Using Clinical Notes with Time Series Data for ICU Management

Swaraj Khadanga
University of Minnesota
Minneapolis, MN
khada004@umn.edu

Karan Aggarwal
University of Minnesota
Minneapolis, MN
aggar081@umn.edu

Shafiq Joty
Nanyang Technological University
Singapore
srjoty@ntu.edu.sg

Jaideep Srivastava
University of Minnesota
Minneapolis, MN
srivasta@umn.edu

Abstract

Monitoring patients in ICU is a challenging and high-cost task. Hence, predicting the condition of patients during their ICU stay can help provide better acute care and plan the hospital's resources. There has been continuous progress in machine learning research for ICU management, and most of this work has focused on using time series signals recorded by ICU instruments. In our work, we show that adding clinical notes as another modality improves the performance of the model for three benchmark tasks: in-hospital mortality prediction, modeling decompensation, and length of stay forecasting that play an important role in ICU management. While the time-series data is measured at regular intervals, doctor notes are charted at irregular times, making it challenging to model them together. We propose a method to model them jointly, achieving considerable improvement across benchmark tasks over baseline time-series model. Our implementation can be found at https://github.com/kaggarwal/ClinicalNotesICU.

1 Introduction

With the advancement of medical technology, patients admitted into the intensive care unit (ICU) are monitored by different instruments on their bedside, which measure different vital signals about patient's health. During their stay, doctors visit the patient intermittently for check-ups and make *clinical notes* about the patient's health and physiological progress. These notes can be perceived as *summarized expert knowledge* about the patient's state. All these data about instrument readings, procedures, lab events, and clinical notes are recorded for reference. Availability of ICU data and enormous progress in machine learning have opened up new possibilities for health care research. Monitoring patients in ICU is a challenging and high-cost task. Hence, predicting the condition of patients during their ICU stay can help plan better resource usage for patients that need it most in a cost-effective way. Prior works (Harutyunyan et al., 2019; Ghassemi et al., 2015; Suresh et al., 2018; Song et al., 2018; Caballero Barajas and Akella, 2015) have focused exclusively on modeling the problem using the time series signals from medical instruments. Expert knowledge from doctor's notes has been ignored in the literature.

In this work, we use clinical notes in addition to the time-series data for improved prediction on benchmark ICU management tasks (Harutyunyan et al., 2019). While the time-series data is measured continuously, the doctor notes are charted at intermittent times. This creates a new challenge to model continuous time series and discrete-time note events jointly. We propose such a multimodal deep neural network that comprises of recurrent units for the time-series and convolution network for the clinical notes. We demonstrate that adding clinical notes improves the performance on in-hospital mortality prediction, modeling decompensation, and length of stay forecasting tasks.

2 Related Work

We provide a review of machine learning approaches for clinical prediction tasks.

Biomedical natural language processing The recent developments in deep learning-based techniques for NLP have been adapted for clinical notes. Convolutional neural networks have been used to predict ICD codes from clinical texts (Mullenbach et al., 2018; Li et al., 2018; Gangavarapu et al., 2019). Rios and Kavuluru (2015) and Baker et al. (2016) used convolutional neural networks to classify various biomedical articles. Pre-trained word and sentence embeddings have also shown good results for sentence similarity tasks (Chen

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6432–6437,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

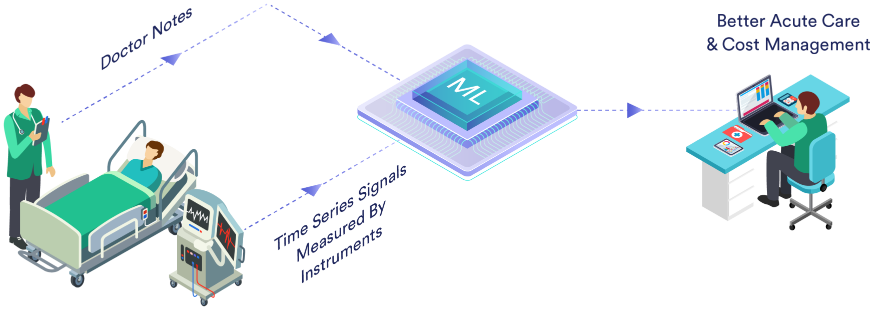

Figure 1: Doctor notes compliments measured physiological signals for better ICU management.

