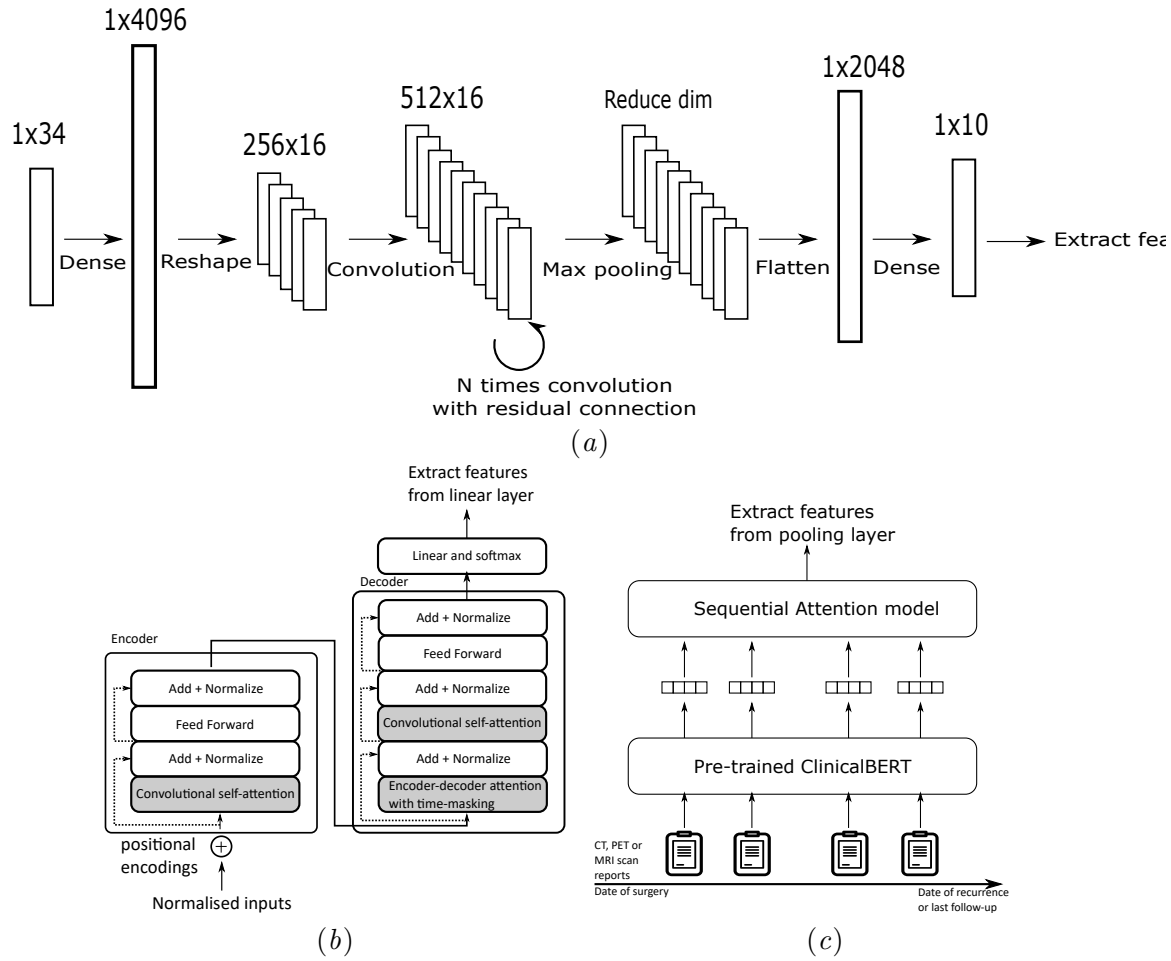

Figure 2: Proposed specialised feature extractors for a) Tabular data; b) Time-series data; c) Clinical text data

model is trained on the original patient labels, and learns which chunk to focus on for the best predictive performance. Subsequently, we extract the pooled output from the attention model as text features for the integration network. We name this model `text-temporalattn` and depict it in Figure 2(c).

