# Predictive models for colorectal cancer recurrence using multi-modal healthcare data

Danliang Ho
National University of Singapore
Singapore
ho.danliang@u.nus.edu

Iain Bee Huat Tan
National Cancer Center Singapore
Singapore
iain.tan.b.h@singhealth.com.sg

Mehul Motani
National University of Singapore
Singapore
motani@nus.edu.sg

## ABSTRACT

Colorectal cancer recurrence is a major clinical problem - around 30-40% of patients who are treated with curative intent surgery will experience cancer relapse. Proactive prognostication is critical for early detection and treatment of recurrence. However, the common clinical approach to monitoring recurrence through testing for carcinoembryonic antigen (CEA) does not possess a strong prognostic performance. In our paper, we study a series of machine and deep learning architectures that exploit heterogeneous healthcare data to predict colorectal cancer recurrence. In particular, we demonstrate three different approaches to extract and integrate features from multiple modalities including longitudinal as well as tabular clinical data. Our best model employs a hybrid architecture that takes in multi-modal inputs and comprises: 1) a Transformer model carefully modified to extract high-quality features from time-series data, and 2) a Multi-Layered Perceptron (MLP) that learns tabular data features, followed by feature integration and classification for prediction of recurrence. It achieves an AUROC score of 0.95, as well as precision, sensitivity and specificity scores of 0.83, 0.80 and 0.96 respectively, surpassing the performance of all-known published results based on CEA, as well as most commercially available diagnostic assays. Our results could lead to better post-operative management and follow-up of colorectal cancer patients.

## CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computer systems organization** → **Neural networks**.

## KEYWORDS

prognostication, colorectal cancer, cancer recurrence, multi-modal, deep learning, representation learning, time-series

**ACM Reference Format:**
Danliang Ho, Iain Bee Huat Tan, and Mehul Motani. 2021. Predictive models for colorectal cancer recurrence using multi-modal healthcare data. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '21), April 8–10, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3450439.3451868

## 1 INTRODUCTION

*Clinical Relevance.* Colorectal cancer is among the top three most commonly diagnosed cancers globally as well as one of the leading causes of cancer-related deaths [18]. Surgery is the mainstay of colorectal cancer treatment when the cancer is caught in a localised stage; nonetheless approximately 30-40% of patients who undergo the procedure eventually develop recurrent disease [7]. Since cancer recurrence can be managed or even treated depending on the extent of disease spread and the patient's condition, timely detection of recurrence has strong clinical significance. Current recommendations for recurrence detection include routine testing for levels of carcinoembryonic antigen (CEA), a blood-based tumour marker, in post-operative patients [5, 11, 27]. CEA is a common clinical marker that has been shown to correlate with the presence and extent of colorectal cancer [23, 37], but systematic studies have reported that CEA by itself is not a strong prognostic marker for colorectal cancer surveillance, with only ~50% sensitivity and ~80% specificity [32, 35]. On the other hand, increasing the intensity of follow-up and CEA testing [30] to improve diagnostic performance adds on to the healthcare burden and diverts resources from other patients. Therefore, there is a pressing need to develop a better tool for accurate detection of colorectal cancer recurrence.

*Technical Significance.* In this work, we compare several approaches for processing information arising from heterogeneous and multi-modal data, a common occurrence in the healthcare setting. Our main contributions are three-fold:

(1) We developed prognostic models for colorectal cancer recurrence using machine learning (ML) and deep learning (DL). Our best models perform significantly better than the commonly used CEA tumour marker as well as most commercially available diagnostic assays (refer to Related Work).
(2) Our models are based on three different approaches in combining tabular clinical information and time-series laboratory values. We showed that complementary information arising from inputs from multiple modalities boosts predictive performance. As far as we know, ours is the only work that applies concepts from multi-modal modelling in the task of predicting colorectal cancer recurrence.
(3) We outlined our methods to modifying the Transformer architecture for use in time-series feature extraction and demonstrated that the modified architecture achieves significant performance gains in multi-modal modelling.

This research work was supported by the Ministry of Education, Singapore, under grant WBS R-263-000-D64-114.
We thank National Cancer Centre Singapore for providing us the datasets used in this study.

*ACM CHIL '21, April 8–10, 2021, Virtual Event, USA*
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8359-2/21/04.
https://doi.org/10.1145/3450439.3451868

# 2 RELATED WORK

## 2.1 Prognostication models in colorectal cancer

The task of prognosticating recurrence is not new and has been traditionally tackled through statistical and epidemiological means such as Cox proportional hazards modelling. Such studies utilise a small number of hand-curated prognostic factors, (e.g pre-operative CEA, tumour stage, depth of invasion, gene signatures) with the aim of stratifying patients into risk groups for optimising post-operative management [42, 43]. There is strong interest in commercialising such models, for example the ColoPrint assay [19] and the OncoDefender-CRC [24], are both commercially available clinical assays for predicting recurrence that utilise a small panel of genomic factors combined with clinicopathological variables. Nonetheless we note that the reported AUROC scores on a validation cohort were rather modest (ColoPrint: 0.626, OncoDefender-CRC: 0.55), as would be expected of models that utilise a limited number of prognostic factors.

In recent years, machine learning techniques have risen in popularity due to their ability to handle datasets with large number of variables as well as high degree of heterogeneity (such as time-series or text). As such, ML models have seen widespread application in various domains including the diagnosis and prognosis of colorectal cancer. Castellanos et al. [4] trained an ensemble model to predict recurrence in Stage II-III colorectal cancer patients, on a dataset containing a large body of information including gene expression, protein-protein interaction, tumour suppressor and driver mutations. They demonstrated that their model predicted better on molecular data than on clinical data alone and their best model achieved an AUROC of 0.786, higher than that of the commercial clinical assays. Xu et al. [41] also used ML algorithms for predicting recurrence, differences included the choice of the algorithm as well as the patient cohort, which consisted of only Stage IV colorectal cancer patients. They reported an AUROC score of 0.761 for their best model.

As of our knowledge, there is very little research on deep learning models for prognosticating recurrence in colorectal cancer. Existing research focused on using convolutional neural networks (CNNs) to extract features from tumour histological sections for prediction. For example, Skrede et al. [34] trained 10 CNNs on more than 12,000,000 images of stained tissue sections to develop a prognostic biomarker that stratified early-stage colorectal patients based on poor versus good outcomes. Jiang et al. [15] also trained multiple CNNs on tissue slides, notably they used InceptionResNet to perform recognition of specific categories such as stroma, mucosa and tumour, and fed the features into a GradientBoosting classifier for the prediction task. Both studies only reported hazard ratios and not AUROC scores thus making comparisons difficult.

We note that the vast majority of these studies used data from a single modality, such as tabular data including clinical and molecular data, or image data for CNN. However, integrating data from multiple sources could provide complementary information not present in a single domain of data, to ultimately improve model performance. As such, in the next section we outline the work that has been performed on creating multi-modal networks.

