CS230

Looking for Low Vision in Electronic Health Records

Sophia Ying Wang
Department of Ophthalmology
Stanford University
sywang@stanford.edu

Sumit Singh
Stanford University
s99singh@stanford.edu

Samba Njie Jr.
Stanford University
snjie@stanford.edu*

Abstract

Although almost 3 million adults in the United States are visually impaired, referral rates to visual rehabilitation services are extremely low, despite the ability of these services to improve quality of life. If predictive algorithms could be developed to identify which patients with visual impairment could most benefit from low vision rehabilitation services, referrals could be made in a more automated fashion to facilitate access to these services. Our project aims to develop predictive algorithms leveraging structured (numeric, coded) and unstructured (free-text notes) clinical information in electronic health records (EHR) data to predict the prognosis of patients with low vision. We use a myriad of approaches on both the structured and unstructured data, including processing the clinical free text using named entity recognition and mapping the words to word embeddings for training. Our best performing models leveraged deep unsupervised feature learning strategies and low-depth, high-dimensional hidden layers architectures, outperforming baselines with an AUROC of 0.76.

1 Introduction

Almost 3 million adults in the United States are estimated to have low vision, defined as visual acuity less than 20/40 in the better-seeing eye. Furthermore, 1.28 million are estimated to have severely low vision—less than 20/200 visual acuity–meeting the definition of legal blindness in most states[1][2]. Low vision rehabilitation comprises multidisciplinary care to deliver a wide variety of interventions to improve patients' daily function and quality of life by maximizing the use of patients' remaining vision. However, the referral rate remains extraordinarily low, leaving almost 90% of patients who may benefit without access or awareness of these critical services[3]. Thus, there is a critical need to develop algorithms that can identify patients who may be candidates for low vision services in an automated manner to better facilitate referrals to low vision services and improve patients' quality of life.

In this paper, we present a novel application of deep learning and natural language processing in the field of ophthalmology using electronic health records (EHR). More specifically, we pose the following problem: given a set of patients, each represented by EHR data comprising both structured ophthalmology patient visits (diagnoses, medications, etc.) and unstructured free text, we predict the probability that their best corrected visual acuity of 20/40 will remain worse than 20/40 for the following one year.

*__Github Link:__ https://github.com/eyelovedata/cs230lowva

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

EHR data has a rich set of patient characteristics used for a wide variety of prediction problems in healthcare. A deep learning model using EHR data has been shown to successfully predict mortality within 3-12 months, thereby identifying those patients who may most benefit from a referral to palliative care services[4]. EHR data has also been leveraged for disease diagnosis classification using stacked denoising autoencoders to embed patient data and use a random forest for prediction, achieving around 0.773 AUROC and 91% accuracy in a variety of disease classification tasks[5].

In ophthalmology, there have been a few studies using EHR data to predict ophthalmic outcomes, yet none have focused on visual acuity on non-image EHR data and free-text notes. Electronic health records of data on military patients, including structured data (diagnoses, post-trauma clinical records, procedure codes, etc.) and free-text physician notes, have been used to classify whether there was an open globe injury by using note word embeddings and an SVM on 26,131 unique patients and 26,131 medical encounters, achieving a precision, recall and F1 of 92.5, 89.8, and 91.1 respectively[6]. Other studies, such as identifying cataract cases from EHR using OCR algorithms to scan OCR documents of text as well as NLP to characterize free text on a database of 13,000 patients with PPVs as high as 99%, yet did not specify the in-house NLP and OCR algorithms they used[7]. To our knowledge, there have been no prior studies predicting visual acuity prognosis using EHR data.

3 Dataset and Features

Data Source: The **Stanford Research Repository (STARR)**[8] captures data from the Stanford EHR, including structured data comprising demographics, diagnosis and procedure codes, medications, as well as unstructured data including free-text notes.

Cohort Description and Size: We have preliminarily identified from STARR over 110,000 unique adult ophthalmology patients who have had over 1.5 million encounters in the ophthalmology department, of whom 10,514 patients had vision worse than 20/40 in their better eye on at least one occasion. Of these, 5612 patients had follow-up with our ophthalmology clinic of at least one year, during which 40.5% never improved to better than 20/40 over that year. We will use in our analyses this subset of 5612 low vision patients with at least one year of follow-up. The outcomes are split 40.5% with low visual acuity over a year (label = 1), and 59.5% who improved (label = 0).