4.5. Integration network

We perform view pooling by proposing the use of a multi-view attention block in our integration network, allowing the model to learn the interdependencies between data views (Figure 1). Our model consists of the following components: a) an interface that receives features extracted from any number of data views, followed by a fully-connected layer that processes each feature set, b) an attention block that performs self-attention across all views, a dropout layer at a rate of 0.1, a layer normalisation operation, c) a residual

connection between the inputs pre and post-attention block, and d) a classification layer for the prediction task. All hidden layers are set to size 20 and we use a single attention head.

5. Training and evaluation

5.1. Binary prediction of recurrence

We train our models using a two-stage approach. The individual feature extractors are first trained on the task of predicting recurrence, using all available patient data for each extractor. We train our time-series feature extractor using Tensorflow (Abadi et al., 2015), and the rest of the deep learning components using Pytorch (Paszke et al., 2017). We use RMSProp (Tieleman and Hinton, 2012) to optimise weights for the time-series feature extractor, and Adam (Kingma and Ba, 2017) for the rest. The training objective was set to minimize cross-entropy loss. Subsequently, we extract features from the second-last layer of each model, and use them as input to train the view-integrator model, on the task of predicting recurrence. We only use overlapping data from all feature extractors to train the view-integrator.

All models except the time-series feature extractor utilised a cyclical learning rate policy (Smith, 2017) that anneals the learning rate from an initial learning rate of 1e-3 to 1e-2 and then down to 1e-5, every epoch. The time-series feature extractor used a learning rate decay policy that started from 1e-4 and decayed at a factor of 0.1 when validation loss fails to improve after 8 epochs. All models were trained for at least 50 epochs, and model weights from the epoch with best validation loss were saved.

5.2. Time-to-recurrence prediction

We keep the individual feature extractors and only retrain our view-integration model for the task of modelling 5-year recurrence-free survival (RFS), as defined by the amount of time lapsed between surgery to date of recurrence diagnosis. We do so by extending the linear Cox proportionate hazards (CoxPH) model to neural network architectures, as proposed by Katzman et al. (2018). The CoxPH equation takes the form:

$$\lambda(t \mid x) = \lambda_0(t) \cdot e^{h(x)}$$

We adjust our model architecture to estimate the log-risk function $h(x)$ in the Cox model using patient covariate data, by changing the last classification layer into a single neuron for regression. The input data are the extracted features from each data view, and the output is a single node that predicts $h(x)$.

We used the package pycox to train and evaluate our model. We set the objective function to be the average negative log partial likelihood of CoxPH, and train the model for 512 epochs with a batch size of 256 samples, on the Adam optimiser. Kaplan-Meier curves and log-rank statistical tests were calculated using the package lifelines.

5.3. Performance comparisons

Our models were tuned for best hyperparameters on the validation dataset and evaluated on the test dataset. We train at least 10 separate models for each model architecture, which

differ in terms of initial weights. For binary prediction of recurrence, we report the average for the following performance metrics: Balanced Accuracy, Recall (or sensitivity), Precision (or Positive Predictive Value), F1 score, Specificity, AUROC and AUPRC. All metrics were computed using default thresholds.

To evaluate our survival model, we calculate the concordance index (C-index), defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs, and the integrated brier score, which performs an integral of the brier score at all available times $t_1 \leq t \leq t_{max}$.

To demonstrate that our model is able to accurately stratify patients into risk groups, we used our trained survival model to predict recurrence risk at 5 years. We divided patients in the test dataset into high-risk versus low-risk based on a threshold value of 0.5, and compared our predictions with the actual recurrence-free survival data using a Kaplan-Meier plot.

5.4. Baseline models

We design three categories of baselines: 1) Those that compare against the performance of the individual data views, 2) Those that compare against the performance of the view integration network, and 3) Other multi-view baselines. We adopt the following abbreviations: 'ts' for time-series, 'tab' for tabular data, 'text' for text data, and the following naming convention for our feature extractors: {data type}-{model name}.