et al., 2018). Recently, there is an interest in the community to use clinical notes for ICU related tasks (Jin et al., 2018; Boag et al., 2018; Liu et al., 2019; Huang et al., 2019). Given the long structured nature of the clinical text, we prefer convolutional neural networks over recurrent networks, as demonstrated in previous studies (Zhang et al., 2016; Boag et al., 2018). The work closest to ours is Jin et al. (2018), who use an aggregated word embeddings of clinical notes for in-hospital mortality prediction.

ICU management related literature ICU management literature has focused exclusively on using time-series measurements for the prediction tasks Harutyunyan et al. (2019); Ghassemi et al. (2015); Suresh et al. (2018); Song et al. (2018); Caballero Barajas and Akella (2015). Recurrent neural networks have been models of choice for these recent works, with additional gains from using attention or multi-task learning (Song et al. (2018)). Xu et al. (2018) accommodated supplemental information like diagnosis, medications, and lab events to improve model performance. We use RNNs for modeling time-series in this work, utilizing the setup identical to Harutyunyan et al. (2019).

Multi-modal learning has shown success in speech, natural language, and computer vision (Ngiam et al. (2011), Mao et al. (2014)). Recently, a lot of work has been done using images/videos with natural language text (Elliott et al. (2016)). We use a similar intuition for utilizing clinical notes with time-series data for ICU management tasks. In the next section, we define the three benchmark tasks we evaluate in this work.

3 Prediction Tasks

We use the definitions of the benchmark tasks defined by Harutyunyan et al. (2019) as the following three problems:

1. **In-hospital Mortality**: This is a binary classification problem to predict whether a patient dies before being discharged from the first two days of ICU data.

2. **Decompensation**: Focus is to detect patients who are physiologically declining. Decompensation is defined as a sequential prediction task where the model has to predict at each hour after ICU admission. Target at each hour is to predict the mortality of the patient within a 24 hour time window.

3. **Length of Stay Forecasting (LOS)**: The benchmark defines LOS as a prediction of bucketed remaining ICU stay with a multi-class classification problem. Remaining ICU stay time is discretized into 10 buckets: $\{0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, 7-8, 8-14, 14+\}$ days where first bucket, covers the patients staying for less than a day (24 hours) in ICU and so on. This is only done for the patients that did not die in ICU.

These tasks have been identified as key performance indicators of models that can be beneficial in ICU management in the literature. Most of the recent work has focused on using RNN to model the temporal dependency of the instrument time series signals for these tasks (Harutyunyan et al. (2019), Song et al. (2018)).

4 Methods

In this section, we describe the models used in this study. We start by introducing the notations used, then describe the baseline architecture, and finally present our proposed multimodal network.

For a patient's length of ICU stay of T hours, we have time series observations, x_t at each time step t (1 hour interval) measured by instruments along with doctor's note n_i recorded at *irregular* time stamps. Formally, for each patient's ICU stay, we have time series data $[x_t]_{t=1}^T$ of length T, and K doctor notes $[N_i]_{i=1}^K$ charted at time $[TC(i)]_{i=1}^K$, where K is generally much smaller than T. For **in-hospital mortality** prediction, m is a binary label at $t = 48$ hours, which indicates whether the person dies in ICU before being discharged. For **decompensation** prediction performed hourly, $[d_t]_{t=5}^T$ are the binary labels at each time step t, which indicates whether the person dies in ICU within the next 24 hours. For **LOS** forecasting also performed hourly, $[l_t]_{t=5}^T$ are multi-class labels defined by buckets of the remaining length of stay of the patient in ICU. Finally, we denote N_T as the concatenated doctor's note during the ICU stay of the patient (*i.e.,*, from $t = 1$ to $t = T$).