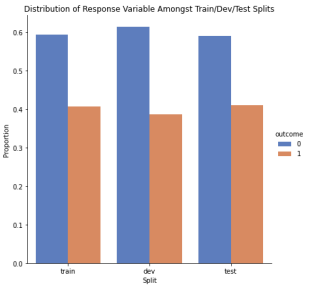

Figure 1: Distribution of Outcomes Per Split

Feature Engineering: Input features included both structured (numeric and coded) fields within the EHR, as well as unstructured data (free-text clinical notes). Structured fields include demographics, diagnosis and procedure codes, certain eye examination findings such as visual acuity (VA), intra-ocular pressure, and others, as well as medications. All numeric fields were standardized to mean 0 and variance 1 for input. For Boolean variables of which there are many (i.e., diagnoses), features with near zero variance were removed. After feature processing there were a total of 560 structured input features.

To preprocess free-text, we used two different approaches:

1. **Named Entity Recognition Tool:** We previously developed regex-based pipeline to identify different sections from clinical notes (such as history of present illness, past medical history, assessment/plan, etc.). We also used CLAMP, a clinical named entity recognition (NER) tool to extract vocabulary terms from free-text notes and map them to standard medical terminologies in the Unified Medical Language System (UMLS)[9]. Every extracted concept as well as its negated concept could be considered as an input feature to our deep learning models, in the form of Boolean variable with 1 for each patient for each concept if it appeared in their notes. Due to the large number of potential features, those with near zero variance were excluded. The addition of these NER features brought the total number of input features to 2223.

2. **Word2Vec:** Word embeddings were pre-trained on PubMed and EMR databases, generating 300-long vectors. We then took each patient note and mapped each word to their embedding vectors and averaged them. This yielded meaningful representations that learned similarity-aware representations in a lower, 300-dimensional space.

Evaluation: We randomly divided the dataset by patient, reserving 300 patients for the held-out test set for final evaluation, 300 patients for the development set, and the remainder for training. We report the standard evaluation metrics of AUROC, AUPRC, overall accuracy, as well as standard measures such as sensitivity, specificity, precision, and F1 score (at the default 50% probability threshold). In particular, we focus on optimizing AUROC as our single-evaluation metric since we want to maximize the true positive rate and minimize the false positive rate to reduce missed referral opportunities. To evaluate what the baseline human-level performance would be, we had a board-certified ophthalmologist (SYW) review a sample of 200 patient charts and predicted whether the vision would recover to better than 20/40 over one year or not, as shown in Table 1.

4 Methods

We pose the problem more formally as follows: given a set of m patients, X, where $X^{(i)} \in X$ is an n-long (random) vector of representing patient i's characteristics (consisting of structured and unstructured patient information) and a set of labels $Y \in \{0,1\}^m$ where:

- $Y^{(i)} = 1$: patient with a best corrected visual acuity of worse than 20/40 **remained worse** than 20/40 for the following year.
- $Y^{(i)} = 0$: patient with a best corrected visual acuity of worse than 20/40 **did not remain worse** than 20/40 for the following year.

4.1 Baseline Machine Learning Models

For these baseline models we included all structured and named entity recognition features.

Logistic Regression with L2 regularization: A simple logistic regression with 5-fold cross validation was applied to the structured and CLAMP NER free-text notes achieving train and test AUROCs of 0.94 and 0.70, respectively, with l2-norm regularization to reduce overfitting. This is a fitting baseline to understand whether an easy linear relationship can be fit to the classification task.

Tree-based Models: If that the learning task is non-linear or overfits, we take one step further to boostrap samples of features and training examples through random forest (bagging) and lightGBM (boosting). We start with a max-depth of 32 for each tree and 100 trees to avoid overfitting, and calibrate further based on training and dev AUROC values to optimize bias-variance tradeoff. After optimization, we orthogonalize our research strategy to focus on a)learning more complex estimators to fit the large feature space more accurately and/or b) reduce the space using embeddings or autoencoders.