5.4.1. For time-series feature extractor

We create strong deep learning temporal models to act as baselines - Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Temporal Convolutional networks (TCN) (Lea et al., 2017). `ts-lstm` has the following settings: bidirectional network with hidden layer size 8, tanh activation, dropout and recurrent dropout rate set to 0.2, initial learning rate set to 1e-3. `ts-tcn` has the following settings: stack of 6 convolutional blocks, each consisting of alternating convolutional and dropout layer, with residual connections between inputs and outputs. We employed causal padding to train model on early inputs. Other settings are: dilation rate 2, filter size 2, hidden layer size 60, relu activation for all layers except the classification layer.

5.4.2. For tabular feature extractor

We implement specialised deep-learning architecture for tabular data analysis, namely TabNet (Arik and Pfister, 2020) and TabResnet (Javier, 2023). Both models were implemented using pytorch-widedeep. Model hyperparameters for `tab-tabnet` are: size of embedding dimensions 32, width of attention embedding 8, num steps 3, num shared Gated Linear Units (GLU) per step 2, num independent GLU per step 2. Model hyperparameters for `tab-tabresnet` are: continuous embedding dimensions 32, block dimension [200, 100, 100], embeddings dropout rate 0.1, block dropout rate 0.1.

5.4.3. For text feature extractor

We implement Huang et al. (2020)'s approach to modelling the longitudinal aspect of the text records, in our model `text-huang`. `text-huang` utilises `ClinicalBERT` as well but

performs longitudinal modelling by first, modelling text chunks as subsequences (similar to us), and second, calculating the predicted probability of recurrence using the probability output for each subsequence. This is performed according to the equation:

$$P(recurrence = 1 \mid h_{patient}) = \frac{P^n_{max} + P^n_{mean} n/c}{1 + n/c}$$

where n is the number of subsequences for each patient and c is a scaling factor set to 2 following (Huang et al., 2020).

5.4.4. For view integration

We construct baselines for which we combine only two out of the three views, using view-specific attention outlined in 4.5. We also compare the effect of attention, by constructing a simple feedforward network baseline that we term `FMA-concatonly`. This simple model receives input from each data view, performs a concatenation of feature vectors to combine all views, and passes it through another feedforward layer before outputting the prediction. All hidden layers are set to size 20.

5.4.5. Other multi-view baselines

We derive baselines using other multi-view approaches. Broadly, we extract hand-crafted features from longitudinal data into a static tabular format using the python package Tsfresh (Christ et al., 2018), combine them with existing tabular data and apply standard machine learning methods. As it is a non-trivial problem to specify static features for longitudinal text data, we only combine time-series and tabular modalities, and leave the task of combining all three views to future work. We implemented off-the-shelf ML classifiers using scikit-learn (Pedregosa et al., 2011). Parameters were selected through extensive grid-search on the validation dataset. We investigated the following models and report the best parameters:

- **Logistic regression** (`LR`), C=0.1 and l2 penalty
- **Support vector machine** (`SVM`) radial basis function (RBF) kernel, C=1, gamma="scale"
- **Multi-layer perceptron** (`MLP`) with two dense layers (70 and 10 nodes), relu activation, dropout (0.3 and 0.15), optimizer=Adadelta

To handle class-imbalanced data, we utilized a weighted loss function, setting weights to the inverse of the corresponding class support.

6. Results

6.1. Performance comparisons

Table 1 shows the performance of our proposed framework on the test dataset, for each of the model components, as well as the our full model `FMA-alldata`. `FMA-alldata` achieves a strong performance of 90% sensitivity, 94.5% specificity, and 97.2% AUROC, and it surpasses all the baselines for combined (`FMA-concatonly`) and individual data views in F1,

Table 1: Comparing performance of individual feature extractors, and our final view-integration model `FMA`, with several baselines. Metrics are reported as average (standard error) of at least 10 model initialisations. Best scores are bolded and second-best scores are underlined.

