Improving information fusion on multimodal clinical data in classification settings

Sneha Jha
Imperial College London
sneha.jha@imperial.ac.uk

Erik Mayer
Imperial College London
e.mayer@imperial.ac.uk

Mauricio Barahona
Imperial College London
m.barahona@imperial.ac.uk

Abstract

Clinical data often exists in different forms across the lifetime of a patient's interaction with the healthcare system - structured, unstructured or semi-structured data in the form of laboratory readings, clinical notes, diagnostic codes, imaging and audio data of various kinds, and other observational data. Formulating a representation model that aggregates information from these heterogeneous sources may allow us to jointly model on data with more predictive signal than noise and help inform our model with useful constraints learned from better data. Multimodal fusion approaches help produce representations combined from heterogeneous modalities, which can be used for clinical prediction tasks. Representations produced through different fusion techniques require different training strategies. We investigate the advantage of adding narrative clinical text to structured modalities to classification tasks in the clinical domain. We show that while there is a competitive advantage in combined representations of clinical data, the approach can be helped by training guidance customized to each modality. We show empirical results across binary/multiclass settings, single/multitask settings and unified/multimodal learning rate settings for early and late information fusion of clinical data.

1 Introduction

A variety of clinical use cases emerge where it is not sufficient to use a single data modality as input to a learning or decision making system (Weber et al., 2014). A single data modality is often known to be insufficient for a clinical purpose. For instance, diagnoses that require imaging data as well as lab values or outcomes that depend on values routinely recorded in the narrative text but not elsewhere. An additional modality can be used to characterize additional features. For instance, information in narrative text that conflicts with or adds specificity to diagnosis or procedure codes or imaging data that can indicate severity of a condition not recorded in structured form or qualitatively mentioned in narrative text. Sometimes, a modality with highly predictive or informative features is particularly expensive or invasive and an alternative source is present that may have features unintelligible or hard to parse for humans. Also, most clinical machine learning systems focus on one clinical prediction task at a time (D'Costa et al., 2020; Ji et al., 2020). However, in real-world systems more than one such task are often performed simultaneously and are interrelated (Yang and Wu, 2021). There is a need to investigate task-specific unified representations of multimodal clinical data in both single-task and multi-task settings to improve decisions in the clinical workflow by demonstrating an increase in predictive power, robustness, and confidence over any single mode of data (Tiulpin et al., 2019). Besides creating and combining efficient representations of data from more than one modality, we also need to study the factors that affect the design and evaluation of these multimodal representations.

2 Multimodal Representations

There are various ways modalities of clinical data can be combined. Multimodal deep learning models integrate information at various possible steps. This can occur in the following ways –

- By finding a common representation for input data for a specific task before modeling. e.g., extracting clinical mentions from narrative text and concatenating it with independent diagnostic signals to form a model input.
- By jointly learning intermediate feature representations for one or more additional modalities, besides the basic input e.g., learning text embeddings from narrative text and using that as an additional input besides the structured

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 154 - 159

data to the same neural network. This is designed for the training algorithm to jointly improve the intermediate embeddings along with the task-specific loss.

- By modeling each modality separately and combining predictions from different models under a task-specific scheme. e.g., aggregating diagnostic predictions from a text modeling and an image modeling system through an averaging scheme or a meta classifier.

As detailed in (Baltrušaitis et al., 2018), multimodal models can be categorized by the fusion techniques based on which they learn the join representations of underlying data. The most common approaches are called early fusion (Chen and Jin, 2015) where individual modality features are combined right after feature extraction and late fusion (Atrey et al., 2010) which combines outputs from unimodal predictors jointly. Early fusion is expected to capture some of the feature-level interactions of each modality and often is easier to model and train. On the other hand, late fusion allows for more flexibility and is expected to model individual modalities better and also handles scenarios where one or more modalities are missing. However, it cannot be expected to capture low level interaction between the modalities. While training late fusion models, the simplest choice is to use the same learning rate across all modalities. But it is both intuitive as well as demonstrable through layer analysis (Yao and Mihalcea, 2022) that learning rates for different modalities can differ a lot and must be handled separately to optimize learning from heterogeneous sources.

3 Methodology

3.1 Data source

We use a publicly available clinical data set - Medical Information Mart for Intensive Care (MIMIC) Johnson et al. (2016) - containing data across various modalities for patients admitted and readmitted to the intensive care unit (ICU). MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality, including post-hospital discharge. It contains highly granular data, including vital signs, laboratory results, and medications.