4.1 Baseline: Time-Series LSTM Model

Our baseline model is similar to the models defined by Harutyunyan et al. (2019). For all the three tasks, we used a Long Short Term Memory or LSTM (Hochreiter and Schmidhuber, 1997) network to model the temporal dependencies between the time series observations, $[x_t]_{t=1}^T$. At each step, the LSTM composes the current input x_t with its previous hidden state h_{t-1} to generate its current hidden state h_t; that is, $h_t = \text{LSTM}(x_t, h_{t-1})$ for $t = 1$ to $t = T$. The predictions for the three tasks are then performed with the corresponding hidden states as follows:

$$\begin{aligned} \hat{m} &= \text{sigmoid}(W_m h_{48} + b_m) \\ \hat{d}_t &= \text{sigmoid}(W_d h_t + b_d) \quad \text{for } t = 5 \dots T \\ \hat{l}_t &= \text{softmax}(W_l h_t + b_l) \quad \text{for } t = 5 \dots T \end{aligned} \tag{1}$$

where $\hat{m}$, $\hat{d}_t$, and $\hat{l}_t$ are the probabilities for in-hospital mortality, decompensation, and LOS, respectively, and W_m, W_d, and W_l are the respective weights of the fully-connected (FC) layer. Notice that the in-hospital mortality is predicted at end of 48 hours, while the predictions for decompensation and LOS tasks are done at each time step after first four hours of ICU stay. We trained the models using cross entropy (CE) loss defined as below.

$$\begin{aligned} \mathcal{L}_{\text{ihm}} &= \text{CE}(m, \hat{m}) \\ \mathcal{L}_{\text{decom}} &= \frac{1}{T} \sum_t \text{CE}(d_t, \hat{d}_t) \\ \mathcal{L}_{\text{los}} &= \frac{1}{T} \sum_t \text{CE}(l_t, \hat{l}_t) \end{aligned} \tag{2}$$

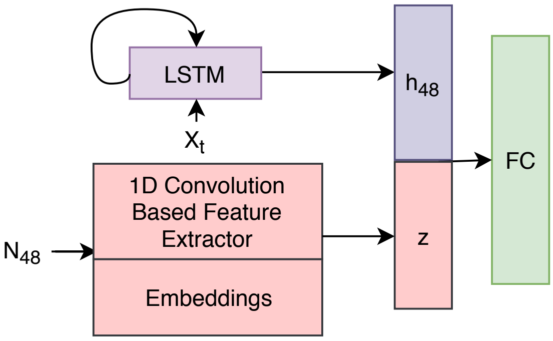

Figure 2: Block diagram from the in-hospital mortality multi-modal network.

4.2 Multi-Modal Neural Network

In our multimodal model, our goal is to improve the predictions by taking both the time series data x_t and the doctor notes n_i as input to the network.

Convolutional Feature Extractor for Doctor Notes. As shown in Fig. 2, we adopt a convolutional approach similar to Kim (2014) to extract the textual features from the doctor's notes. For a piece of clinical note N, our CNN takes the word embeddings $\mathbf{e} = (e_1, e_2, \dots, e_n)$ as input and applies 1D convolution operations, followed by maxpooling over time to generate a p dimensional feature vector $\hat{z}$, which is fed to the fully connected layer along side the LSTM output from time series signal (described in the next paragraph) for further processing. From now onwards, we denote the 1D convolution over note N as $\hat{z} = \text{Conv1D}(N)$.

Model for In-Hospital Mortality. This model takes the time series signals $[x_t]_{t=1}^T$ and all notes $[N_i]_{i=1}^K$ to predict the mortality label m at $t = T$ ($T = 48$). For this, $[x_t]_{t=1}^T$ is processed through an LSTM layer just like the baseline model in Sec. 4.1, and for the notes, we concatenate ($\otimes$) all the notes N_1 to N_K charted between $t = 1$ to $t = T$ to generate a single document N_T. More formally,

$$\begin{aligned} N_T &= N_1 \otimes N_2 \otimes \cdots \otimes N_K \\ h_t &= \text{LSTM}(x_t, h_{t-1}) \quad \text{for } t = 1 \dots T \\ \hat{z} &= \text{Conv1D}(N_T) \\ \hat{m} &= \text{sigmoid}(W_1 h_{48} + W_2 \hat{z} + b) \end{aligned} \tag{3}$$

We use pre-trained word2vec embeddings (Mikolov et al., 2013) trained on both MIMIC-III clinical notes and PubMed articles to initialize our methods as it outperforms other embeddings as shown in (Chen et al., 2018). We also freeze the embedding layer parameters, as we did not observe any improvement by fine-tuning them.