4.2 Advanced Neural Network Models

Palliative Care Model: We next implemented a model previously described[4] which predicts mortality within 3-12 months in order to better deploy palliative care resources for patients with poor prognosis. This model was developed using Stanford EHR data consisting of structured feature inputs, and given the similarity of the prediction task and the underlying data, we replicated the architecture of this model, comprising 18 hidden layers of 512 dimensions each, with SeLU activation. We tried this model with both structured and structured+text features. We used Adam optimizer with default learning rate of 0.0001, batch size of 32, and performed early stopping when a 15% validation portion of the train set showed signs of overfitting.

DeepPatient Model: This is a state-of-the-art model developed by Miotto et al[5], which uses a 3-layer stacked denoising autoencoder with LDA and NegEx used for free-text representation trained on 700,000 Mount Sinai EHR (structured and free-text) data, which was in turn fed into a random forest classifier for disease classification. Our problem, with less observations, uses a 3-layer stacked denoising autoencoder architecture, with PubMed-trained word2vec embeddings for the notes, lightGBM for the classification task, and an Adam optimizer, outperforming the

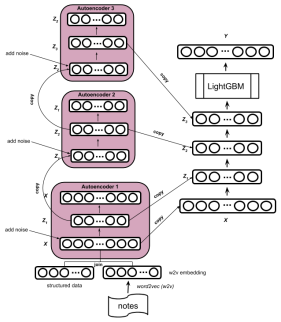

Figure 2: DeepPatient (Stacked Denoising Autoencoder) Architecture

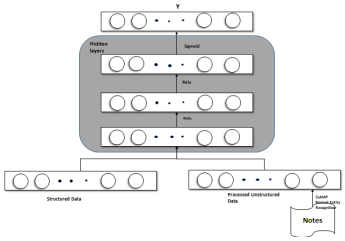

Figure 3: Optimized *ISeeU* Architecture

original implementation for our dataset. Due to a data-limited environment, the feature-learned lower-dimensional representations generated by the unsupervised autoencoder models and the similarity-aware word2vec embeddings learn the classification task much more easily to improve AUROC.

ICU Model: This model was developed originally to predict the mortality outcome of patients admitted to the ICU. We implemented a model inspired by Che et. Al (2016) in the form of a DNN that was implemented with two hidden layers double the size of the input layer (4446) and one prediction/output layer. Multiple strategies were used to reduce overfitting such as training for 250 epochs with early stopping criterion based on the validation loss. SGD with Adam36 was used along with a weight regularizer and dropout. This model was applied to the structured data in conjunction with the unstructured free text notes that were represented by the CLAMP named entity recognition model[10].

Optimized ICU Outcome Model: The model created by Torrez et al.[11] was utilized to predict the mortality outcome of children admitted to the ICU. Originally, the model described using a 3 layer DNN followed by a GRU gate, however the authors noted that the combination model did not do well to model the large amount of static features that are normally found in EHR data[11]. The trade-off in performance by the added GRU gate and a negligible rise in accuracy were not optimal as we aim to eventually build a product around our model. The final model utilized the layer sizes described in the paper for each of the three layers through a ReLU gate, replacing the final GRU gate with a sigmoid. We then utilized strategies to reduce overfitting similar to the Che et al. model described above; training for 250 epochs with early stopping criterion based on the loss on validation set, SGD with Adam36 as well as a weight regularizer and dropout[10]. In addition to these techniques we also manually tuned our model with the focus of maximizing the AUROC statistic, eventually achieving an AUROC of 0.76.

5 Results

Starting from Table 1, we see that our baseline machine learning models perform with similar accuracy and AUROC. Even with 0.001 l2-regularization, we achieve a train AUROC of 0.95 and test of 0.70, which meant that we are likely to have overfit to our test set. The tree-based models (RF and LGBM) have the same test AUROC, even with smaller tree depths and reducing the number of trees. The manual model, reviewed by Dr. Wang, performs higher recall in a sample of 200 patients, but machine learning methods already improve the accuracy by a lot. This means that we are predicting the negative class more strongly using machine learning methods.

As for Table 2, our deep learning algorithms show higher AUROCs than baselines. Our deepest model trained (palliative care) performed similarly to our baseline models. Despite regularization strategies, we believe that such a complex model on a small training set may have led to its baseline performance. Instead, we reduced the number of layers with the ICU model and the DeepPatient model, increasing test AUROC metrics to 0.75 and 0.76, respectively .

