

Multimodal Ensemble Approach to Incorporate Various Types of Clinical Notes for Predicting Readmission

Bonggun Shin, Julien Hogan, Andrew B. Adams, Raymond J. Lynch, Rachel E. Patzer, and Jinho D. Choi[†]

Abstract—Electronic Health Records (EHRs) have been heavily used to predict various downstream clinical tasks such as readmission or mortality. One of the modalities in EHRs, clinical notes, has not been fully explored for these tasks due to its unstructured and inexplicable nature. Although recent advances in deep learning (DL) enables models to extract interpretable features from unstructured data, they often require a large amount of training data. However, many tasks in medical domains inherently consist of small sample data with lengthy documents; for a kidney transplant as an example, data from only a few thousand of patients are available and each patient’s document consists of a couple of millions of words in major hospitals. Thus, complex DL methods cannot be applied to these kind of domains. In this paper, we present a comprehensive ensemble model using vector space modeling and topic modeling. Our proposed model is evaluated on the readmission task of kidney transplant patients, and improves 0.0211 in terms of c-statistics from the previous state-of-the-art approach using structured data, while typical DL methods fails to beat this approach. The proposed architecture provides the interpretable score for each feature from both modalities, structured and unstructured data, which is shown to be meaningful through a physician’s evaluation.

I. INTRODUCTION

Predicting post-discharge rehospitalization is one of the major research areas in health-informatics, because the improvement of the prediction model could lead to better patient outcomes and efficient usages of medical resources [1, 2]. According to Jones et al. [3], about a half of surgical readmissions may be preventable, indicating potential positive effects of the prediction model on both patients and medical institutes. There have been many approaches to predict a readmission using electronic health records (EHRs), the majority of which are based only on structured data, such as demographic information, lab test values, and vital signs [4, 5, 6].

McAdams-Demarco et al. [7] found that a readmission of post-transplantation is a complex event consisting of various causes such as infections, rejection, and exacerbation of comorbidities. For this reason, multi-modal features are more desirable when designing a prediction model, implying

that there is likely room for improvement in previous models by incorporating these rich untapped data. Prior research has attempted to derive patterns from unstructured clinical notes, as the field of natural language processing (NLP) is advancing. These attempts enriched the model by extracting valuable information from unstructured data in predicting clinical outcomes [8], or identifying patient phenotype cohorts [9, 10, 11]. Moreover, recent successes in deep learning (DL) have encouraged use of deep neural networks in clinical NLP problems, and many of them have shown its superiority in various downstream clinical tasks [11, 12, 13]. However, these DL based models cannot be directly applied to many practical clinical problems due to peculiarities of these clinical datasets:

- Small sized samples - Many clinical downstream tasks consist of less than a few thousand data samples, which makes a model prone to be overfitted.
- Missing note types - Since not all types of notes are available for each patient, we need to impute the missing note modality, engendering an inevitable performance loss.
- Target-irrelevant sentences - Patients have multiple lengthy documents, consisting of a large number of words, however, only small portion of which are relevant to the target task. Therefore, important information tends to be diluted due to many non-informative words (or sentences).

To overcome these issues, we propose an ensemble framework that doesn’t require imputation of missing modalities, consisting of simple classifiers to circumvent overfitting, using vector space modeling and topic modeling as a feature to make it robust to long documents.

Our framework is evaluated on the Emory Kidney Transplant Dataset (EKTD), which task is to predict post-discharge rehospitalization at 30 days, being associated with poor outcomes of a patient. The dataset comprises 80 structured variables along with three different types of clinical notes. Our experiments show that the proposed framework outperforms the previous state-of-the-art approaches by 0.0211 in terms of c-statistics. Not only that, our research further adds interpretability to the data by effectively incorporating discriminating indices [14] to the trained model. To the best of our knowledge, this is the first time that an interpretable ensemble model is introduced for a readmission prediction problem with multiple modalities of input data.

B. Shin is with Department of Computer science, Emory University, Atlanta, GA 30303, USA. bonggun.shin@emory.edu

J. Hogan is with Department of Surgery, Emory University, Atlanta, GA 30303, USA. julien.hogan@emory.edu

A. B. Adams is with Department of Surgery, Emory University, Atlanta, GA 30303, USA. andrew.b.adams@emory.edu

R. J. Lynch is with Department of Surgery, Emory University, Atlanta, GA 30303, USA. ray.lynch@emoryhealthcare.org

R. E. Patzer is with Department of Surgery, Emory University, Atlanta, GA 30303, USA. rpatzer@emory.edu

J. D. Choi is with Department of Computer science, Emory University, Atlanta, GA 30303, USA. jinho.choi@emory.edu

[†]To whom correspondence should be addressed

Modality	Patients	Notes	Common Patients
Structured	2,060	N.A.	2,060
Consultations	2,282	21,854	1,354
Progress	2,444	202,296	1,415
Selection Conf. Ref.	2,843	3,512	2,033

TABLE I

STATISTICS OF STRUCTURED AND UNSTRUCTURED DATASET.