3.2 Data modalities

The following structured and semi-structured modalities typically found in inpatient settings in clinical data were extracted and compiled at a patient level -

- **Clinical Notes** - Free-form narrative text is entered by clinicians and nurses during the stay of a patient and these usually summarize reasons for admission, details of treatment, nutrition, and the patient's symptoms and diagnoses. These clinical notes are temporally ordered.

- **Tabular Data** – Metadata such as sex, age, height, weight at admission, the type of the ICU, and other tabular inputs were recorded for each patient. Values such as weight, may vary during the patient's stay and are potentially part of the time series data set as well.

- **Time Series data** – Various temporal physiological variables, such as diastolic blood pressure, systolic blood pressure, oxygen saturation, were recorded for each patient. These physiological variables are recorded irregularly, and they are important indicators of the patient's condition during the hospital stay.

We add the two different kinds of structured data from the MIMIC-III dataset to the clinical text. Data is preprocessed as in (Harutyunyan et al., 2019), excluding ICU stays with missing events or missing length-of-stay and excluding patients younger than eighteen years of age since both clinical dynamics and clinical documentation of paediatrics facilities are significantly different from those of adults.

3.3 Experiments

The experiments focus on the following two tasks -

- **In-hospital mortality prediction** : To predict death by the end of the hospital stay based on first 48 hours of observations. To prevent mortality is the primary aim and a number of task formulations as in (Harutyunyan et al., 2019) and (Khadanga et al., 2019) attempt to predict

patient survival at the end of the hospital stay. The hypothesis is that observations from the first 48 hours of a patient's stay in the ICU include crucial clues towards the probability of survival.

- **Phenotyping** : To predict a patient's phenotype at the time of discharge in terms of billing codes. This is a multilabel classification task and the target label set is derived from the billing code at a patients discharge, which is then converted to 25 labels following the procedure from (Harutyunyan et al., 2019).

The above two are standard tasks in clinical prediction settings and allowed us to compare directly with prior work such as (Harutyunyan et al., 2019) and (Khadanga et al., 2019). They provide two representative tasks in binary and multiclass, multilabel settings. Multitask learning is a particularly useful direction to explore in clinical settings with potential to capture dependencies between tasks, especially in low-data regimes. It was first proposed in the clinical prediction setting by (Caruana et al., 1995), where they used future lab values as auxiliary targets during training improving prediction of mortality among pneumonia patients.
Multimodal embeddings are used as the input to the two task- specific components. Each layer per task is a fully connected network with h_t hidden units, a dropout layer with dropout probability α_t, a ReLU activation, and an output layer matching the shape of the individual component's respective task. Each task-specific component shares the base multimodal embedding but is independent of the other layers. The multimodal encoder is comprised of one child encoder per input modality. In the early fusion setup, the multimodal embedding is a concatenation of the outputs from each child encoder. In the late fusion setting, for each time step, the model structure is a linear layer with 512 hidden states with ReLU activation projected to a 128-dimensional linear layer to predict the output class. For the phenotyping task each of the 25 output neurons has a sigmoid activation. The results for late fusion multimodal learning have been reported only in single-task settings.

Each task-specific component can employ a task-specific loss. To learn across both tasks simultaneously in the multitask experiments, we take the weighted sum of all the losses resulting to form the multitask loss. The current experiments use cross entropy loss for both tasks. To find multitask weights, we used uncertainty weighting described in (Kendall et al., 2018).
Time series encoder.Given a patient's ICU stay of length of T hours, the time series data is resampled with 1 hour interval to obtain $[TS_t]$ from t = 1 to t = T. The time series encoder is an LSTM (Hochreiter and Schmidhuber, 1997). The input $[TS_t]$ at time step t is directly the input to an LSTM model (Hochreiter and Schmidhuber, 1997) along with the previous states, and the next hidden state is the extracted feature, denoted by f_t^n .

$$f_t^n = LSTM(TS_t, f_{t-1}^n) \tag{1}$$

The experiments use a 1-layer LSTM with 256 hidden units as the time series encoder.
Clinical Text Encoder. For each ICU stay, there are N clinical notes recorded at irregular intervals.The chart time of these notes are $[Time(i)_{i=1}^N]$ where $N <= T$ The convolutional model TextCNN in (Kim, 2014) is used to extract features from textual clinical notes. To create embeddings from N_t notes collected at time $Time_i$, the CNN model gives us the feature matrix z per clinical note. A weighted average of all notes, weighted by recency produces a feature vector for a record.