Model	Bal Accuracy	Sensitivity	Specificity	PPV	F1	AUROC	AUPRC
Combined models using FMA							
`FMA-alldata`	**0.924 (0.003)**	0.903 (0.006)	0.945 (0.005)	0.795 (0.016)	**0.845 (0.009)**	**0.972 (0.002)**	**0.944 (0.002)**
`FMA-tabts`	0.801 (0.014)	0.731 (0.032)	0.872 (0.022)	0.604 (0.049)	0.646 (0.026)	0.899 (0.011)	0.758 (0.035)
`FMA-tabtext`	0.883 (0.008)	0.863 (0.019)	0.904 (0.006)	0.679 (0.011)	0.758 (0.009)	0.946 (0.006)	0.821 (0.027)
`FMA-tstext`	0.921 (0.004)	**0.922 (0.016)**	0.921 (0.013)	0.744 (0.029)	0.818 (0.013)	0.963 (0.004)	0.849 (0.010)
`FMA-concatonly`	0.898 (0.010)	0.838 (0.026)	0.959 (0.011)	**0.849 (0.037)**	0.834 (0.016)	0.943 (0.002)	0.923 (0.003)
Other multi-view baselines							
`MV-SVM`	0.832 (0.073)	0.733 (0.147)	0.931 (0.015)	0.726 (0.058)	0.725 (0.102)	0.906 (0.055)	0.817 (0.093)
`MV-LR`	0.799 (0.057)	0.711 (0.111)	0.886 (0.038)	0.623 (0.084)	0.661 (0.085)	0.896 (0.049)	0.805 (0.084)
`MV-NN`	0.806 (0.067)	0.667 (0.129)	0.945 (0.030)	0.762 (0.108)	0.705 (0.108)	0.872 (0.067)	0.789 (0.095)
Individual feature extractors							
`ts-transformer`	0.724 (0.018)	0.556 (0.052)	0.893 (0.017)	0.581 (0.027)	0.542 (0.024)	0.864 (0.004)	0.608 (0.003)
`ts-lstm`	0.696 (0.031)	0.462 (0.076)	0.930 (0.014)	0.578 (0.069)	0.483 (0.066)	0.863 (0.001)	0.606 (0.009)
`ts-tcn`	0.759 (0.017)	0.647 (0.038)	0.871 (0.007)	0.542 (0.015)	0.587 (0.025)	0.870 (0.010)	0.643 (0.029)
`tab-cnn`	0.721 (0.016)	0.653 (0.046)	0.790 (0.018)	0.454 (0.030)	0.518 (0.019)	0.742 (0.018)	0.450 (0.023)
`tab-tabresnet`	0.623 (0.006)	0.379 (0.024)	0.866 (0.001	0.419 (0.023)	0.388 (0.011)	0.683 (0.011)	0.397 (0.019)
`tab-tabnet`	0.500 (0.003)	0.009 (0.004)	**0.991 (0.005)**	0.163 (0.105)	0.016 (0.008)	0.521 (0.023)	0.223 (0.014)
`text-temporalattn`	0.887 (0.015)	0.878 (0.039)	0.896 (0.001)	0.668 (0.016)	0.753 (0.013)	0.968 (0.002)	0.871 (0.004)
`text-huang`	0.772 (0.014)	0.653 (0.020)	0.951 (0.006)	0.594 (0.033)	0.747 (0.018)	0.929 (0.005)	0.780 (0.014)

AUROC and AUPRC by a significant margin. The effect of adding data views is evident, as combining more data views always resulted in a better modelling performance compared to combining less views. Our integration model which utilised view-wise attention also contributed to the good performance, achieving at least 2 percentage-point improvement in most metrics as compared to a simple feature concatenation model `FMA-concatonly`.

Between views, the `text-temporalattn` model achieves scores that significantly surpass all other feature extractor models for most metrics except specificity. Furthermore, view integration models that utilise the text view (`FMA-tabtext` and `FMA-tstext`) also performs better, indicating that the text view contains information that is more predictive for our problem. Lastly, it is clear that in comparison to machine learning multi-view baselines, generally the `FMA-` models significantly exceeds the former's performance.

For individual feature extractors, our proposed models achieve good performance relative to their corresponding baselines. Specifically, `tab-cnn` and `text-temporalattn` obtained top scores across all metrics, while `ts-transformer` achieves second-place in its category when we consider AUROC and AUPRC. We chose to continue with `ts-transformer` as it has the best validation performance when we first tested it.