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV	F1
Vanilla LR	0.70	0.60	0.68	0.51	0.80	0.64	0.70	0.57
Random Forest	0.71	0.63	0.68	0.48	0.82	0.66	0.70	0.55
LightGBM	0.71	0.62	0.68	0.50	0.80	0.65	0.70	0.56
Manual Review	-	-	0.60	0.73	0.51	0.53	0.72	0.61

Table 1: Baseline Model Results

Model	AUROC	AUPRC	Accuracy	Sensitivty	Specificty	PPV	NPV	F1
PC (Struct)	0.70	0.58	0.65	0.54	0.72	0.57	0.69	0.55
PC (Struct + NER)	0.70	0.59	0.70	0.50	0.84	0.68	0.71	0.58
AE (Struct)	0.70	0.62	0.66	0.56	0.72	0.58	0.70	0.57
AE (Struct + NER)	0.71	0.63	0.68	0.54	0.77	0.62	0.71	0.58
ICU (Struct + NER)	0.73	0.70	0.70	0.62	0.76	0.62	0.70	0.62
DP (Struct + W2V)	0.75	0.69	0.72	0.55	0.83	0.69	0.72	0.61
ICUv2 (Struct + NER)	**0.76**	0.68	0.70	0.59	0.77	0.75	0.70	0.76

Table 2: Advanced Model Results

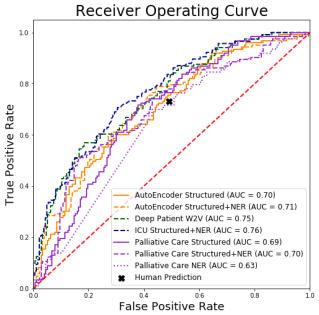

Figure 4: AUROC of Advanced Models

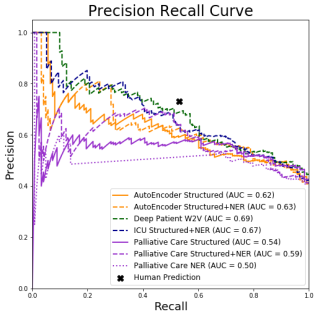

Figure 5: AUPRC of Advanced Models

6 Discussion/Conclusion

One of our best performing models used as feature inputs clinical notes with words mapped to word embeddings. This approach outperformed all other approaches using a combination of structured features and named entity recognition features from the notes, suggesting that future work to improve performance should focus on word embedding types of models. Among our other advanced neural network models included an architecture previously used to predict mortality for palliative care referrals, which was extremely deep (18 layers of 512 nodes each), and a model used in the ICU setting which was originally very wide(2 layers, 4446 nodes each) but was later optimized (3 layers, 100 nodes each). Between these models, the ICU model appeared to perform best.

We also learned that in data-limited environments, tree-based ensembling methods work best, leveraging the statistical power of bootstrapping to estimate the distribution of our data. Despite 100+ experiments on various models, we still cannot surpass an AUROC of 0.80, which implies that the true distribution has yet to be unseen. We also learned the power of unsupervised, feature learned representations and the ability of regularization techniques such as early stopping to choose the best state in which a model has reduced the loss. Lastly, we saw the strength of word2vec as a natural language representation of free-text, which formed the basis of deep learning models to learn sequential meaning.

7 Future Work

Given a small training set, one way to better infer the true data distribution is using transfer learning on both the structured and free-text data. Models such as BlueBERT[12], a BERT model trained on PubMed clinical notes, can side-step this data-limited constraint. This can also allow us to learn the sequential representation of word2vec embeddings via the power of the attention layers in transformers, instead of averaging out word embeddings per note as we do now.

8 Contributions

Thanks to Dr. Sophia Wang for defining the cohort, feature engineering, manual rating for human-level performance baseline, figures, and palliative care model. Sumit Singh has implemented the ICU outcome model, optimized ICU outcome model, and baseline feature engineering work. Samba Njie has performed baseline, autoencoder, and DeepPatient modeling work, exploratory analyses, added figures, and converted the report to LaTeX.

9 Appendix

9.1 Demographic Data Distributions

Below, we feature a few exploratory analyses performed on demographic data to provide an intuitive understanding of our cohort of study. Amongst Figures 6 and 7, which feature the distribution of gender and race across train, dev, and test splits, we see similar distributions across these splits. We are primarily interested in whether the dev and test splits are the same so we maximize our generalization potential for these models to model the real-world scenario when this tool is implemented.