$$weight(t,i) = exp(-\lambda(t - Time_i)) \tag{2}$$

$$f_t = \gamma \sum_{i=1}^{N} weight(t,i).z_i \tag{3}$$

Here, λ is a scaling factor and γ is a normalization term. The embeddings are generated using word2vec embeddings (Mikolov et al., 2013). The TextCNN model has three 1-D kernels of size 2,3 and 4 with 128 filters each.
Tabular Data Encoder. To process the tabular inputs, we learn an embedding table for each categorical input dimension as in (De Brébisson et al., 2015) and individual features are concatenated to form one tabular embedding. All features are represented as 32-dimensional embeddings.

In the early fusion setup, the default Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of $1e$-4 with early stopping. The mortality prediction task uses $h_t = 108$. The phenotyping task uses $h_t = 512$. In the late fusion setup, we ran the following sets of experiments -

- **Unified learning rate across modalities** : We use the default Adam optimizer with a learning rate of $3e$-4 with early stopping. The mortality prediction task uses $h_t = 108$. The phenotyping task uses $h_t = 512$.
- **Adapted learning rate per modality** : We use the best fine-tuned learning rate per modality for each of them while training the late-fusion model.

We observed better results with the AdamW optimizer (Loshchilov and Hutter, 2017) but report results using the default Adam optimizer to be able to at least partially compare with (Harutyunyan et al., 2019) since learning rates can vary a lot based on the optimizer used.

4 Results

Since the mortality prediction is a binary classification problem, we report the AUC-PR numbers, which is a standard evaluation metric. Because diseases can co-occur and a majority of patients often have more than one diagnosis, phenotyping is a multilabel classification problem, which requires the performance to be reported by averaging across labels or examples. These labels have varying base rates. In imbalanced tasks, (Lipton et al., 2014) show that if the predictive features for rare labels are lost, which is possible due to feature selection, macro F1 is an unsuitable metric. We report the macro AUC-ROC, which is the unweighted mean of AUC-ROC for each label. We also add a weighted average AUC-ROC metric accounting for base rate of the diseases in Table 3. The phenotyping task does not categorically benefit from the multitask setting. The model is trained to jointly predict 25 labels which in itself might have a regularizing effect akin to multitask learning and the additional regularization expected from adding the in-hospital mortality prediction task may be unable to provide further significant improvement over the single-task setting.

We follow (Harutyunyan et al., 2019) as the baseline for the MultiTask TimeSeries set-up and (Khadanga et al., 2019) for the SingleTask Notes + TimeSeries set-up. We show results of the early fusion runs in Table 1 and the late fusion runs in Table 2.

We also report error bounds for the experiments by choosing observations at the 2.5th percentile and the 97.5th percentiles and reporting the median. This was computed by drawing 5000 samples with replacement 100 times from the test set.

5 Related Work

We refer to (Harutyunyan et al., 2019) that uses a single modality of time series in a multi-task setting using LSTMs and channel-wise LSTMs. Similarly, (Khadanga et al., 2019) presents a unimodal model with clinical notes only for individual task settings but they also report additional results in a multimodal setting using both time series and text data without using multitask learning. We reuse some of the multitasking configuration for the MIMIC dataset described in (Huang et al., 2020). There are also available comparisons against baseline logistic regression and random forest models in (Zhang et al., 2020). All of these use only a unified learning rate across modalities. A number of works note the need of exploiting modality-specific features such as (Wang et al., 2015; Liu et al., 2018) for combining text with other modalities such as image and audio. In the late-fusion setting, a closely related work is (Yao and Mihalcea, 2022) that investigates modality-specific learning rates. They do not investigate a multitask setting and also study modalities structurally different from ours. Another closely related work is (Fujimori et al., 2019) that take up the issue of potential overfitting to certain modalities. Their approach is via early stopping and is still closer to our unified learning rate set up. It is also worth noting that the typical modalities in clinical data are very domain-specific and even well-studied modalities such as text in general-domain NLP often behave differently in the clinical domain (Rumshisky et al., 2020).