6.2. Modelling recurrence-free survival

Our model demonstrates strong performance when predicting RFS for CRC patients. On the test dataset, our model achieves a C-index of 0.960 ± 0.01, and an Integrated Brier Score of 0.036 ± 0.01. We also show that our model accurately stratifies patients into risk-groups that are associated with the actual RFS. Figure 3 shows a Kaplan-Meier plot of

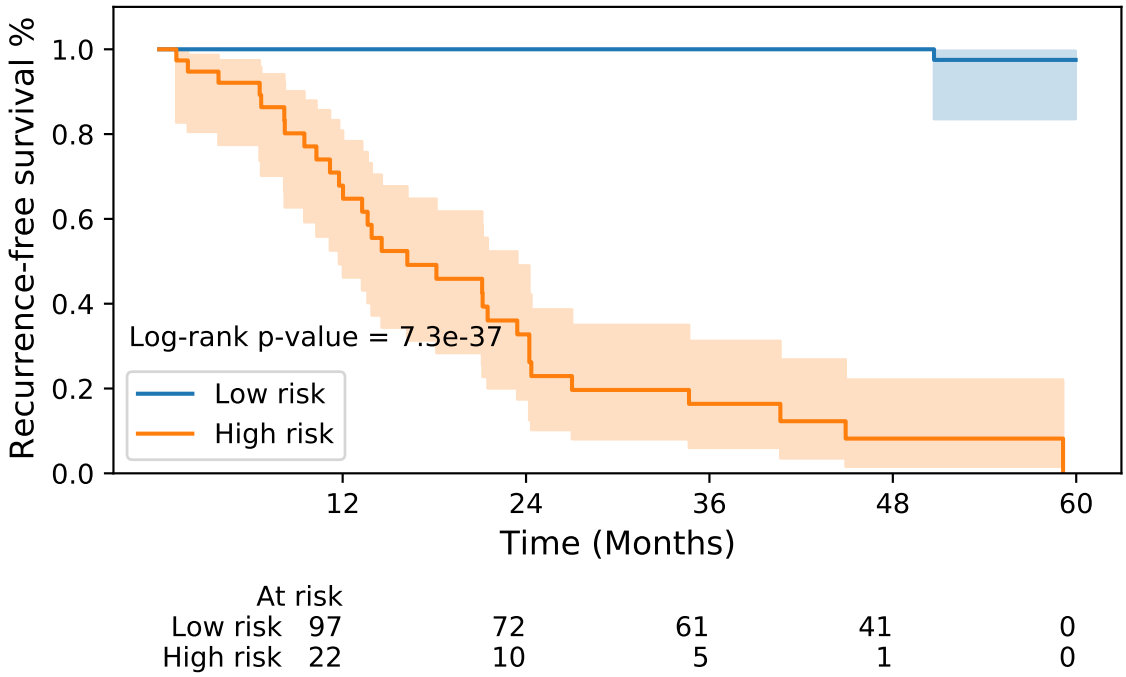

Figure 3: Recurrence-free survival plot

patients stratified by our model according to predicted risk. Our model correctly identified significant differences in survival trends between the two risk groups.

6.3. Model forecasting

The previous section highlights our model's capabilities in predicting binary outcomes to recurrence, as well as continuous outcomes for time-to-recurrence. Next we turn our attention to forecasting X months ahead, whether recurrence would occur. This perspective of modelling confers the important clinical advantage of allowing for early detection of at-risk individuals, paving the way for early intervention and prevention in the clinic. Nonetheless, developing good forecasting models with longitudinal data is a non-trivial task, as it is necessary to extrapolate future outcomes based on limited past data (where a greater forecasting horizon means less information to work with). In this section, we present preliminary work on developing a forecasting model based on just the time-series module.

To perform forecasting, we drop all raw timepoints that fall within X months to the recurrence event, where X is an integer between 1 and 6, to obtain 6 datasets. For each set of data, we perform data preprocessing and normalisation on the remaining timepoints separately, and train 10 initialisations of the `ts-transformer` model. We name each model `ts-transformer-`X`mth` and we output average scores and standard errors for our evaluation metrics.