## 2.2 Multi-modal learning in healthcare

Healthcare data is highly heterogeneous and consists of vast repositories of data from multiple sources and modalities, including laboratory time-series data, high-dimensional genomics datasets, unstructured text data from electronic health records and imaging data from scans or histological sections, in addition to the usual clinical data. Considering the richness of healthcare big data, there are obvious benefits to developing models that consider information from different and complementary sources for more informed decision making. Nonetheless multi-modal data processing and modelling is challenging for standard machine learning networks, as decisions have to be made on how to best integrate the data, for example through data transformations or feature extraction into a compatible structure.

The advent of deep neural networks created inroads into realising the potential of multi-modal learning applied to healthcare problems. This is due to their ability to inherently model underlying data distributions without prior assumption, as well as extract relevant features from high-dimensional data, thus solving the problem of data integration. Nonetheless, much of the existing literature in this area focused on the single aspect of combining image analysis with tabular data such as clinical or genomics datasets [36]. This is perhaps understandable as healthcare has seen much success with using deep learning, often with large pre-trained networks, for medical image processing and image analysis [10, 26, 31].

In comparison, as of our knowledge there are fewer works that focus on temporal data when dealing with multi-modal datasets. This is surprising as much of available healthcare data exists as time-series, for example data collected from laboratory measurements and clinical equipment for the purpose of monitoring patients over a period of time [16, 29]. These studies are mostly situated in the problem of predicting clinical outcomes using structured and temporal information available within EHRs, and the general consensus is that incorporating longitudinal data contributed to better performance than models that utilised only structured information [9, 44]. Models that utilised the full temporal information also outperformed models that aggregated the same temporal features as inputs [33, 44]. In addition, Ghassemi et al. [9] showed that in the task of mortality prediction within an ICU setting, as the prediction horizon (length of time until event) increased, the baseline tabular features became less predictive of mortality, but the longitudinal data were able to maintain or even improve the accuracy of prediction. All in all, longitudinal data adds utility to the prediction problem by conferring additional information not available in the static tabular data. Furthermore it is more likely than static data to be predictive of diseases with a long time-course to onset, notably chronic diseases such as Alzheimer's Disease [8], as well as cancer recurrence, as fluctuation patterns within longitudinal data could act as a more accurate reflection of the underlying physiological state of an individual.

Nonetheless, it is a non-trivial problem to jointly model both static and temporal information simultaneously, especially with standard machine learning techniques. Most studies utilise some form of aggregation of temporal features by time-bins, to effectively combine with structured data. For example Ghassemi et al. extracted features from cumulative 12-hour bins of clinical notes, while Zhao et al. computed statistics of each measurement in yearly bins. There is some recent work that exploited the power of deep-learning sequential modelling to do away with manual feature aggregation: a) Lee et al.'s work [22] where they extracted feature representations from longitudinal data using Gated Recurrent Units (GRUs) and integrated the resultant feature vectors with a logistic regression classifier. They showed that their approach worked on both simulated data as well as real-world clinical data, and b) El-Sappagh et al.'s work [8] in extracting time-series features using bidirectional Long Short-Term Memory (LSTM) networks, combined with features extracted from images, in the task of predicting Alzheimer's Disease progression.

Notwithstanding the success of such approaches, we note that there are several methods to performing multi-modal learning and these papers have only proposed a single architecture. Furthermore, LSTMs and GRUs are standard architectures for sequential data learning and have since been superseded by better-performing architectures including temporal CNNs (TCNs) [21] and attention architectures [39]. In our paper, we present and compare several different approaches to multi-modal learning, including the use of state-of-the-art architectures for longitudinal data processing.

# 3 DATASETS AND PRE-PROCESSING

## 3.1 Study Cohort

Our study cohort consisted of 882 patients diagnosed with Stage I-III colorectal cancer, referred to National Cancer Center Singapore for post-operative management, following surgical resection of the primary tumour. Written consent was obtained from all patients. Institutional ethics approval was obtained for this study.

## 3.2 Dataset Description

The dataset consisted of the following clinical information collected for each patient: a) tabular data on 65 variables that are potentially prognostic for recurrence, including demographic information, tumour characteristics, laboratory test results and treatment parameters, as well as b) time-series data on measured levels of CEA. Patients were followed-up post-operation with a frequency between 1 to 3 months on average, for a median follow-up duration of 40 months. (See Table 1 for details on dataset characteristics). All collected data were de-identified; each patient was assigned a unique serial number upon study entry, and all personal identifiers were removed prior to data analysis.

**Table 1: Dataset characteristics**

| | Category | Numbers | Proportion |
|---|---|---|---|
| *Demographic factors* | | | |
| Patient number | | 882 | |
| | 60.48 | | |
| Gender | Male | 540 | 0.61 |
| | Female | 342 | 0.39 |
| Race | Chinese | 744 | 0.84 |
| | Malay | 71 | 0.08 |
| | Indian | 24 | 0.03 |
| | Others | 43 | 0.05 |
| *Tumour characteristics* | | | |
| Location | Sigmoid | 246 | 0.28 |
| | Rectum | 247 | 0.28 |
| | Rectosigmoid | 129 | 0.15 |
| | Others | 260 | 0.30 |
| TNM Stage | I | 35 | 0.04 |
| | II | 217 | 0.25 |
| | III | 621 | 0.70 |
| *Laboratory typing* | | | |
| Molecular | KRAS + | 147 | 0.17 |
| | BRAF + | 18 | 0.02 |
| | NRAS + | 13 | 0.01 |
| | P53 + | 72 | 0.08 |
| MSS Status | MSI | 38 | 0.04 |
| | MSS | 569 | 0.65 |
| | Unknown | 275 | 0.31 |
| *Treatment parameters* | | | |
| Surgery | Laparoscopic | 304 | 0.34 |
| | Not available | 288 | 0.33 |
| | Open | 192 | 0.21 |
| | Others | 98 | 0.11 |
| Therapy | Xelox | 390 | 0.44 |
| | Xeloda | 173 | 0.20 |
| | Not available | 148 | 0.17 |
| | Not given | 92 | 0.10 |
| | Folfox | 23 | 0.03 |
| | Others | 56 | 0.06 |
| *CEA time-series data* | | | |
| Median follow-up (months) | | 40 | |
| Median datapoints per patient | | 14 | |
| Outcomes | Recurrence | 202 | 0.23 |
| | No Recurrence | 680 | 0.77 |

## 3.3 Pre-processing Steps

*3.3.1 Tabular data.* We scrubbed the dataset to remove major errors and inconsistencies attributed to misspellings, letter case, extra white space, and categories that were semantically similar. We imputed missing data according to the following criteria:

- Missing dates (day or month, never year) – Imputed with '01' or January respectively
- MNAR (Missing-Not-At-Random) data – Likely MNAR data were identified using domain knowledge (e.g. *"grade_of_synchronous_tumour"* was more likely to be left blank when synchronous tumour was absent and filled when present). These data were subsequently imputed with a new category *'not_available'*.
- All other missing data – Imputed with multiple rounds of imputation using MICE [3]

We processed free-text descriptions in our dataset using rule-based matching to extract relevant features. For example, chemotherapy details were coded as free-text and information on therapy, dosage and additives had to be extracted manually. In a few cases, we split features to reduce the number of categories. For example, the tumour location field ("right descending colon") could be split into general location ("descending colon") and relative position ("right").