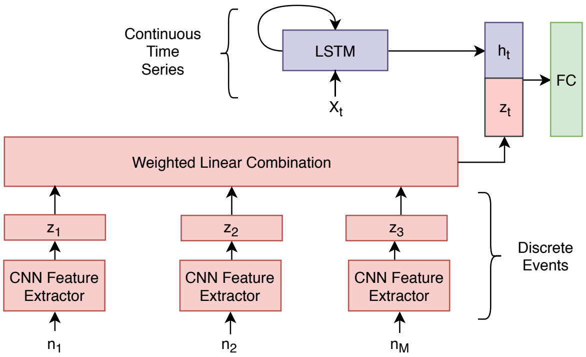

Figure 3: Block diagram from decompensation and length of stay prediction multi-modal network.

Model for Decompensation and Length of Stay. Being sequential prediction problems, modeling decompensation and length-of-stay requires special technique to align the discrete text events to continuous time series signals, measured at 1 event per hour. Unlike in-hospital mortality, here we extract feature maps z_i by processing each note N_i independently using 1D convolution operations. For each time step $t = 1, 2 \ldots T$, let z_t denote the extracted text feature map to be used for prediction at time step t. We compute z_t as follows.

$$\begin{aligned} z_i &= \text{Conv1D}(N_i) \quad \text{for } i = 1 \ldots K \\ w(t,i) &= \exp[-\lambda * (t - CT(i))] \\ z_t &= \frac{1}{M} \sum_{i=1}^{M} z_i w(t,i) \end{aligned} \tag{4}$$

where M is the number of doctor notes seen before time-step t, and λ is a decay hyperparameter tuned on a validation data. Notice that z_t is computed as a weighted sum of the feature vectors, where the weights are computed with an exponential decay function. The intuition behind using a decay is to give preference to recent notes as they better describe the current state of the patient.

The time series data x_t is modeled using an LSTM as before. We concatenate the attenuated output from the CNN with the LSTM output for the prediction tasks as follows:

$$\begin{aligned} h_t &= \text{LSTM}(x_t, h_{t-1}) \\ \hat{d}_t &= \text{sigmoid}(W_d^1 h_t + W_d^2 z_t + b) \\ \hat{l}_t &= \text{softmax}(W_l^1 h_t + W_l^2 z_t + b) \end{aligned} \tag{5}$$

Both our baselines and multimodal networks are regularized using dropout and weight decay. We used Adam Optimizer to train all our models.

5 Experiments

We used MIMIC-III (Johnson et al., 2016) dataset for all our experiments following Harutyunyan et al. (2019)'s benchmark setup for processing the time series signals from ICU instruments. We use the same test-set defined in the benchmark and 15% of remaining data as validation set. For in-hospital mortality task, only those patients are considered who were admitted in the ICU for at least 48 hours. However, we dropped all clinical notes which doesn't have any chart time associated and also dropped all the patients without any notes. Owing to this step, *our results are not directly comparable to the numbers reported by Harutyunyan et al. (2019)*. Notes which have been charted before ICU admission are concatenated and treated as one note at $t = 1$. After preprocessing, the number of patients for in-hospital mortality is 11,579 and 22,353 for the other two tasks.

For in-hospital mortality task, best performing baseline and multimodal network have 256 hidden units LSTM cell. For convolution operation, we used 256 filters for each of kernel size 2, 3 and 4. For decompensation and LOS prediction, we used 64 hidden units for LSTM and 128 filters for each 2,3 and 4 size convolution filters. The best decay factor λ for text features was 0.01. We implement our methods with TensorFlow Abadi et al. (2015)[1]. All our models were regularized using 0.2 dropout and 0.01 weight decay coefficient. We run the experiments 5 times with different initialization and report the mean and standard deviations.