Table 2 shows results that model performance declines from the original `ts-transformer` model as the forecasting horizon increases, with most metrics showing a general downward

Table 2: Model forecasting X months ahead, using only time-series data. Metrics are reported as average (standard error) of at least 10 model initialisations

Model	Bal Accuracy	Sensitivity	F1	PPV	Specificity	AUROC	AUPRC
ts-transformer	0.724 (0.018)	0.556 (0.052)	0.542 (0.024)	0.581 (0.027)	0.893 (0.017)	0.864 (0.004)	0.608 (0.003)
ts-transformer-1mth	0.717 (0.019)	0.580 (0.057)	0.499 (0.022)	0.481 (0.026)	0.854 (0.020)	0.836 (0.001)	0.517 (0.006)
ts-transformer-2mth	0.717 (0.019)	0.583 (0.058)	0.497 (0.023)	0.472 (0.021)	0.851 (0.020)	0.841 (0.001)	0.519 (0.005)
ts-transformer-3mth	0.721 (0.023)	0.603 (0.064)	0.490 (0.028)	0.436 (0.007)	0.839 (0.019)	0.833 (0.001)	0.453 (0.001)
ts-transformer-4mth	0.741 (0.019)	0.648 (0.054)	0.518 (0.021)	0.451 (0.012)	0.834 (0.018)	0.834 (0.001)	0.448 (0.005)
ts-transformer-5mth	0.741 (0.018)	0.655 (0.054)	0.517 (0.019)	0.449 (0.013)	0..826 (0.020)	0.834 (0.003)	0.491 (0.008)
ts-transformer-6mth	0.741 (0.019)	0.659 (0.054)	0.503 (0.017)	0.423 (0.010)	0.822 (0.018)	0.824 (0.001)	0.449 (0.009)

trend. However the decline is not as drastic as expected, as when we forecast up to 6 months in advance, we retain sensitivity while AUROC and specificity falls slightly (7 and 4 percentage points respectively). PPV and AUPRC showed the greatest performance decline with a drop of 15 percentage points, which shows that dropping timepoints had a significant impact on model precision.

7. Discussion

7.1. Clinical implications

Our study demonstrates that multi-view information combining longitudinal CEA readings, longitiduinal radiological text reports and single-timepoint clinical information is sufficient to strongly predict recurrence and time-to-recurrence in CRC patients. Our full model achieves sensitivity, specificity, AUROC and AUPRC scores of 0.90, 0.95, 0.97 and 0.95 respectively, exceeding all known models developed for the same task, including the reported performance of both CEA-alone in the clinic (sensitivity $\sim$0.5, specificity $\sim$0.8), as well as state-of-the-art CRC recurrence prognostication models (AUROC of Castellanos et al. (2017)'s best model: 0.82). We hope that our model's strong performance together with stringent baselines lend credibility towards future consideration for potential deployment in the clinic.

7.2. Technical implications

Our individual feature extractors are neural networks that we have carefully developed to give good representations for that particular data format. For example our tabular feature extractor introduces spatial structure that can be exploited by a CNN network, allowing the model to learn from less data, while our text extractor uses a combination of ClinicalBERT and temporal attention modelling to process longitudinal clinical text. We show that our models are highly competitive compared to state-of-the-art baselines on this CRC recurrence problem. We believe that our work provides a good starting point for future experiments on model generalisation to other datasets and clinical problems with similar data formats.

We also show that adding more data views, as well as dynamically combining feature representations based on the importance of each view, resulted in performance gains compared to lesser views and simple feature concatenation as an integration method. This

concurs with existing literature that having more information helps the modelling process, and view-wise attention may serve to highlight important views and avoid overfitting on the additional data. One aspect we could improve upon is to consider a modelling approach whereby instead of highlighting the entire view, one could instead highlight different parts of individual views. This approach would be analogous to the situation whereby a clinician extracts useful data from different data sources when coming to a decision. We intend to work on this as part of our future work.

Additionally we show that having good feature representations lends favourably to downstream tasks beyond simple prediction. We show that modifying the training approach from minimizing cross-entropy to minimizing the negative likelihood of CoxPH, following the use of already-extracted feature representations, give rise to strong survival models that can predict time-to-recurrence. Our work points towards the strength and versatility of deep learning for multi-view modelling, and shows support to the continued usage of neural networks for representation learning and data integration for various downstream tasks.