*3.3.2 Time-series data.* We included only CEA measurements with timestamps that occurred after surgery, and prior to a confirmed recurrence diagnosis. We scrubbed the dataset for major errors such as non-numerical CEA values and duplicated data, as well as linear interpolation for missing data imputation.

In Approach 1 we extracted tabular features from the CEA time-series prior to feeding into the models (see Section 5.1 for more details). Models developed in all other approaches are able to account for time-dependencies from the time-series dataset directly. For these approaches, we performed month-wise re-sampling, so as to create evenly spaced time intervals, followed by zero-padding to account for variable lengths within the time-series.

*3.3.3 Data transformations and scaling.* We transformed all numerical data via logarithmic transformation to account for long-tailed distribution within the dataset features, followed by min-max scaling. All categorical data was transformed using one-hot encoding, to obtain our final dataset with 548 features. Prior to feeding data into our models, the dataset was split into 60% for training, 20% for validation, and 20% as a held-out test dataset, using stratified sampling so that all classes were proportionately represented.

# 4 EXPERIMENTAL SETUP

## 4.1 Overview

We explored the following approaches to tackle the problem of multi-modality:

- Approach 1: Shallow ML models that use pre-extracted time-series features combined with tabular data as model input
- Approach 2a: Deep-learning hybrid models that perform automated feature extraction and data integration
- Approach 2b: Modification of the Transformer architecture to adapt to time-series data and boost the performance of our multi-modal architecture
- Approach 3: Unsupervised approach to learn latent representations from multi-modal data using autoencoders

Methods and results of each approach are described subsequently.

## 4.2 Model Evaluation

All developed models were tuned for best hyperparameters on a validation dataset and evaluated on a separate, held-out set. We obtained generalisation performance through the following method:

(1) We trained 20 separate models for each model architecture. Models differ only in terms of initial weights and are similar in all other aspects.
(2) We created 5 sets of simulated data by bootstrap resampling of the held-out set. Each bootstrap sample was set to 100 patients from the held-out set, with replacement.
(3) We computed sample statistics over the 20 models each with 5 bootstrap samples to obtain 100 estimates and reported the average.

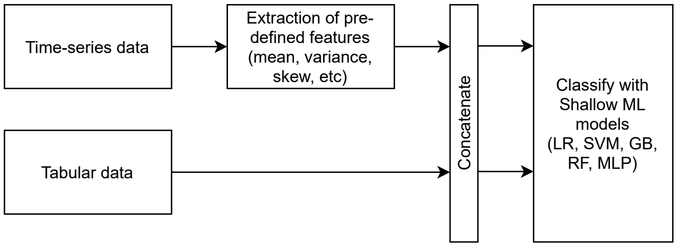
Time-series data
Extraction of pre-defined features (mean, variance, skew, etc)
Tabular data
Concatenate
Classify with Shallow ML models (LR, SVM, GB, RF, MLP)

**Figure 1: Outline of our approach when modelling with shallow ML**

We reported the following performance metrics: Precision, Recall (or sensitivity), Balanced Accuracy, Specificity, and Area Under Receiver Operating Characteristic (AUROC). Balanced accuracy is the average accuracy across all classes and, in the two-class scenario, can be computed as follows:

$$\text{Balanced Accuracy} = (\text{Sensitivity} + \text{Specificity})/2$$

Balanced accuracy accounts for both classes separately and is thus a fairer measure of performance when dealing with imbalanced datasets such as ours. Furthermore, balanced accuracy encapsulates specificity and sensitivity, two important and commonly used measures of a clinical assay's performance. As such, while we still prefer models with high AUROC scores as aligned with standard machine learning literature, we also consider models with high balanced accuracy scores favourably.

# 5 APPROACH 1: MODELLING WITH PRE-EXTRACTED TIME-SERIES FEATURES

## 5.1 Methods

We used the Python package *tsfresh* [6] to extract relevant and meaningful features time-series data using known characterization methods (e.g. mean, variance, coefficients from Fourier and Wavelet transformations). We selected only features that were statistically significant accounting for multiple hypothesis testing. The processed features were directly concatenated with tabular data and fed into the models as a single dataset. A simple schematic of our approach is shown in Figure 1.

We investigated five different shallow ML classifiers that encompassed different flavours of ML models. Parameters were selected through extensive grid-search on a validation dataset.

- logistic regression (LR), with parameters C=0.1 and l2 penalty
- support vector machine (SVM) with radial basis function (RBF) kernel and parameters C=1 and gamma="scale"
- gradient boosting (GB) with learning rate set to 0.2, max depth=2, max features=285, min samples at leaf node=8, number of estimators=45
- random forest (RF) with criterion='entropy', max depth=4, max features=150, number of estimators=46
- a simple multi-layer perceptron (MLP) with two dense layers with 70 and 10 nodes, relu activation, dropout set at 0.3 and 0.15 respectively, and Adadelta optimiser

Our choice of models was based on selecting a variety of ML approaches. LR is a linear model, SVM with RBF is a non-linear model,

**Table 2: Results of shallow ML models on pre-extracted time-series combined with tabular inputs.**
**P: Precision, R: Recall, Spec: Specificity, Bal Acc: Balanced Accuracy, AUC: AUROC.**

| Models | P | R | Spec | Bal Acc | AUC |
|---|---|---|---|---|---|
| *Baselines* | | | | | |
| SVM-tsfresh | 0.508 | 0.606 | 0.856 | 0.731 | 0.821 |
| SVM-tabular | 0.482 | 0.743 | 0.803 | 0.773 | 0.779 |
| *Combined data* | | | | | |
| LR | 0.622 | 0.687 | 0.895 | 0.791 | 0.886 |
| SVM | **0.714** | 0.696 | **0.933** | 0.814 | **0.895** |
| GB | 0.696 | 0.681 | 0.927 | 0.804 | 0.881 |
| RF | 0.522 | **0.814** | 0.818 | **0.816** | 0.886 |
| MLP | 0.514 | 0.803 | 0.816 | 0.809 | 0.882 |

GB and RF are ensemble learning models using tree-based classifiers, and lastly MLP is a bilayer neural network.

