Predicting Gastrointestinal Bleeding Events from Multimodal In-Hospital Electronic Health Records Using Deep Fusion Networks

Chen-Ying Hung, Ching-Heng Lin, Chi-Sen Chang, Jeng-Lin Li, and Chi-Chun Lee

***Abstract*—Applying machine learning (ML) methods on electronic health records (EHRs) that accurately predict the occurrence of a variety of diseases or complications related to medications can contribute to improve healthcare quality. EHRs by nature contain multiple modalities of clinical data from heterogeneous sources that require proper fusion strategy. The deep neural network (DNN) approach, which possesses the ability to learn classification and feature representation, is well-suited to be employed in this context. In this study, we collect a large in-hospital EHR database to develop analytics in predicting 1-year gastrointestinal (GI) bleeding hospitalizations for patients taking anticoagulants or antiplatelet drugs. A total of 815,499 records (16,757 unique patients) are used in this study with three different available EHR modalities (disease diagnoses, medications usage, and laboratory testing measurements). We compare the performances of 4 deep multimodal fusion models and other ML approaches. NNs result in higher prediction performances compare to random forest (RF), gradient boosting decision tree (GBDT), and logistic regression (LR) approaches. We further demonstrate that deep multimodal NNs with early fusion can obtain the best GI bleeding predictive power (area under the receiver operator curve [AUROC] 0.876), which is significantly better than the HAS-BLED score (AUROC 0.668).**

I. Introduction

In-hospital EHRs are valuable data sources of the existing healthcare system, and ML techniques are a set of highly effective data-driven predictive algorithms capable of learning powerful hidden relationship between the desired outcome and a variety of clinical variables derived from large databases. The longitudinal nature of EHR along with its variety in the health-related information offers an enormous possibility in deploying ML techniques for clinical practices. These routinely collected in-hospital EHRs could provide high performing predictive analytics for health applications such as evaluation of treatment efficacy or complications. It could be applied for a wide range of diseases that would positively impact a patient's clinical outcomes directly.

Anticoagulants and antiplatelet drugs are usually used in the prevention and treatment of ischemic stroke or heart diseases. The use of these drugs and their complications (such as GI bleeding) are likely to increase as the population ages. Hence, accurately assessing the risks of GI bleeding occurrences in patients taking these medications is important as it would help physicians to properly balance the trade-off between the benefit of treatments and the risk of bleeding. All of the current clinical scores used for predicting GI bleeding events, such as the QBleed algorithms (using 21 variables with an AUROC 0.77 [1]), the HAS-BLED score (using 9 variables with an AUROC 0.72 [2]) or a recently developed model by Shimomura et al. (using 5 variables with an AUROC 0.65 [3]), showed only moderate predictive power. In this work, our aim is to develop GI bleeding predictive algorithms from in-hospital EHRs using techniques of ML and DNN.

In-hospital EHRs cover data from multiple information domains, and the types of these records vary in structures. Additional challenge is that this large amount of structured and semi-structured data produces thousands of potential predictive variables. The use of DNNs, which automatically learn complex feature relationships at multiple levels of abstraction [4], allows us to address challenges of modeling many variables simultaneously to obtain high AUROC performances. For example, researchers have applied deep learning methods to derive predictive algorithms with accuracy beyond current clinical scores in applications such as prediction of stroke [5,6,7] and inpatient mortality [8]. In our previous work, we have demonstrated that using DNNs on electronic medical claims can accurately predict stroke events, reaching a state-of-the art AUROC 0.92 [7].

Furthermore, an advantage of DNN technique is its ability to integrate multimodal data sources of heterogeneous types that can jointly be optimized to achieve further improved prediction accuracy. Specifically, there are two strategies commonly used for multimodal fusion: early fusion and late fusion. The early fusion approach learns relationships between features and class discrimination to model the interaction between modalities. On the other hand, late fusion handles these modalities as independent streams until the end [9].

In this study, we apply DNN with multimodal fusion strategy on a large in-hospital EHR database to derive analytics for GI bleeding hospitalizations prediction, specifically targeting for patients that have previously been treated with anticoagulants or antiplatelet drugs. The performances of different multimodal fusion strategies and other conventional ML algorithms are compared in this work. Our results show the benefit of early fusion approach compared to late fusion method, and the deep multimodal fusion network outperform single modal strategy. Finally, DNNs with early fusion strategy are capable of obtaining the highest predictive accuracy (AUROC 0.876) for 1-year GI bleeding events prediction.