II. APPROACHES

A. Datasets

The proposed framework is evaluated on the Emory Kidney Transplant Dataset (EKTD), after Institutional Review Board approval. It consists of 2,060 patients of structured data (80 predictors), and various number of patients depending on the type of clinical notes as summarized in Table I. We utilized three common clinical notes: Consultations including all notes redacting for outpatient consultations during the year prior to transplantation, Progress including all notes written during the transplant admission and Selection Conference summarizing the result of the pre-transplant screening and justifying the waiting-list registration. As shown by the Table I, some patients don't have all types of notes, indicating the need of imputation, if a typical ensemble classifier is used. The target values, patient 30-days readmission outcomes are also recorded as a binary variable. Of the final population, 633 (30.7%) were rehospitalized after 30 days.

B. Baseline Model with Structured Dataset

Structured data include demographic and clinical characteristics of both the recipient and the donor, features related to labs results during the transplant admission. Many lab test values are time series, but only the last value at the discharge is used to form a fixed length feature. All non-binary categorical variables are transformed into dummy-binary variables, increasing the feature length into 92. Standardized normalization are applied to each feature using the mean and standard deviation calculated from the training samples. The baseline model is trained using only this structured features.

C. Feature Representation for Unstructured Dataset

EKTD contains three types of clinical notes. Notes written after the discharge are excluded for creating a fair model. These notes are preprocessed using the ELIT tokenizer¹, and all non-alphabetic tokens and typical English stopwords are removed from the notes. If a patient has multiple notes in a specific note type, those notes are merged into one document to transform it to a fixed sized feature vector. In this paper, vector space modeling and topic modeling are used as a vectorization method.

1) *Vector Space Modeling*: We use a popular vector space modeling, the term frequency-inverse document frequency (TF-IDF), since it can effectively filter out corpus specific stopwords, which are not covered by conventional English stopwords. TF-IDF model is fitted using only training data for each fold, and the test data is vectorized using the fitted TF-IDF model. The resulting TF-IDF vector has the size of $|V_n|$, where V_n is the number of vocabulary of the note n .

¹<https://github.com/elitcloud/elit>

2) *Topic Modeling*: Topic modeling is another way of document vectorization method. We use a topic distribution generated by Latent Dirichlet Allocation (LDA) [15] to represent each type of notes of a patient. As the successful clinical note processing work [16] suggested, we used 50 topics. We set the hyperparameter, $\alpha = \frac{5}{\text{numberTopic}}$ after trying various values on each validation set. A final topic distribution is drawn from a MCMC chain after trained 3,000 iterations. The resulting LDA vector has the size of the number of topics, 50.

D. Incorporating Unstructured Modalities

A naive way of incorporating unstructured modalities is a logistic regression with naively concatenated features (Fig. 1(a)), where each feature comes from different modalities; A structured feature ($x_s \in R^{92}$), TFIDF vectors ($x_{tfidf,n} \in R^{|V_n|}$), and LDA vectors ($x_{lda,n} \in R^{50}$), where n represents a note type. Note that this method requires imputation because all modalities should be concatenated together to form a fixed size vector. However, the proposed multi-modal approach, the averaged sigmoids method (Fig. 1(b)), doesn't require imputation of the missing modalities, because each model for each modality is trained separately. Predictions from these separate models are averaged into one final probability value. Therefore, if some of the modalities are missing, the proposed method just takes an average without those modalities.

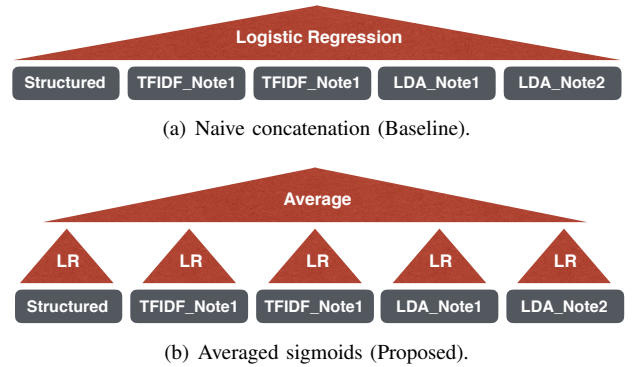


Fig. 1. Multimodal ensemble models.