For baseline comparisons, we ran SVM models with either the pre-extracted time-series features (SVM-tsfresh) or the structured tabular data (SVM-tabular).

All ML models were run using *sklearn* [28] and trained on class weights that were inversely proportional to the class frequencies, to account for imbalanced learning.

## 5.2 Results

Our results shown in Table 2 confirmed that incorporating data from both the time-series and tabular modalities resulted in better model performance than from a single modality, with at least a 6-percentage point increment in AUROC scores for models that utilised combined data as compared to the baselines. SVM trained on combined data achieved the best performance, with an AUROC score of 0.895 and decent recall and specificity of 0.696 and 0.933 respectively. We note that RF also performed well, topping all models in terms of recall and balanced accuracy, but we favour SVM for overall good performance across all metrics.

While we achieved good results with the current approach, these models were constrained by the need to perform manual feature extraction prior to modelling. Next we exploited the ability of deep neural networks to perform automated feature learning.

# 6 APPROACH 2A: AUTOMATED LEARNING WITH DEEP-LEARNING HYBRID ARCHITECTURES

## 6.1 Methods

We created neural networks with hybrid architectures to process and integrate data from multiple modalities. A high-level overview of the proposed network design is shown in Fig. 2. Stage 1 processes time-series data to output temporal features, and Stage 2 combines the tabular and temporal features to make the final prediction. Specifically, the models were trained as follows:

- Stage 1: Modelling of sequential data to extract features from time-series. We explored the performance of 3 different models: LSTM, temporal CNN (TCN) and a Transformer-based model for this task. We trained each network on binary prediction of recurrence, and output the learnt temporal feature representations in the form of a fixed-size vector from the last fully-connected layer (FCN).
- Stage 2: Feature integration and classification. We created a separate Multi-layered Perceptron (MLP) that accepted as input the feature representations from Stage 1, as well as the tabular data. The network performs two functions: 1) extract features from the tabular inputs, and 2) concatenate features from each modality and learn integrative feature representations to make the final prediction.

All deep-learning models were run with Tensorflow [1]. Unless otherwise specified, all deep learning models were trained to minimise binary crossentropy loss, and utilised Adam optimiser with an initial learning rate set to 1e-4, which decayed at a factor of 0.1 when validation loss failed to improve after 8 epochs. All models were trained for at least 100 epochs, halting training early when no improvement was observed on the validation loss after 20 epochs.

### 6.1.1 *Long Short-Term Memory (LSTM) network.*

We employed a stacked LSTM [13] network (16 node bidirectional layer followed by 8 node unidirectional layer), with activation function set to tanh, dropout and recurrent dropout rate set to 0.2. Zero-padded timesteps were skipped during model training and evaluation by virtue of a masking layer implemented on top of the LSTM. Initial learning rate was set to 1e-3. To coerce the model to learn the minority class better on an imbalanced dataset, we trained the network using balanced class weights.

### 6.1.2 *Temporal Convolutional Network (TCN).*

Referencing architecture first described by Lea et al. [21], we employed temporal 1D-CNNs that apply convolutions across the temporal domain. To increase the size of the receptive field we stacked 6 convolutional blocks, using a dilation rate that increases at a factor of 2 at each consecutive layer (l1 is 1, l2 is 2, l3 is 4… and so on). Each convolutional block consists of 2 layers of alternating 1D-convolutional and dropout layer as well as a residual connection that combines each block's input and output signal. We employed causal padding to maintain constant dimensions after each convolution operation. Other model specifications are as follows: filter size 2, hidden layer size 60, relu activations for all layers except for classification, He uniform weight initialisation [12], and dropout rate 0.1.

### 6.1.3 *Transformer-based architecture.*

Our model is based on the work by Vaswani et al. [39] and consists of:

- Positional encoding of the time-series inputs to indicate the position of each time-step in the sequence. This is done as follows:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

where $pos$ is the timepoint, $i$ is the data dimension, and $d_{model}$ is the hidden layer dimension. The positional information is combined with input data through direct summation, followed by linear projection through a Dense layer, to act as input into the encoder.

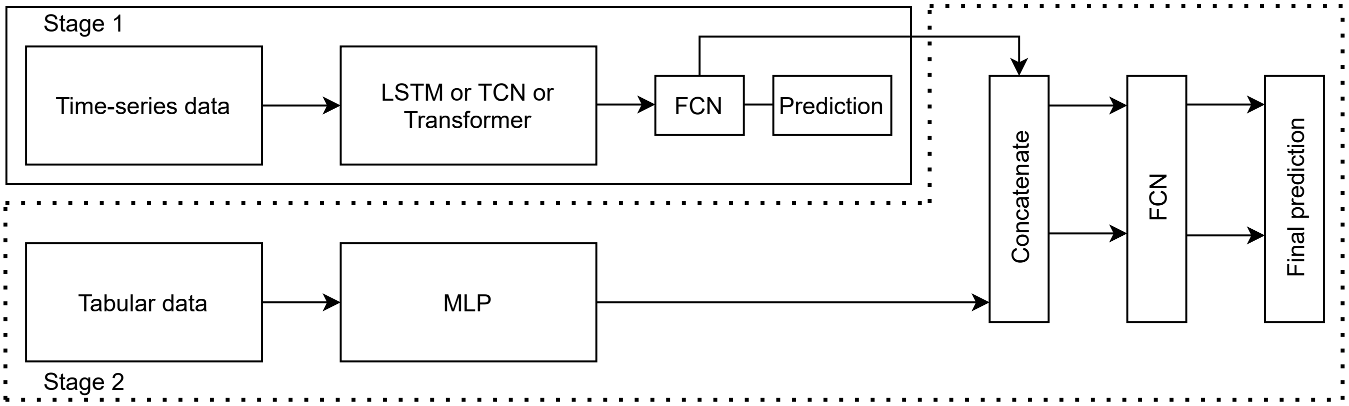

**Figure 2: Schematic of the network design and training for our deep-learning hybrid architecture**

- Scaled dot-product self-attention to model time-series dependencies. This is calculated as:

$$attention(Q, K, V) = \text{softmax}\left(\frac{QWK^T W}{\sqrt{d_k}}\right) VW$$

  where $W$ is a linear projection and $Q$, $K$, $V$ represent the query, key and value matrices respectively. We calculate self-attention using multiple heads, which simply refers to performing the self-attention calculation multiple times (each time being a single head), followed by concatenating all the output heads to obtain the final output.
- An encoder block comprising sequentially stacked layers of multi-headed self-attention, dropout, layer normalisation, position-wise feed-forward network, and residual connections that combine the block's input and output signal.
- A decoder block that takes in the output of the encoder. It differs from the encoder in only two aspects. The first is that we use an attention mask to mask out future inputs to preserve the auto-regressive property of the model and ensure that the decoder can only see past tokens. The second is that in addition to self-attention on the decoder outputs, the model also performs encoder-decoder attention, where the query vector came from the previous decoder layer, while the key and value vectors are created from the encoder output.
- An average pooling layer to reduce the output dimension, followed by a sigmoid classifier to make the final decision.