CCL is the corresponding author for this work. He is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan. (phone: +88635162439. e-mail: cclee@ee.nthu.edu.tw)

CCL, CYH, JLL are with the MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

CYH, JLL are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan.

CYH is with the Department of Internal Medicine, Taipei Veterans General Hospital, Hsinchu Branch, Hsinchu, Taiwan.

CHL is with the Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan.

CSC is with the Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan.

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

TABLE 1. INFORMATION OF THE STUDY COHORT

	All records	Records in training time period	Records in testing time period
Time period	2006-2015	2007-2012	2014
No of patients	16,757	14,406	11,640
No of records	815,499	551,156	127,929
No of records with GI bleeding events	4,111	2,623	720

II. METHODS

A. Database and study population

The database for this study is extracted from the EHRs of Taichung Veterans General Hospital. All patients who used anticoagulants or antiplatelet drugs (Anatomical Therapeutic Chemical code: B01AA03, B01AE07, B01AF02, B01AF01, B01AC06, B01AC04, B01AC07, B01AC05, B01AC23, B01AC24) more than 4 months during 2006 to 2015 were identified from the EHR system of the hospital. The database contains de-identified EHR data from 23,631 patients treated in the inpatient and outpatient departments (with a total of 46,389 inpatient records and 5,505,898 outpatient records). These records contain information of patient's demographics, disease diagnoses, medications use, and laboratory testing measurements. The Institutional Review Board of Taichung Veterans General Hospital institutional approved the study.

We design a cohort for predicting 1-year GI bleeding hospitalizations. Patients aged 18 to 90 years are identified from the outpatient database. Patients are not eligible for enrollment if they had any types of GI bleeding in the past 1 year before enrollment or had an insufficient follow-up time period (1 year before enrollment and 1 year follow-up period after enrollment). We further remove records that have inadequate numbers of available clinical variables. Following this exclusion criteria, our final dataset includes a total of 16,757 patients (815,499 records, see Table 1). In order to perform 5-fold cross validation properly, these patients are randomly divided into 5 groups. In order to validate the real-world use of these ML models, records of each group are further divided into 2 subgroups with non-overlapping time periods: records in training time period (2007-2012, a total of 551,156 records) and records in testing time period (2014, a total of 11,640 records).

B. Outcome definition

In this work, our aim is to predict the risks of GI bleeding occurrences in patients using anticoagulants or antiplatelet drugs. To ensure the diagnostic validity, the outcome event is defined as any GI bleeding (ICD-9-CM code: 530.7, 531.0, 531.2, 531.4, 531.6, 532.0, 532.2, 532.4, 532.6, 533.0, 533.2, 533.4, 533.6, 534.0, 534.2, 534.4, 534.6, 535.01, 535.11, 535.21, 535.31, 535.41, 535.51, 535.61, 535.71, 537.83, 537.84, 562.02, 562.03, 562.13, 569.3, 569.85, 578) recorded in the hospital discharge diagnoses in the inpatient database.

C. Feature engineering

We utilize data from the EHRs within 1 years prior to our enrollment to generate features as input to DNN. We first gather the following measurements from the records of an individual patient at the enrollment time:

TABLE 2. A TOTAL OF 4,050 FEATURES EXTRACTED

Measurement Dimension	Temporal Dimension	No of Features
Demographics: gender and age		2
Disease diagnosis: 914 in total	In past 1 month In past 3 months In past 6 months In past 1 year (4 in total)	3,656
Medication use: 18 in total		72
Laboratory biomarker: 16 in total		64 mean, 64 median, 64 standard deviation, 64 maximum, and 64 minimum values

- Demographic measurements: gender and age.
- Disease diagnosis measurements: A list of diagnoses is classified by the ICD-10-CM codes and mapped into binary values (914 in total). In the records, the diagnoses of diseases were coded using the ICD-9-CM code. We convert the ICD-9-CM code to ICD-10-CM code using the code-converting sheet provided by the National Health Insurance Bureau as previously done in our prior works [5,6,7]
- Medication usage measurements: A list of relevant medications (which may cause or prevent GI bleeding, e.g. proton-pump inhibitors, steroid, and painkillers) is mapped into binary values (18 in total).
- Laboratory biomarker measurements: A total of 16 laboratory measurements of widely used biomarkers are recorded as continuous values.