Lastly we show through our preliminary study the possibility of modifying our model to perform forecasting and early detection of disease. We observe that model performance indeed declines with a greater forecasting horizon if we utilise a simple drop-out approach. While the decline is not as drastic as expected, nonetheless we are interested in developing approaches to mitigate the performance decline, such as through missing data imputation or better extrapolation techniques.

Limitations Our work is limited in the following ways: 1) Ours is based on a retrospective cohort and we cannot assume that our model will generalise in the wild. This is an especially pertinent issue considering tendencies of real-world data to undergo dataset drift. 2) While our modelling approach for both feature extractors and view-integrator is based on data format and not clinical problem, we have not validated our proposed approach on other datasets. We intend to do this as part of future work. 3) Our work on developing forecasting models for early intervention is at the early stages. We are focusing our efforts in this area and intend to report upon this in the future.

Acknowledgments

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002), and by National University of Singapore and National University Health System under the Health Innovation Programme (WBS A-0009026-01-00).

References

Martın Abadi, Ashish Agarwal, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015.

Okechinyere J. Achilonu, June Fabian, et al. Predicting Colorectal Cancer Recurrence and Patient Survival Using Supervised Machine Learning Approach: A South African Population-Based Study. *Frontiers in Public Health*, 9:838, 2021.

Shotaro Akaho. A kernel method for canonical correlation analysis, 2007. arXiv:cs/0609071.

Behnaz Alafchi, Ghodratollah Roshanaei, et al. Joint modelling of colorectal cancer recurrence and death after resection using multi-state model with cured fraction. *Scientific Reports*, 11(1):1016, 2021.

Sercan O. Arik and Tomas Pfister. TabNet: Attentive Interpretable Tabular Learning. *arXiv:1908.07442 [cs, stat]*, 2020.

Vadim Borisov, Tobias Leemann, et al. Deep Neural Networks and Tabular Data: A Survey. *arXiv:2110.01889 [cs]*, 2022.

Jason Castellanos, Qi Liu, et al. Predicting colorectal cancer recurrence by utilizing multiple-view multiple-learner supervised learning. *Journal of Clinical Oncology*, 35 (4_suppl):635–635, 2017.

Maximilian Christ, Nils Braun, et al. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77, 2018.

Yanwei Fu, Yanwen Guo, et al. Multi-View Video Summarization. *Multimedia, IEEE Transactions on*, 12:717–729, 2010.

Oscar G. F. Geessink, Alexi Baidoshvili, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cellular Oncology (Dordrecht)*, 42(3):331–341, 2019.

A. L. Goldberger, L. A. Amaral, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23): E215–220, 2000.

Roger Higdon, Rachel K. Earl, et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *Omics: A Journal of Integrative Biology*, 19(4):197–208, 2015.

Danliang Ho and Mehul Motani. Machine and Deep Learning Methods for Predicting Immune Checkpoint Blockade Response. *Proceedings of the 2nd Machine Learning for Health symposium*, pages 512–529, 2022.

Danliang Ho, Iain Bee Huat Tan, et al. Predictive models for colorectal cancer recurrence using multi-modal healthcare data. *Proceedings of the Conference on Health, Inference, and Learning*, pages 204–213, 2021a. ISBN: 978-1-4503-8359-2 Place: New York, NY, USA.

Danliang Ho, Iain Bee Huat Tan, et al. Prognosticating Colorectal Cancer Recurrence using an Interpretable Deep Multi-view Network. *Proceedings of Machine Learning for Health*, pages 97–109, 2021b.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

P. Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17:331–347, 1961.

Harold Hotelling. Relations Between Two Sets of Variates. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, Springer Series in Statistics, pages 162–190. Springer, New York, NY, 1992.

Kexin Huang, Jaan Altosaar, et al. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, 2020. arXiv:1904.05342 [cs].

Javier. pytorch-widedeep, 2023. original-date: 2017-10-21T08:11:44Z.

Dan Jiang, Junhua Liao, et al. A machine learning-based prognostic predictor for stage III colon cancer. *Scientific Reports*, 10(1):10333, 2020.

Alistair E. W. Johnson, Tom J. Pollard, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.

Jared L. Katzman, Uri Shaham, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, 2017.

Adrienne Kline, Hanyin Wang, et al. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):1–14, 2022.