E. Feature Importance

We apply the discriminative index (DI) [14] to each modality to identify corresponding key features. The DI algorithm simultaneously utilizes input features and associated coefficients to calculate the contribution score, WX. Shin et al. [14] showed that the scores calculated from both inputs and weights are more accurate than just considering weights. For example, even if a weight value is big, it could not be one of the important features, because the reason why the weight became big is that the corresponding input feature values might simply in a range of very small numbers. By comparing the WX scores of two different cohorts (negative and positive samples), proper feature impact scores can be calculated as described in Algorithm 1.

Algorithm 1: DI [14]

Input: X, Y, θ, c **Output:** Sorted Feature List

```

1  $\hat{X}^{True} = \text{avg}(X^{True});$ 
2  $\hat{X}^{False} = \text{avg}(X^{False});$ 
3  $WX^{True} = \theta^T \cdot \hat{X}^{True};$ 
4  $WX^{False} = \theta^T \cdot \hat{X}^{False};$ 
5 for  $k \in \{1, \dots, K\}$  do
6    $DI_k \leftarrow |WX_k^{True} - WX_k^{False}|$ 
7 return  $\text{argsort}(DI)$ 

```

III. EXPERIMENTS

We present the results of the proposed ensemble framework contrasting other frameworks and show the effectiveness of integration of multimodal features for a readmission prediction task. In addition, we provide the interpretability of the framework by showing the usefulness of the selected top-10 features of each modality. 5-fold cross validation was used to evaluate all approaches. For each fold, we held out a 20% of the patients as a test set. The remaining 80% of patients were used to vectorizations and training the models.

A. Training Details

The number of features of structured modality is 92, three topic modeling modalities are 50, and three vector space modeling modalities vary depending on the training data of each subfold. For instance, the first subfold finds the optimal vector dimensions as 30,275, 33,501, and 20,598 for Consultations, Progress, and Selection Conference, respectively.

B. Performance Comparison

The experiments are designed to validate that the proposed framework successfully ameliorates three hurdles of practical clinical problems (Section I) including EKTD. The first issue is overfitting. As discussed by Shin et. al [11], unless the model is dealing with large dataset, deep learning has no edge compared to a simple logistic regression model. Reflecting this lesson, we pick a logistic regression as a classifier for all frameworks. The second issue is possible performance degradation due to imputation of missing modalities. We contrast two ensemble models, one that requires imputation (Fig 1(a)), and another one that doesn't (Fig 1(b)) to show that how our framework (Fig 1(b)) effectively handles missing modalities. The last problem is information dilution caused by lengthy documents. We compare the proposing vectorization method to the recent advanced vector modelings, word vectors [17] and document vectors [18], which shows the proposed one better capture the useful information in a low signal to noise environment. For word vectors, we used the pre-trained biomedical word2vec [19], and document vectors are trained on each training set of each note type. We excluded convolutional [11] and recurrent neural network [20] based models, because the size of the model exceeds the memory capacity due to the huge number of tokens per one patients, ranging from one to ten millions.

Table II shows that the proposed ensemble framework, consisting of vectorspace modeling, topic modeling and averaged sigmoid ensemble classifier (Fig 1(b)), outperforms the other frameworks. This framework successfully integrates three types of clinical notes with structured dataset, improving 0.0211 compared to the structured only model. Moreover, 95% confidence interval of the proposed framework indicates that our framework consistently outperforms the baseline by more than 0.01 margin.

Method	Avg. c-stats	95% CI	Delta
Structured Only	0.6523	(0.6218, 0.6829)	-
Avg.W2V (Concat)	0.6561	(0.631, 0.682)	0.0038
Avg.W2V (Avg.Sig.)	0.6597	(0.635, 0.684)	0.0074
Doc2Vec (Concat)	0.6522	(0.624, 0.681)	-0.0001
Doc2Vec (Avg.Sig.)	0.6491	(0.624, 0.674)	-0.0032
TFIDF-LDA (Concat)	0.6669	(0.6488, 0.6850)	0.0146
TFIDF-LDA (Avg.Sig.)	0.6734	(0.6635, 0.6834)	0.0211

TABLE II

AVERAGES AND 95% CONF. INTERVALS OF C-STATS OF ALL FIVE FOLDS. TOP MOST ROW REPRESENTS THE SCORES OF THE BASELINE WITH ONLY STRUCTURED FEATURES. OTHERS ARE COMBINATIONS OF TWO MULTIMODAL ENSEMBLES WITH DIFFERENT VECTORIZATIONS.