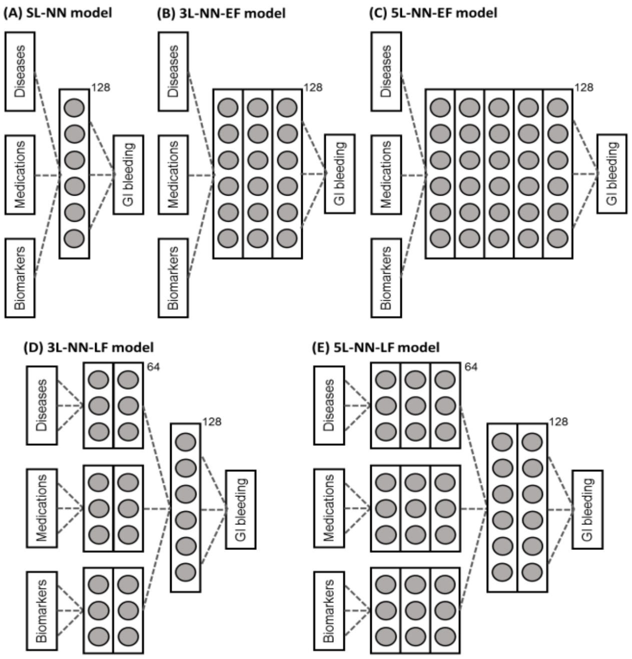
In order to generate the final feature vector that captures both the clinical measurements and temporal information, we utilize time stamp of these measurements. We extract a total of 4,050 features from the dataset (Table 2). These features can be abstracted as combinations of the measurement dimension and the temporal dimension. The temporal dimension that we use covers 4 time periods (1 month, 3 months, 6 months and 1 year). For biomarker measurements, we fill in a normal laboratory value in each time period (to avoid missing value) before calculating the mean, median, standard deviation, maximum, and minimum values within the selected time period (for example, mean value of hemoglobin levels in the past 1 year). In the medication use measurements, we compute the total number of specific medication classes recorded during these time periods. In the disease diagnosis measurements, we use the total number of times that a specific diagnosis is made during these time periods. We additionally perform feature selection to identify the most discriminative features as a data preprocessing before training our models. We use simple Pearson correlation method to select the most relevant 8, 16, 32, and 64 features to perform experiments.

D. Single modal architecture

In each measurement domain (disease, medication, and biomarker), we compare the performances of 3 ML approaches (using the scikit-learn version 0.18.0 packages) with 3 NN models (using the Keras toolbox):

- Random Forest (RF): using 100 trees with Gini index as the criteria for learning the tree splitting.

Figure 1. The structure of SL-NN and deep multimodal fusion models.



- Gradient Boosted Decision Tree (GBDT): using 100 boosted decision trees with binomial loss function.
- Logistic Regression (LR): using L2-regularization with strength 1.0.
- Single hidden Layer NN (SL-NN, Figure 1A): using a multilayer feed-forward perceptron with one hidden layer. The number of neurons per hidden layer is 128, and hyperbolic tangent is used as the activation function. We use dropout with a retention rate of 50% for hidden units. During the training process, the parameters of the model are randomly initialized. The final layer is trained to minimize cross entropy loss between the output and the true labels.
- 3-Layer NN (3L-NN): with 3 fully connected hidden layers. Other architecture is the same as SL-NN.
- 5-Layer NN (5L-NN): with 5 fully connected hidden layers. Other architecture is the same as SL-NN.

E. Multimodal fusion strategies

Early Fusion: The extracted features are combined into a single representation. This fusion scheme integrates unimodal features representations in the initial hidden layers of NN.

- 3-Layer NN with Early Fusion (3L-NN-EF, Figure 1B): with three 128-dimensional hidden layers. Other architecture is the same as SL-NN.
- 5-Layer NN with Early Fusion (5L-NN-EF, Figure 1C): with five 128-dimensional hidden layers. Other architecture is the same as SL-NN.