Except for self-attention calculations which employ linear mapping, as well as the classifier layer, all other layers utilise relu activation. Weights are initialised with He uniform distribution, and dropout rate is set to 0.1.

*6.1.4 Feature integration and classification.* We trained a separate neural network to process the tabular inputs, as well as perform feature integration and classification in an end-to-end fashion. The network receives both the temporal features from sequential models described previously, as well as the tabular inputs. Tabular data features are extracted using a Dense layer with 50 hidden nodes, dropout rate set to 0.3, relu activation and He uniform weight initialisation. The network performs a direct concatenation of the resultant fixed-vector output with the temporal features. The integrated high-dimensional features are linearly transformed using a fully-connected network, and subsequently fed into a sigmoid classification layer to perform the final prediction.

**Table 3: Comparing neural network hybrid architectures with shallow ML models. Baselines consist of a) stacked LSTM trained on only time-series features (`LSTM-ts`), b) SVM trained on extracted and processed time-series features using *tsfresh* (`SVM-tsfresh`). Nomenclature for hybrid architectures consist of: <time-series extractor>-<tabular extractor> where we differed only the time-series extractor.**
**P: Precision, R: Recall, Spec: Specificity, Bal Acc: Balanced Accuracy, AUC: AUROC.**

| Models | P | R | Spec | Bal Acc | AUC |
|---|---|---|---|---|---|
| *Baselines* | | | | | |
| `LSTM-ts` | 0.639 | 0.659 | 0.909 | 0.784 | 0.873 |
| `SVM-tsfresh` | 0.508 | 0.606 | 0.856 | 0.731 | 0.821 |
| *Hybrid architectures* | | | | | |
| `LSTM-mlp` | 0.493 | **0.858** | 0.78 | 0.819 | 0.884 |
| `TCN-mlp` | **0.853** | 0.786 | **0.966** | **0.876** | 0.913 |
| `Transformer-mlp` | 0.815 | 0.765 | 0.956 | 0.861 | **0.916** |

## 6.2 Results

Table 3 shows that deep-learning architectures generally outperformed shallow ML models. This can be seen in the comparisons between the baselines (`LSTM-ts` achieved 5-percentage point increment in AUC compared to `SVM-tsfresh`) as well as comparing between multi-modal models. In particular, both our top-ranking models, `TCN-mlp` and `Transformer-mlp`, achieved significant performance gains in all evaluation metrics across the board as compared to `SVM`, our best performer from Approach 1 (see Table 2).

Not surprisingly, `LSTM-mlp` was the worst performer among all hybrid architectures, with scores that did not even surpass `SVM` except for recall and balanced accuracy. In fact, `LSTM-mlp` seemed to have traded off specificity for better recall scores, likely attributed to the imbalanced learning setup where the model was penalised too little for predicting the majority class wrongly. We observed that training the model without class weights made the model overfit to the majority class resulting in an overall worse performance than when class weights were added. While it was possible that the performance of LSTM-mlp could be improved through tuning class weights, we note that it was still not likely to surpass that of TCN-mlp and Transformer-mlp. Since the models differed only in the architecture for time-series processing, it was likely that the TCN and Transformer architecture extracted higher quality temporal features that had a direct impact on the final performance of the hybrid model. We also refer to the substantial body of work that show both TCNs and Transformers outperform recurrent neural networks across a diverse range of tasks and datasets [2, 17, 20].

Given the fame of the Transformer architecture, we were surprised that Transformer-mlp was outperformed by TCN-mlp in almost all metrics except the AUROC. We wondered whether it was because the Transformer was originally created for text processing, and not for time-series. In our next experiment we explore modifications to the Transformer architecture to adapt to the specific demands of time-series modelling.

## 7 APPROACH 2B: MODIFICATIONS TO TRANSFORMER ARCHITECTURE

### 7.1 Methods

Originally developed for natural language processing, the Transformer architecture captures long-term dependencies in sequential text data through dot-product self-attention mechanisms that highlight important pair-wise relationships between words. However, by applying canonical self-attention directly to time-series data, the model captures important relationships between long-distances but may be less sensitive to the local context in surrounding time-points. Furthermore, the influence of long-range dependencies in time-series data is likely to be less significant compared to the surrounding context. We wondered whether these reasons affected Transformer-mlp's performance in Approach 2a and therefore explored the following modifications to our Transformer implementation described in Section 6.1.3:

(1) Convolutional self-attention: when creating the attention matrices, instead of creating query, key and value vectors out of Dense layers, we employed 1D-CNNs that convolves across the temporal dimension.
(2) Localised attention: we applied a local mask in the decoder that masks out future timesteps and also additionally limits the amount of backward attention, restricting the decoder to only focus on short-term patterns

We trained our model with a similar approach described in Approach 2a on multi-modal inputs. We also performed an ablation study to elucidate the effects of each mechanism.

We note that both convolutional self-attention and local attention, as part of a proposed sparse attention approach for reducing memory requirements, have been described in [25].

**Table 4: Ablation study on how modifications to the Transformer architecture affect multi-modal modelling. ConvSA: Convolutional self-attention, LA: Local attention.**
**P: Precision, R: Recall, Spec: Specificity, Bal Acc: Balanced Accuracy, AUC: AUROC.**

| ConvSA | LA | P | R | Spec | Bal Acc | AUC |
|---|---|---|---|---|---|---|
| × | × | 0.815 | 0.765 | 0.956 | 0.861 | 0.916 |
| ✓ | × | 0.825 | 0.764 | 0.96 | 0.862 | 0.927 |
| × | ✓ | 0.767 | 0.726 | 0.944 | 0.835 | 0.895 |
| ✓ | ✓ | 0.828 | **0.797** | 0.959 | **0.878** | **0.946** |

### 7.2 Results

Results of our ablation study are presented in Table 4. We observe that the convolutional self-attention mechanism had a beneficial effect on model performance, irrespective of whether local attention was applied or not. In fact, the combination of convolutional self-attention and local attention contributed to the best performance we have seen so far, surpassing not only the unmodified Transformer, but also TCN-mlp from Approach 2a (see Table 3) across all metrics. On the other hand, the effect of local attention was variable: it had an additive effect when combined with convolutional self-attention, but a detrimental effect when used on the original dense vectors. Our results are in line with that published in [25] in that convolutional self-attention consistently outperforms canonical self-attention on time-series datasets, while limiting the attention context (be it through sparse attention as proposed in [25], or through local attention in our work), may not always improve the model's performance. While more experiments are required to confirm the effect of local attention, we postulate that local attention could highlight short-term dependencies captured by the convolutional self-attention vectors while ignoring distant noise, thus resulting in better performance.