C. Feature Analysis

Algorithm 1 is applied to all seven modalities, and Top-10 important features are presented in Figure 2. Top predictors from the structured data included mostly labs results, such as hemoglobin, albumin, and creatinine level which is a marker of the function of the transplanted kidney. The overall predictive accuracy of this model was low and consistent with previously published predictive models of 30-day readmission. Other important predictors are the quality of the immunological matching between the donor and the recipient (HLA_MISMATCH). The 6 models based on clinical notes captured relevant predictors of hospital readmission, including assorted patients comorbidities. All models' selected terms are related to diabetes or diabetes-related complications (pancreas, diabetes, insulin, and mellitus). Similarly, all LDA models reported topics related to cardio-vascular complications (carotid, coronary stent, and coronary artery bypass grafting) and digestive neoplasia (colonoscopy, polyp, and sigmoid).

Severity markers were also extracted as major predictors within the progress notes indicating the need for admission in the intensive care unit (ICU), respiratory failure (oxygen) or the need for intravenous medications to either lower (nicardipine) or increase (phenylephrine) blood pressure (mmHg). Of specific interest were the topics related to socio-economic status or social support. Indeed, if patients comorbidities are usually captured in classic structured databases, social-economic features are often poorly recorded. The need for a social evaluation of the patients expressed either in the selection conference notes (social team needed) or in the consultations as demonstrated by the presence of social workers names within the notes (Licensed Master Social Worker (LMSW), Licensed Clinical Social Worker (LCSW))

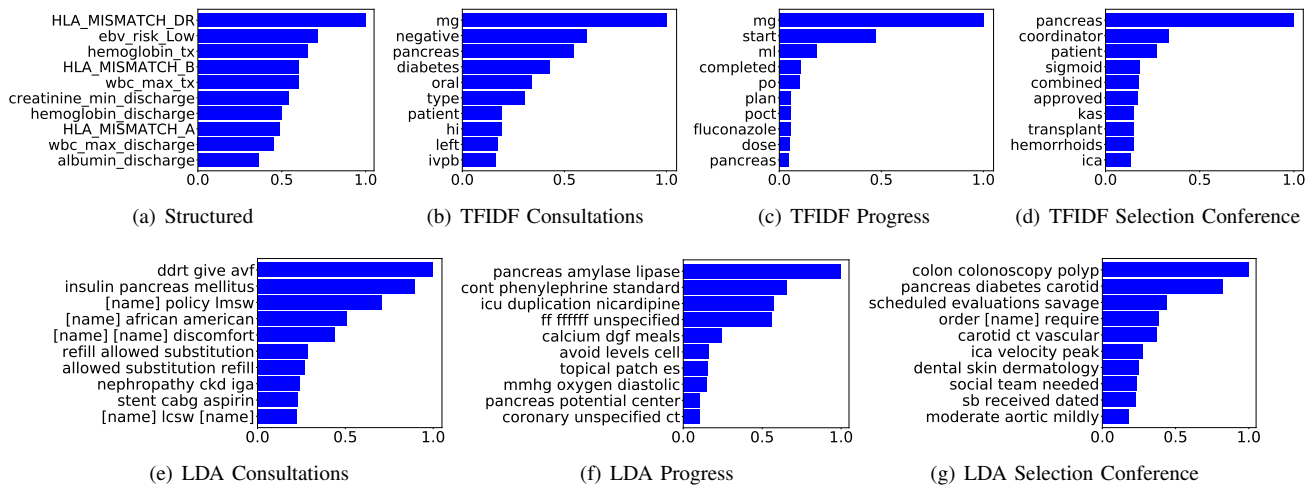


Fig. 2. Top 10 important features, which are min-max normalized. For LDA modalities, top three frequent words for each topic are listed on the y-axis. All names of a physician or a social worker are replced to “[name]”.

was included as a top predictive feature in 2 out of 3 LDA-based models. Similarly, topics related to medication delivery and adherence were also captured (refill allowed substitution). This finding is extremely interesting since much emphasis is currently made on the impact of adherence on clinical outcome and underline the potential of NLP in this field.

IV. CONCLUSIONS

This paper proposes a multi-modal ensemble framework with vector space and topic modeling features that effectively integrates structured and unstructured dataset for predicting readmissions at 30 days. Our experiments show that this framework not only adequately handles missing modality, but properly catches useful information from a very long document. In addition, we introduced a way to interpret the prediction results, which could potentially be valuable in medical actions. The physicians evaluation showed that the provided framework is meaningful in that it not only allows the extraction of both classic predictors previously reported in other studies, but also the predictive features covers fields that are usually poorly covered in structured databases such as socio-economic status, social support or medication delivery and adherence.