C. Lea, M. D. Flynn, et al. Temporal Convolutional Networks for Action Segmentation and Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012, 2017.

Garam Lee, Byungkon Kang, et al. MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework. *Frontiers in Genetics*, 10, 2019.

Gang Liu, Chuanpeng Dong, et al. Integrated Multiple "-omics" Data Reveal Subtypes of Hepatocellular Carcinoma. *PLOS ONE*, 11(11):e0165457, 2016. Publisher: Public Library of Science.

Donghuan Lu, Karteek Popuri, et al. Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Scientific Reports*, 8(1):5697, 2018.

Leila Mahmoudi, Ramezan Fallah, et al. A bayesian approach to model the underlying predictors of early recurrence and postoperative death in patients with colorectal Cancer. *BMC Medical Research Methodology*, 22(1):269, 2022.

Reetesh K. Pai, Imon Banerjee, et al. Quantitative Pathologic Analysis of Digitized Images of Colorectal Carcinoma Improves Prediction of Recurrence-Free Survival. *Gastroenterology*, 163(6):1531–1546.e8, 2022. Publisher: Elsevier.

Adam Paszke, Sam Gross, et al. Automatic differentiation in PyTorch. In *2017 Conference on Neural Information Processing Systems*, page 4, 2017.

Fabian Pedregosa, Gaël Varoquaux, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Catherine R. Planey and Olivier Gevaert. CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Genome Medicine*, 8(1):27, 2016.

Jong Pil Ryuk, Gyu-Seog Choi, et al. Predictive factors and the prognosis of recurrence of colorectal cancer within 2 years after curative resection. *Annals of Surgical Treatment and Research*, 86(3):143–151, 2014.

Malihe Safari, Hossein Mahjub, et al. Specific causes of recurrence after surgery and mortality in patients with colorectal cancer: A competing risks survival analysis. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 26:13, 2021.

Bin Shan, Weichong Yin, et al. ERNIE-ViL 2.0: Multi-view Contrastive Learning for Image-Text Pre-training, 2022. arXiv:2209.15270 [cs].

Bethany Shinkins, Brian D. Nicholson, et al. The diagnostic accuracy of a single CEA blood test in detecting colorectal cancer recurrence: Results from the FACS trial. *PLoS ONE*, 12(3), 2017.

Ravid Shwartz-Ziv and Amitai Armon. Tabular Data: Deep Learning is Not All You Need. *arXiv:2106.03253 [cs]*, 2021.

Ole-Johan Skrede, Sepp De Raedt, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks, 2017. arXiv:1506.01186 [cs].

Caspar G. Sørensen, William K. Karlsson, et al. The diagnostic accuracy of carcinoembryonic antigen to detect colorectal cancer recurrence – A systematic review. *International Journal of Surgery*, 25:134–144, 2016.

T Tieleman and G Hinton. Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude, 2012.

Wen-Chien Ting, Yen-Chiao Angel Lu, et al. Machine Learning in Prediction of Second Primary Cancer and Recurrence in Colorectal Cancer. *International Journal of Medical Sciences*, 17(3):280–291, 2020.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(1):1–67, 2011.

Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

Rongfang Wang, Jinkun Guo, et al. Locoregional recurrence prediction in head and neck cancer based on multi-modality and multi-view feature expansion. *Physics in Medicine and Biology*, 67(12), 2022.

World Health Organisation WHO. Cancer.

Yucan Xu, Lingsha Ju, et al. Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. *Scientific Reports*, 10(1):1–9, 2020.

Xiaoqiang Yan, Shizhe Hu, et al. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

Zhenpei Yang, Zhile Ren, et al. MVS2D: Efficient Multi-view Stereo via Attention-Driven 2D Convolutions, 2021. arXiv:2104.13325 [cs].

Patrick. E. Young, Craig M. Womeldorph, et al. Early Detection of Colorectal Cancer Recurrence in Patients Undergoing Surgery with Curative Intent: Current Status and Challenges. *Journal of Cancer*, 5(4):262–271, 2014.

Shunyu Zhang, Yaobo Liang, et al. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000, Dublin, Ireland, 2022. Association for Computational Linguistics.