6 Limitations and Future Work

This work investigates one way to adapt learning rates to modalities. There can be more adaptive strategies that take a priori clinical knowledge about a modality into account, which is a possible topic of future work. The late fusion methods discussed are also occasionally unstable during training. It is also conceivable that clinical text with different linguistic structure (e.g. short, more standardised radiology reports vs longer, less structured progress reports) behave differently when combined with other modalities. Further investigation is required to mitigate these issues. The current work aims to show the effect of adding modalities and adapting parameters specific to useful modal-

Task	Modality	IHMortality	Phenotype
SingleTask	Notes	0.517±0.052	0.712±0.004
SingleTask	TimeSeries	0.423±0.052	0.788±0.004
SingleTask	Notes + Tabular	0.519±0.04	0.72±0.007
SingleTask	Notes + TimeSeries	0.580±0.05	0.796±0.005
SingleTask	Notes + TimeSeries + Tabular	0.570±0.051	**0.814**±0.005
MultiTask	TimeSeries	0.423±0.052	0.77±0.005
MultiTask	TimeSeries + Tabular	0.526±0.003	0.781±0.002
MultiTask	Notes + TimeSeries	**0.601**±0.05	0.773±0.005
MultiTask	Notes + TimeSeries + Tabular	0.599±0.051	0.813±0.004

Table 1: Effect of multimodal learning with early fusion

RateAcrossModality	Modality	IHMortality	Phenotype
-	Notes	0.517±0.052	0.712±0.004
-	Time series	0.423±0.052	0.788±0.004
unified	Notes + TimeSeries	0.590±0.049	0.802±0.005
multimodal	Notes + TimeSeries	0.614±0.047	0.803±0.004
unified	Notes + TimeSeries + Tabular	0.601±0.051	0.815±0.005
multimodal	Notes + TimeSeries + Tabular	**0.62**±0.050	**0.817**±0.004

Table 2: Effect of multimodal learning with late fusion with varying learning rate across modalities

Task	Modality	Phenotype
SingleTask	Notes	0.707±0.003
SingleTask	TimeSeries	0.781±0.005
SingleTask	Notes + Tabular	0.73±0.006
SingleTask	Notes + TimeSeries	0.789±0.007
SingleTask	Notes + TimeSeries + Tabular	0.808±0.002
MultiTask	TimeSeries	0.767±0.006
MultiTask	TimeSeries + Tabular	0.772±0.002
MultiTask	Notes + TimeSeries	0.766±0.004
MultiTask	Notes + TimeSeries + Tabular	**0.812**±0.004

Table 3: Effect of multimodal learning with early fusion on phenotype (AUC-ROC weighted by label prevalence)

ities. Future work will also address comparisons not possible with existing baselines. More complex models with advanced architecture can be applied in a modular fashion in both single task and multi-task settings.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Rich Caruana, Shumeet Baluja, and Tom Mitchell. 1995. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, 8.

Shizhe Chen and Qin Jin. 2015. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56.

Alister D'Costa, Stefan Denkovski, Michal Malyska, Sae Young Moon, Brandon Rufino, Zhen Yang, Taylor Killian, and Marzyeh Ghassemi. 2020. Multiple sclerosis severity classification from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 7–23, Online. Association for Computational Linguistics.

Alexandre De Brébisson, Étienne Simon, Alex Auvolat, Pascal Vincent, and Yoshua Bengio. 2015. Artifi-

cial neural networks applied to taxi destination prediction. ECMLPKDDDC'15, page 40–51, Aachen, DEU. CEUR-WS.org.

Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. 2019. Modality-specific learning rate control for multimodal classification. In *Asian Conference on Pattern Recognition*, pages 412–422. Springer.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96. ArXiv: 1703.07771.

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

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):1–9.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, Online. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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):1–9.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, Hong Kong, China. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors. 2020. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online.

A. Tiulpin, S. Klein, S. Bierma-Zeinstra, J. Thevenot, Esa Rahtu, J. Meurs, E. Oei, and S. Saarakkala. 2019. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. *Scientific Reports*.

Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. 2015. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 347–354.

Griffin M. Weber, Kenneth D. Mandl, and Isaac S. Kohane. 2014. Finding the Missing Link for Big Biomedical Data. *JAMA*, 311(24):2479–2480.

Bo Yang and Lijun Wu. 2021. How to leverage multimodal ehr data for better medical predictions? *arXiv preprint arXiv:2110.15763*.

Yiqun Yao and Rada Mihalcea. 2022. Modality-specific Learning Rates for Effective Multimodal Additive Late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834, Dublin, Ireland. Association for Computational Linguistics.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. 2020. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, 20(1):1–11.