## 8 APPROACH 3: UNSUPERVISED FEATURE LEARNING WITH AUTOENCODERS

### 8.1 Methods

In our last approach, we performed representational feature learning via denoising autoencoders and used the learnt features for classification. A high-level schematic of how we implemented our framework is shown in Figure 3. Our denoising autoencoder consists of the following components:

- Perturbation of the input data $X$ with gaussian noise to obtain a partially corrupted version $\tilde{x}$
- An encoder unit which consists of stacked hidden layers with progressively reduced dimensionality. It learns a compressed latent representation $Y$ through deterministic mapping:

$$Y = g(W\tilde{x} + b)$$

  where $W$ and $b$ are the weights and biases respectively and are learnable parameters during training, and $g(\cdot)$ is a non-linear transformation such as relu or tanh.
- A decoder unit, which mirrors the encoder structure with layers stacked in progressively increasing dimensions. It

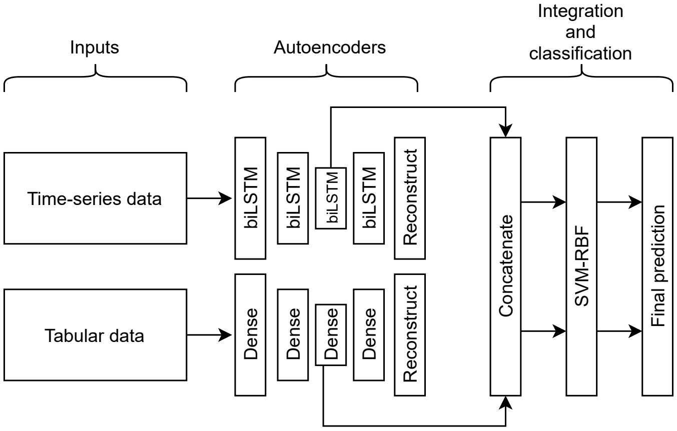

**Figure 3: Schematic of how we used autoencoders for unsupervised feature learning**

takes in the representation $Y$ to create a reconstructed signal $\hat{X}$, calculated as follows:

$$\hat{X} = g(W'Y + b')$$

where again $W'$ and $b'$ are the set of weights and biases learnt by the decoder.

During the training process, the algorithm searches for the set of parameters that optimally reconstruct the original data $X$ through minimising the reconstruction loss $L(\tilde{x}, \hat{X})$. In doing so the network is encouraged to learn a strong and compact representation $Y$ that best captures the main axes of variation in $X$. The denoising function of the autoencoder serves to not only prevent the network from simply learning an identity mapping between the input and output data, but also learn more robust features leading to better classification performance.

We trained two separate denoising autoencoders to learn features from the time-series and structured data respectively:

- Time-series autoencoder consists of stacked bidirectional LSTMs with tanh activation. The last layer of the decoder is a time-distributed Dense layer that serves to reconstruct the signal.
- Tabular autoencoder consists of stacked Dense layers that utilises relu activation and dropout to prevent overfitting.

Both models were optimized by gradient descent with momentum via RMSprop with an initial learning rate of 1e-3, and trained to minimise reconstruction error through mean-squared error loss function. Similar to the deep learning architectures described above, we trained each model for at least 100 epochs, employed early stopping based on no improvement in validation loss, and learning rate decay.

For our classification task, we extracted the learnt encodings $Y$ from each autoencoder in fixed-vector format, and combined them through direct concatenation. The integrated representations were fed into an SVM with a radial basis kernel for training and evaluation using methodology aligned with the approaches described in this paper. We refer to this integrated multi-modal autoencoder as `AE-mm`.

**Table 5: Comparing feature representation learning using autoencoders with our hybrid architectures. `AE-mm`: multimodal autoencoder that extract and integrate features from time-series and tabular data. Subsequent models are best models from Approach 2.**
**P: Precision, R: Recall, Spec: Specificity, Bal Acc: Balanced Accuracy, AUC: AUROC.**

| Models | P | R | Spec | Bal Acc | AUC |
|---|---|---|---|---|---|
| `AE-mm` | 0.600 | 0.845 | 0.866 | 0.855 | 0.915 |
| `LSTM-mlp` | 0.493 | **0.858** | 0.78 | 0.819 | 0.884 |
| `TCN-mlp` | **0.853** | 0.786 | **0.966** | 0.876 | 0.913 |
| `Transformer-mlp-mod` | 0.828 | 0.797 | 0.959 | **0.878** | **0.946** |

## 8.2 Results

Although `AE-mm` did not surpass the performance of our modified Transformer architecture with convolutional self-attention and local masking, it performed well for all evaluation metrics, ranking second in terms of AUROC and third in terms of balanced accuracy. We found it interesting that `AE-mm`'s performance was significantly better than `LSTM-mlp` with a 3.1 percentage-point difference, as both used stacks of bidirectional LSTMs and Dense layers but in different combinations. We believe this was most likely a result of the difference in training approaches. Through learning compact and strong feature representations, autoencoders could possibly filter out noisy and anomalous data which would have otherwise affected the classification boundary and hence the model's performance. Furthermore, `AE-mm` has the potential to be modified to account for missing modality data, and we will discuss this point later.

# 9 REFLECTIONS

*Clinical Implications.* Our paper demonstrates that longitudinal CEA tumour marker readings combined with routinely collected clinical information such as demographic data, tumour staging and treatment parameters strongly predict recurrence in colorectal cancer patients, with our best model achieving sensitivity, specificity and AUROC scores of ~0.8, ~0.96 and ~0.95. We note that this exceeds the reported performance of both CEA-alone in the clinic (with sensitivity ~0.5, specificity ~0.7), several commercially available diagnostic assays such as the ColoPrint and the OncoDefender-CRC (with AUROC scores of 0.63 and 0.55 respectively), as well as recent research works that utilise large genomics, proteomics or image datasets [4, 15, 34]. While the results are not directly comparable as the datasets used are different, the strong performance of our models is heartening. We believe that this highlights the advantage of leveraging on deep neural networks that integrate complementary information from multiple data sources to solve clinical problems.

*Model Complexity.* We observed that deep neural network architectures generally outperform shallow ML models that utilise pre-extracted features. There are also performance gains associated with increasing model complexity within the various neural network architectures, e.g., as we move from LSTMs to TCNs to Transformers. This is likely attributed to higher quality features learnt by the deep neural networks combined with the ability to perform more complex modelling, which directly contributed to the model's performance. Nonetheless, model complexity cannot explain the performance gains associated with the modified Transformer as the number of parameters does not significantly increase with addition of convolutional layers and local masking.