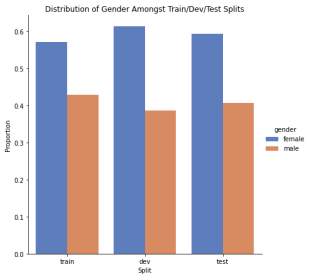

Figure 6: Distribution of Gender Per Split

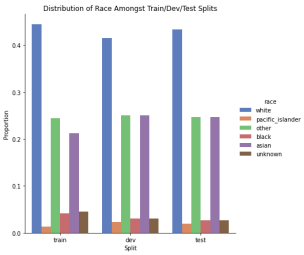

Figure 7: Distribution of Race Per Split

9.2 Exploratory Analysis on Feature Correlations

In Figure 8, we see a different visualization exercise featuring our feature space. Here, we observed how correlated our binary outcome variable is with each of the coded and numeric structured features to aid feature selection. Since our structured feature matrix is real-valued and our response variable binary, we sought to use a **Point-Biserial correlation measure** to estimate the correlation. This alleviates any normality assumptions of the data and specifically computes the correlation coefficient and p-values between binary and real-valued vectors. We see in Figure 8 that there are no meaningfully high correlations.

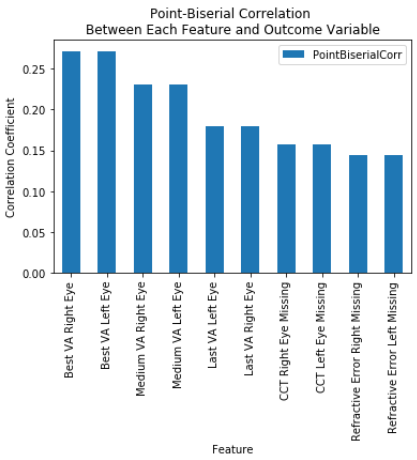

Figure 8: Feature Correlations with Outcome Response

9.3 Other Tried Models

Single Autoencoder Model: Given that our data lives in a high-dimensional space, we wanted to try a model that would reduce the dimensionality of the feature space and place similar patients closer together. This feature learning problem is solved by an autoencoder with a single layer. The

autoencoder network was trained on 2,223 input features and a 512-dimensional latent space with l2-regularization coefficient of 0.001, SeLU activation, and a binary cross-entropy loss with mini-batch size of 32 for 50 epochs and early stopping (patience = 5). Figure 9 summarizes the architecture.

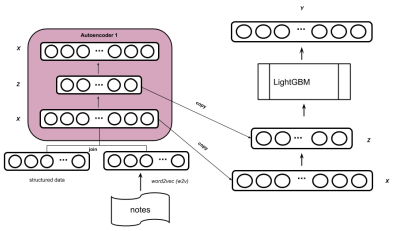

Figure 9: Single Autoencoder Architecture

9.4 ROC and Precision-Recall Curves of Baseline Models

For completeness, we also add on to the Figure 4 Advanced Neural Network AUROC and AUPRC curves by showing the performances of our baseline models. As you can see, the performances are similar, yet even the best performing machine learning models, here the LightGBM trained on structured only data with a 0.73 AUROC on structured data test set, does not outperform the best neural network architectures described in the Results section.