6 Results

We use Area Under Precision-Recall (AUCPR) metric for in-hospital mortality and decompensation tasks as they suffer from class imbalance with only 10% patients suffering mortality, following the benchmark. Davis and Goadrich (2006) suggest AUCPR for imbalanced class problems. We use Cohen's linear weighted kappa, which measures the correlation between predicted and actual multi-class buckets to evaluate LOS in accordance with with Harutyunyan et al. (2019).

We compared the multimodal network with the baseline time series LSTM models for all three tasks. Results from our experiments are documented in Table 1. Our proposed multimodal net-

[1] The code can be found at `https://github.com/kaggarwal/ClinicalNotesICU`

In-Hospital Mortality

	AUCROC	AUCPR
Baseline (No Text)	0.844	0.487
Text-Only	0.793	0.303
MultiModal - Avg WE	0.851	0.492
MultiModal - 1DCNN	**0.865**	**0.525**

Decompensation

	AUCROC	AUCPR
Baseline (No Text)	0.892	0.325
Text-Only	0.789	0.081
MultiModal - Avg WE	0.902	0.311
MultiModal - 1DCNN	**0.907**	**0.345**

Length of Stay

	Kappa
Baseline (No Text)	0.438
Text Only	0.341
MultiModal - Avg WE	0.449
MultiModal - 1DCNN	**0.453**

Table 1: Mean metrics for all three tasks. Standard deviations: IHM (AUCROC < 0.004, AUCPR < 0.015), Decompensation (AUCROC < 0.008, AUCPR < 0.008), and LOS (Kappa < 0.003).

work outperforms the time series models for all three tasks. For in-hospital mortality prediction, we see an improvement of around 7.8% over the baseline time series LSTM model. The other two problems were more challenging itself than the first task, and modeling the notes for the sequential task was difficult. With our multimodal network, we saw an improvement of around 6% and 3.5% for decompensation and LOS, respectively.

We did not observe a change in performance with respect to results reported in benchmark (Harutyunyan et al., 2019) study despite dropping patients with no notes or chart time. In order to understand the predictive power of clinical notes, we also train text only models using CNN part from our proposed model. Additionally, we try average word embedding without CNN as another method to extract feature from the text as a baseline. Text-only-models perform poorly compared to time-series baseline. Hence, text can only provide additional predictive power on top of time-series data.

7 Conclusion

Identifying the patient's condition in advance is of critical importance for acute care and ICU management. Literature has exclusively focused on using time-series measurements from ICU instruments to this end. In this work, we demonstrate that utilizing clinical notes along with time-series data can improve the prediction performance significantly. In the future, we expect to improve more using advanced models for the clinical notes since text summarizes expert knowledge about a patient's condition.

References

Martín 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 Mané, 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 Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Simon Baker, Anna Korhonen, and Sampo Pyysalo. 2016. Cancer hallmark text classification using convolutional neural networks. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 1–9.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. Whats in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26.

Karla L Caballero Barajas and Ram Akella. 2015. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Desmond Elliott, Douwe Kiela, and Angeliki Lazaridou. 2016. Multimodal learning and reasoning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*.

Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, and Sowmya Kamath. 2019. Tags: Towards automated classification of unstructured clinical nursing notes. In *International Conference on Applications of Natural Language to Information Systems*, pages 195–207. Springer.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

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*, 3:160035.

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.

M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. Wang. 2018. Automated icd-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1.

Ning Liu, Pan Lu, Wei Zhang, and Jianyong Wang. 2019. Knowledge-aware deep dual networks for text-based mortality prediction. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1406–1417. IEEE.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

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.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Anthony Rios and Ramakanth Kavuluru. 2015. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267. ACM.

Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Harini Suresh, Jen J Gong, and John V Guttag. 2018. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 802–810. ACM.

Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573. ACM.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 795–804.