*Comparing Hybrid and Autoencoder models.* Between the two deep learning methods, while we note that the hybrid model achieved generally better results compared to the autoencoder approach in our dataset, it is difficult to recommend one over the other as each possesses distinct advantages. We have shown that in the hybrid model, the quality of features extracted from each modality can be enhanced by modifying or substituting the corresponding architecture, such as substituting LSTMs with Transformer-based approaches, or even by modifying Transformer for further performance gains. Furthermore, the network can be easily extended to include other data modalities through leveraging on large transfer networks, for example BERT and ResNet for text and images respectively. However its main limitation lies in its inability to account for and model with missing data within a particular modality. Missing entries in a modality typically occur in a large contiguous block, in other words, the entire set of data in a modality is not collected for one or more individuals. Because it does not fulfil assumptions for missing-at-randomness, the data cannot simply be imputed via traditional imputation approaches.

In contrast, multi-modal autoencoders that learn feature representations of available data from each modality have been successfully applied to the task of learning with missing modalities [14, 38, 40]. For example, Jaques et al. [14] described an approach where, inspired by denoising autoencoders, they predicted reasonable values for missing modality data from sensors by training a network to recover the values of data that was purposefully masked out, using available information from other modalities. Their approach differs from ours in that they use a single autoencoder that receives a combined input from all modalities, whereas ours is more reminiscent of the real-world situation whereby multi-modal inputs cannot be easily combined due to different data formats and it is necessary for feature fusion to occur within the model itself. We are interested in further developing this idea to create models that both work with heterogeneous data formats as well as consider missing modality information, perhaps by marrying these two approaches.

*Study limitations.* Our work is limited in the following areas. We note that ours is a predictive model based on retrospective patient data and thus there is no guarantee of model performance on a prospective cohort. It will also increase clinical relevance if our model also possesses the ability to proactively forecast recurrence months before onset. We note that similar to existing multi-modal architectures, our model lacks interpretability; quantifying feature contributions at the individual and modality level may improve the trustworthiness of the model by shedding light on whether the model is outputting predictions based on reasonable explanations. Lastly, regarding measure of generalisation performance, due to time constraints we could only perform 5 bootstrap evaluations; we note that 50-200 bootstrap samples are recommended for obtaining reliable estimates of uncertainty and hope to do so given more time.

*Data and code availability.* Due to issues regarding data sensitivity and medical confidentiality, the dataset used in this work cannot be released. However in the spirit of method reproducibility, we intend to release a Jupyter Notebook documenting code for all our models, as well as demonstrate how to run the code on a synthetic dataset, both available at: https://github.com/phu5ion/colorectal-multimodal.

## REFERENCES

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271 [cs]* (April 2018). http://arxiv.org/abs/1803.01271 arXiv: 1803.01271.

[3] Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45, 1 (Dec. 2011), 1–67. https://doi.org/10.18637/jss.v045.i03 Number: 1.

[4] Jason Castellanos, Qi Liu, R. Daniel Beauchamp, and Bing Zhang. 2017. Predicting colorectal cancer recurrence by utilizing multiple-view multiple-learner supervised learning. *Journal of Clinical Oncology* 35, 4_suppl (Feb. 2017), 635–635. https://doi.org/10.1200/JCO.2017.35.4_suppl.635 Publisher: Wolters Kluwer.

[5] A. Castells, X. Bessa, M. Daniels, C. Ascaso, A. M. Lacy, J. C. García-Valdecasas, L. Gargallo, F. Novell, E. Astudillo, X. Filella, and J. M. Piqué. 1998. Value of postoperative surveillance after radical surgery for colorectal cancer: results of a cohort study. *Diseases of the Colon and Rectum* 41, 6 (June 1998), 714–723; discussion 723–724. https://doi.org/10.1007/BF02236257

[6] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307 (Sept. 2018), 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

[7] Laura A. M. Duineveld, Kristel M. van Asselt, Willem A. Bemelman, Anke B. Smits, Pieter J. Tanis, Henk C. P. M. van Weert, and Jan Wind. 2016. Symptomatic and Asymptomatic Colon Cancer Recurrence: A Multicenter Cohort Study. *Annals of Family Medicine* 14, 3 (May 2016), 215–220. https://doi.org/10.1370/afm.1919

[8] Shaker El-Sappagh, Tamer Abuhmed, S. M. Riazul Islam, and Kyung Sup Kwak. 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing* 412 (Oct. 2020), 197–215. https://doi.org/10.1016/j.neucom.2020.05.087

[9] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining* 2014 (Aug. 2014), 75–84. https://doi.org/10.1145/2623330.2623742

[10] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. 2018. NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine* 158 (May 2018), 113–122. https://doi.org/10.1016/j.cmpb.2018.01.025

[11] R. A. Graham, S. Wang, P. J. Catalano, and D. G. Haller. 1998. Postsurgical surveillance of colon cancer: preliminary cost analysis of physician examination, carcinoembryonic antigen testing, chest x-ray, and colonoscopy. *Annals of Surgery* 228, 1 (July 1998), 59–63. https://doi.org/10.1097/00000658-199807000-00009

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

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

[14] N. Jaques, Sara Taylor, Akane Sano, and Rosalind W. Picard. 2017. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. https://doi.org/10.1109/ACII.2017.8273601

[15] Dan Jiang, Junhua Liao, Haihan Duan, Qingbin Wu, Gemma Owen, Chang Shu, Liangyin Chen, Yanjun He, Ziqian Wu, Du He, Wenyan Zhang, and Ziqiang Wang. 2020. A machine learning-based prognostic predictor for stage III colon cancer. *Scientific Reports* 10, 1 (June 2020), 10333. https://doi.org/10.1038/s41598-020-67178-0 Number: 1 Publisher: Nature Publishing Group.

[16] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (May 2016), 160035. https://doi.org/10.1038/sdata.2016.35 Number: 1 Publisher: Nature Publishing Group.

[17] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang. 2019. A Comparative Study on Transformer vs RNN in Speech Applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 449–456. https://doi.org/10.1109/ASRU46091.2019.9003750

[18] NaNa Keum and Edward Giovannucci. 2019. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews. Gastroenterology & Hepatology* 16, 12 (Dec. 2019), 713–732. https://doi.org/10.1038/s41575-019-0189-8

[19] Scott Kopetz, Josep Tabernero, Robert Rosenberg, Zhi-Qin Jiang, Víctor Moreno, Thomas Bachleitner-Hofmann, Giovanni Lanza, Lisette Stork-Sloots, Dipen Maru, Iris Simon, Gabriel Capellà, and Ramon Salazar. 2015. Genomic Classifier ColoPrint Predicts Recurrence in Stage II Colorectal Cancer Patients More Accurately Than Clinical Factors. *The Oncologist* 20, 2 (Feb. 2015), 127–133. https://doi.org/10.1634/theoncologist.2014-0325

[20] Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 641–652. https://www.aclweb.org/anthology/C18-1054

[21] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. 2017. Temporal Convolutional Networks for Action Segmentation and Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1003–1012. https://doi.org/10.1109/CVPR.2017.113 ISSN: 1063-6919.