Late Fusion: The late fusion approaches learn representation separately from each unimodal features, and then combine these learned unimodal representations into a multimodal representation at the layers prior to the final softmax output.

TABLE 3: THE PERFORMANCES (AUROC) OF MODELS TRAINING WITH DISEASE DOMAIN INFORMATION

No of features	8	16	32	64
RF	0.565	0.544	0.608	0.565
GBDT	0.537	0.534	0.474	0.552
LR	0.659	0.761	0.758	0.769
SL-NN	0.687	0.765	0.768	0.789
3L-NN	***0.700***	***0.766***	***0.772***	***0.793***
5L-NN	***0.751***	***0.766***	***0.772***	***0.796***

- 3-Layer NN with Late Fusion (3L-NN-LF, Figure 1D): using a neural network with two 64-dimensional hidden layers for each domain features, follow-by one fully connected 128-dimensional hidden layer.
- 5-Layer NN with Late Fusion (5L-NN-LF, Figure 1E): using a neural network with three 64-dimensional hidden layers for each domain features, follow-by two layers of fully connected 128-dimensional layer units.

Experimental procedures

In this study, we examine our GI bleeding prediction task performance of each model using 5-fold subject independent cross validation. In each cross validation fold, records (in training time period) from 80% of patients are used as training sets and records (in testing time period) from the rest 20% as testing sets. In order to speed up the training process, we apply a simple normalization approach by scaling the feature values to a range between 0 and 1. Down-sampling is performed to guarantee an almost identical class distribution between bleeding and non-bleeding cases. The optimization algorithm used to train the network is based on RMSprop. We also conduct additional experiments by increasing the selected features amount (from 8, 16, 32, to 64 features) to examine the effect of the number of features have on our proposed approaches. We report the AUROC as the measure of performances of these algorithms.

III. RESULTS

Our study includes a total of 16,757 unique patients with 815,499 records. The average age is 65.6 years and 59.2% records come from male patients. A total of 4,111 records have 1-year GI bleeding hospitalization events. The event rate is 0.50% (4,111/815,499) per record. Table 1 shows the numbers of patients and records in training and testing time periods. There are 551,156 records in the training time period (2007-2012), and 127,929 records in the testing time period (2014). The event rate is 0.47% (2,623/551,156) and 0.56% (720/127,929) per record in the training time period and testing time period, respectively. A total of 4,050 features are generated from the datasets (3,656 features for disease domain, 72 features for medication domain, and 320 for biomarker domain, see Table 2). After feature selection, we select 8, 16, 32, and 64 features for the development of models.

The performances of 6 single modal approaches (RF, GBDT, LR, SL-NN, 3L-NN and 5L-NN) are compared with different number of features used. Table 3 shows the performances of these models trained with disease domain. The 3L-NN and 5L-NN obtain the highest AUROCs when training with 64 features (0.793 and 0.796, respectively).

TABLE 4: THE PERFORMANCES (AUROC) OF MODELS TRAINING WITH MEDICATION DOMAIN INFORMATION

No of features	8	16	32	64
RF	0.593	0.591	0.643	0.643
GBDT	0.672	0.672	0.695	0.722
LR	0.685	0.708	0.716	0.719
SL-NN	0.727	0.750	***0.762***	***0.744***
3L-NN	***0.730***	***0.755***	***0.761***	***0.755***
5L-NN	***0.729***	***0.750***	0.755	0.743

Table 4 shows the performances of models training with medication domain information. The 3L-NN and 5L-NN have higher AUROCs than other models while training with 8 and 16 features. Meanwhile, the SL-NN and 3L-NN achieve the highest AUROCs while training with 32 features (0.762 and 0.761). Table 5 shows the performances of models training with biomarker domain information. The 3L-NN and 5L-NN obtain the highest AUROCs while training with 16 features (0.853 and 0.854). Overall, NNs obtain higher AUROC values compared to RF, GBDT and LR. The 3L-DNN has the highest predictive performance, followed by the 5L-DNN, and SL-NN methods. These findings also indicate that biomarker domain provides the most information for GI bleeding prediction than disease or medication domains.