REFERENCES

- [1] S. F. Jencks, M. V. Williams, and E. A. Coleman, “Rehospitalizations among patients in the medicare fee-for-service program,” *New England Journal of Medicine*, vol. 360, no. 14, pp. 1418–1428, 2009.
- [2] M. Harhay, E. Lin, A. Pai, M. Harhay, A. Huverserian, A. Mussell, P. Abt, M. Levine, R. Bloom, J. Shea, *et al.*, “Early rehospitalization after kidney transplantation: assessing preventability and prognosis,” *American Journal of Transplantation*, vol. 13, pp. 327–335, 2013.
- [3] C. E. Jones, R. H. Hollis, T. S. Wahl, B. S. Oriel, K. M. Itani, M. S. Morris, and M. T. Hawa, “Transitional care interventions and hospital readmissions in surgical populations: a systematic review,” *The American Journal of Surgery*, vol. 212, no. 2, pp. 327–335, 2016.
- [4] M. F. Levy, L. Greene, M. A. Ramsay, L. W. Jennings, K. J. Ramsay, J. Meng, H. T. Hein, R. M. Goldstein, B. S. Husberg, T. A. Gonwa, *et al.*, “Readmission to the intensive care unit after liver transplantation,” *Critical care medicine*, vol. 29, pp. 779–786, 2016.
- [5] K. L. Covert, J. N. Fleming, C. Staino, J. P. Casale, K. M. Boyle, N. A. Pilch, H. B. Meadows, C. R. Mardis, J. W. McGillicuddy, S. Nadig, *et al.*, “Predicting and preventing readmissions in kidney transplant recipients,” *Clinical transplantation*, vol. 30, no. 7, pp. 779–786, 2016.
- [6] R. Leal, H. Pinto, A. Galvão, L. Rodrigues, L. Santos, C. Romãozinho, F. Macário, R. Alves, M. Campos, A. Mota, *et al.*, “Early rehospitalization post-kidney transplant due to infectious complications: Can we predict the patients at risk?” in *Transplantation proceedings*, vol. 49, no. 4, Elsevier, 2017, pp. 783–786.
- [7] M. McAdams-DeMarco, M. Grams, E. Hall, J. Coresh, and D. Segev, “Early hospital readmission after kidney transplantation: patient and center-level associations,” *American Journal of Transplantation*, vol. 12, no. 12, pp. 3283–3288, 2012.
- [8] M. Staff, “Can data extraction from general practitioners electronic records be used to predict clinical outcomes for patients with type 2 diabetes?” *Journal of Innovation in Health Informatics*, vol. 20, pp. 3621–3627, 2017.
- [9] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, “A review of approaches to identifying patient phenotype cohorts using electronic health records,” *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 3621–3627, 2014.
- [10] S.-M. Zhou, M. A. Rahman, M. Atkinson, and S. Brophy, “Mining textual data from primary healthcare records: Automatic identification of patient phenotype cohorts,” in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 3621–3627.
- [11] B. Shin, F. H. Chokshi, T. Lee, and J. D. Choi, “Classification of radiology reports using neural attention models,” *Neural Networks (IJCNN), 2014 International Joint Conference on*, 2017.
- [12] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, “Predictive risk modelling for early hospital readmission of patients with diabetes in india,” *International Journal of Diabetes in Developing Countries*, vol. 36, no. 4, pp. 519–528, 2016.
- [13] E. Craig, C. Arias, and D. Gillman, “Predicting readmission risk from doctors’ notes,” *NIPS 2017 Workshop on ML for Health*, 2017.
- [14] B. Shin, S. Park, , Y. Choi, K. Kang, and K. Kang, “Wx: a neural network-based feature selection algorithm for next-generation sequencing data,” *bioRxiv*, p. 221911, 2017.
- [15] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, 2003.
- [16] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, “Unfolding physiological state: Mortality modelling in intensive care units,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 75–84.
- [17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013.
- [18] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *ICML*, 2014, pp. 1188–1196.
- [19] S. Moen and T. S. S. Ananiadou, “Distributional semantics resources for biomedical text processing,” in *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, 2013, pp. 39–43.
- [20] P. Grnarova, F. Schmidt, S. L. Hyland, and C. Eickhoff, “Neural document embeddings for intensive care patient mortality prediction,” *arXiv preprint arXiv:1612.00467*, 2016.