[22] Garam Lee, Byungkon Kang, Kwangsik Nho, Kyung-Ah Sohn, and Dokyoon Kim. 2019. MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework. *Frontiers in Genetics* 10 (2019). https://doi.org/10.3389/fgene.2019.00617 Publisher: Frontiers.

[23] Joo Han Lee and Seong-Wook Lee. 2017. The Roles of Carcinoembryonic Antigen in Liver Metastasis and Therapeutic Approaches. https://doi.org/10.1155/2017/7521987 ISSN: 1687-6121 Pages: e7521987 Publisher: Hindawi Volume: 2017.

[24] Peter F Lenehan, Lisa A Boardman, Douglas Riegert-Johnson, Giovanni De Petris, David W Fry, Jeanne Ohrnberger, Eugene R Heyman, Brigitte Gerard, Arpit A Almal, and William P Worzel. 2012. Generation and external validation of a tumor-derived 5-gene prognostic signature for recurrence of lymph node-negative, invasive colorectal carcinoma. *Cancer* 118, 21 (Nov. 2012), 5234–5244. https://doi.org/10.1002/cncr.27628

[25] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Advances in Neural Information Processing Systems* 32 (2019), 5243–5253. https://papers.nips.cc/paper/2019/hash/6775a0635c302542da2c32aa19d86be0-Abstract.html

[26] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. 2017. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. In *Information Processing in Medical Imaging (Lecture Notes in Computer Science)*, Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen (Eds.). Springer International Publishing, Cham, 348–360. https://doi.org/10.1007/978-3-319-59050-9_28

[27] Gershon Y. Locker, Stanley Hamilton, Jules Harris, John M. Jessup, Nancy Kemeny, John S. Macdonald, Mark R. Somerfield, Daniel F. Hayes, Robert C. Bast, and ASCO. 2006. ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 24, 33 (Nov. 2006), 5313–5327. https://doi.org/10.1200/JCO.2006.08.2644

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[29] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data* 5, 1 (Sept. 2018), 180178. https://doi.org/10.1038/sdata.2018.178 Number: 1 Publisher: Nature Publishing Group.

[30] John N. Primrose, Rafael Perera, Alastair Gray, Peter Rose, Alice Fuller, Andrea Corkhill, Steve George, and David Mant. 2014. Effect of 3 to 5 Years of Scheduled CEA and CT Follow-up to Detect Recurrence of Colorectal Cancer: The FACS Randomized Clinical Trial. *JAMA* 311, 3 (Jan. 2014), 263. https://doi.org/10.1001/jama.2013.285718

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Lecture Notes in Computer Science)*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[32] Bethany Shinkins, Brian D. Nicholson, John Primrose, Rafael Perera, Timothy James, Sian Pugh, and David Mant. 2017. The diagnostic accuracy of a single CEA blood test in detecting colorectal cancer recurrence: Results from the FACS trial. *PLoS ONE* 12, 3 (March 2017). https://doi.org/10.1371/journal.pone.0171810

[33] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B. Ellis, Erwin P. Bottinger, and John V. Guttag. 2015. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of biomedical informatics* 53 (Feb. 2015), 220–228. https://doi.org/10.1016/j.jbi.2014.11.005

[34] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S. Hveem, Knut Liestøl, John Maddison, Hanne A. Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, Inger Nina Farstad, Enric Domingo, David N. Church, Arild Nesbakken, Neil A. Shepherd, Ian Tomlinson, Rachel Kerr, Marco Novelli, David J. Kerr, and Håvard E. Danielsen. 2020. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet* 395, 10221 (Feb. 2020), 350–360. https://doi.org/10.1016/S0140-6736(19)32998-8 Publisher: Elsevier.

[35] Caspar G. Sørensen, William K. Karlsson, Hans-Christian Pommergaard, Jakob Burcharth, and Jacob Rosenberg. 2016. The diagnostic accuracy of carcinoembryonic antigen to detect colorectal cancer recurrence – A systematic review. *International Journal of Surgery* 25 (Jan. 2016), 134–144. https://doi.org/10.1016/j.ijsu.2015.11.065

[36] Xiao Tan, Andrew T. Su, Hamideh Hajiabadi, Minh Tran, and Quan Nguyen. 2021. Applying Machine Learning for Integration of Multi-Modal Genomics Data and Imaging Data to Quantify Heterogeneity in Tumour Tissues. In *Artificial Neural Networks*, Hugh Cartwright (Ed.). Springer US, New York, NY, 209–228. https://doi.org/10.1007/978-1-0716-0826-5_10

[37] Guojun Tong, Wei Xu, Guiyang Zhang, Jian Liu, Zhaozheng Zheng, Yan Chen, Pingping Niu, and Xuting Xu. 2018. The role of tissue and serum carcinoembryonic antigen in stages I to III of colorectal cancer—A retrospective cohort study. *Cancer Medicine* 7, 11 (Oct. 2018), 5327–5338. https://doi.org/10.1002/cam4.1814

[38] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 4971–4980. https://doi.org/10.1109/CVPR.2017.528

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[40] Cheng Wang, Mathias Niepert, and Hui Li. 2018. LRMM: Learning to Recommend with Missing Modalities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3360–3370. https://doi.org/10.18653/v1/D18-1373

[41] Yucan Xu, Lingsha Ju, Jianhua Tong, Cheng-Mao Zhou, and Jian-Jun Yang. 2020. Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. *Scientific Reports* 10, 1 (Feb. 2020), 1–9. https://doi.org/10.1038/s41598-020-59115-y Number: 1 Publisher: Nature Publishing Group.

[42] Yajun Yu, Megan Carey, William Pollett, Jane Green, Elizabeth Dicks, Patrick Parfrey, Yildiz E. Yilmaz, and Sevtap Savas. 2019. The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Medicine* 17 (July 2019). https://doi.org/10.1186/s12916-019-1379-5

[43] Syed Nabeel Zafar, Chung-Yuan Hu, Rebecca A. Snyder, Amanda Cuddy, Y. Nancy You, Lisa M. Lowenstein, Robert J. Volk, and George J. Chang. 2020. Predicting Risk of Recurrence After Colorectal Cancer Surgery in the United States: An Analysis of a Special Commission on Cancer National Study. *Annals of Surgical Oncology* 27, 8 (Aug. 2020), 2740–2749. https://doi.org/10.1245/s10434-020-08238-7

[44] Juan Zhao, QiPing Feng, Patrick Wu, Roxana A. Lupu, Russell A. Wilke, Quinn S. Wells, Joshua C. Denny, and Wei-Qi Wei. 2019. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Scientific Reports* 9 (Jan. 2019). https://doi.org/10.1038/s41598-018-36745-x