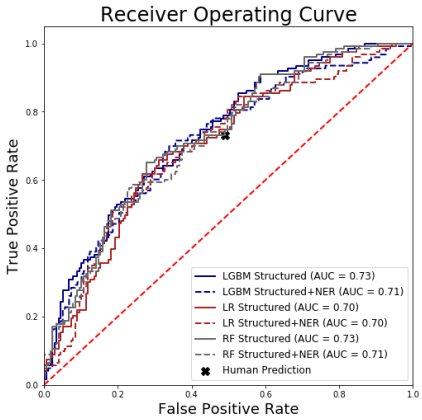

Figure 10: AUROC of Baseline Models

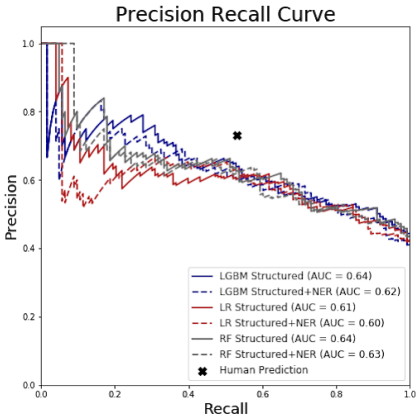

Figure 11: AUPRC of Baseline Models

9.5 Other NLP Techniques

Besides word2vec and CLAMP named entity recognition, topic modeling is also used to generate a set of features from the unstructured free-text clinical notes. We experimented with the use of **Latent Dirichlet Allocation**, a generative topic model where given a number of topics and a corpus of documents, estimates the distribution of topics for each document by using Dirichlet and Multinomial distribution as priors[13]. The Python package *gensim* was capitalized to implement LDA, as well as cleaning the corpus of stopwords. Bigrams and unigrams were initially generated and fed into the LDA model, and was trained on 20-40 topics. To evaluate this model, we did a "post-generative evaluation" by casting the features per document/patient note as an embedding and predicting the outcome variable just on this "LDA embedding" so to speak. It performed poorly, and so due to the interest of time, no further experimentation has been tried to tune or explore further topic modeling approaches so far.

References

[1] Fontenot JL, and Bona MD, and Kaleem MA, and et al. "Vision Rehabilitation Preferred Practice Pattern". In: *Ophthalmology* 125.1 (2018), pp. 228–278. DOI: 10.1016/j.ophtha.2017.09.030.

[2] Chan T., and Friedman DS, and Bradley C, and Massof R. "Estimates of Incidence and Prevalence of Visual Impairment, Low Vision, and Blindness in the United States". In: *JAMA Ophthalmol* 136.1 (2018), pp. 12–19. DOI: 10.1001/jamaophthalmol.2017.4655.

[3] Coker MA, Huisingh CE, McGwin G Jr, et al. "Rehabilitation Referral for Patients With Irreversible Vision Impairment Seen in a Public Safety-Net Eye Clinic". In: *JAMA Ophthalmol* 136.4 (2018), pp. 400–408. DOI: 10.1001/jamaophthalmol.2018.0241.

[4] Anand Avati and Kenneth Jung and Stephanie Harman and Lance Downing and Andrew Y. Ng and Nigam H. Shah. "Improving Palliative Care with Deep Learning". In: *CoRR* abs/1711.06402 (2017). arXiv: {1711.06402}. URL: %7Bhttp://arxiv.org/abs/1711.06402%7D.

[5] Riccardo Miotto, Li Li, and Brian Kidd. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". In: *Scientific Reports* 6 (May 2016), p. 26094. DOI: 10.1038/srep26094.

[6] Apostolova E, White HA, Morris PA, Eliason DA, Velez T. "Open Globe Injury Patient Identification in Warfare Clinical Notes". In: *AMIA Annu Symp Proc* 136.4 (2017), pp. 403–410. DOI: doi.org/10.1167/tvst.9.2.13.

[7] Peggy Peissig et al. "Importance of multi-modal approaches to effectively identify cataract cases from electronic health records". In: *Journal of the American Medical Informatics Association : JAMIA* 19 (Mar. 2012), pp. 225–34. DOI: 10.1136/amiajnl-2011-000456.

[8] Henry Lowe et al. "An Integrated Standards-Based Translational Research Informatics Platform. AMIA ... Annual Symposium proceedings / AMIA Symposium". In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2009 (Nov. 2009), pp. 391–5.

[9] Ergin Soysal et al. "CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines". In: *Journal of the American Medical Informatics Association* (Nov. 2017), ocx132. DOI: 10.1093/jamia/ocx132.

[10] Zhengping et al Che. "Interpretable Deep Models for ICU Outcome Prediction." In: *AMIA ... Annual Symposium proceedings* null (2016), pp. 371–380.

[11] William Caicedo-Torres and Jairo A. Gutiérrez. "ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU". In: *CoRR* abs/1901.08201 (2019). arXiv: 1901.08201. URL: http://arxiv.org/abs/1901.08201.

[12] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets". In: *CoRR* abs/1906.05474 (2019). arXiv: 1906.05474. URL: http://arxiv.org/abs/1906.05474.

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.