Table 6 shows the disease prediction performances of models training with different multimodal fusion strategies (disease, medication, and biomarker domains) when using different number of features. Overall, the early fusion models (3L-NN-EF and 5L-NN-EF) have higher AUROCs than other approaches. The 3L-NN-EF and 5L-NN-EF models obtain the highest AUROCs while training with 16 features from each domain (0.872 and 0.876), which are significantly higher than the HAS-BLED score (AUROC 0.668) or model proposed by Shimomura (AUROC 0.633) in our testing datasets. Early fusion is slightly better than late fusion approach. Of noted, most conventional ML models (GBDT, LR, and NNs) show a degradation when using too many correlated features (i.e., a decrease in accuracy when training with 64 features).

IV. DISCUSSION

In summary, we demonstrate that DNNs with early fusion technique outperform late fusion and other ML approaches. An encouraging AUROC of 0.876 is achieved by using 5L-NN-EF. Our results also show that performances of NNs are superior to that of RF, GBDT, and LR in both single modal and multimodal condition. Using more feature dimensions does not obtain a higher AUROC performance in this study. Interactions between feature factors may limit the performance of these ML models. We demonstrate that DNNs can leverage implicit correlations among features in different modalities in EHRs. The disadvantage of the late fusion approach may result from loss of information when learning each feature representation separately. In the task of predicting GI bleeding events, the information of three domains are complementary to each other. These underlying relationship between these modalities may play a role in the overall improvement of the outcome prediction.

TABLE 5: THE PERFORMANCES (AUROC) OF MODELS TRAINING WITH BIOMARKER DOMAIN INFORMATION

No of features	8	16	32	64
RF	0.734	0.675	0.710	0.639
GBDT	0.762	0.761	0.693	0.673
LR	0.848	0.847	0.838	0.823
SL-NN	0.846	0.848	0.839	0.839
3L-NN	***0.850***	***0.853***	***0.843***	***0.844***
5L-NN	***0.854***	***0.853***	***0.845***	***0.846***

TABLE 6: THE PERFORMANCES (AUROC) OF MODELS TRAINING WITH MULTIMODAL INFORMATION

No of features in each domain	8	16	32	64
RF	0.653	0.601	0.666	0.713
GBDT	0.670	0.697	0.747	0.743
LR	0.727	0.808	0.812	0.765
SL-NN	0.834	0.868	0.866	0.864
3L-NN-EF	***0.855***	***0.872***	***0.871***	***0.865***
5L-NN-EF	***0.864***	***0.876***	***0.873***	***0.865***
3L-NN-LF	0.850	0.867	0.867	0.855
5L-NN-LF	0.854	0.860	0.862	0.853

V. CONCLUSIONS

In this evaluation of applying ML on in-hospital EHRs, algorithms based on DNNs with early fusion achieve high AUROCs for prediction of 1-year GI bleeding hospitalizations and outperform the commonly used HAS-BLED score. Further prospective research is necessary to understand the potential impact of our algorithms on healthcare quality.

REFERENCES

[1] J. Hippisley-Cox, and C. Coupland, "Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores," BMJ, vol. 349, pp. g4606, July 28, 2014.

[2] R. Pisters, D.A. Lane, R. Nieuwlaat, C.B. de Vos, H.J. Crijns, et al., "A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey," Chest, vol. 138, pp. 1093-100, Nov. 2010.

[3] A. Shimomura, N. Nagata, T. Shimbo, T. Sakurai, S. Moriyasu, et al., "New predictive model for acute gastrointestinal bleeding in patients taking oral anticoagulants: A cohort study," J Gastroenterol Hepatol, vol. 33, pp. 164-71, Jan. 2018.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-44, May 28, 2015.

[5] C.Y. Hung, W.C. Chen, P.T. Lai, C.H. Lin, and C.C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," Conf Proc IEEE Eng Med Biol Soc, pp. 3110-3, July 2017.

[6] C.Y. Hung, C.H. Lin, and C.C. Lee, "Improving young stroke prediction by learning with active data augmenter in a large-scale electronic medical claims database," Conf Proc IEEE Eng Med Biol Soc, pp. 5362-5, July 2018.

[7] C.Y. Hung, C.H. Lin, T.H. Lan, G.S. Peng, and C.C. Lee, "Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database," PLoS One, vol. 14, pp. e0213007, Mar 13, 2019.

[8] R. Alvin, O. Eyal, C. Kai, M.D. Andrew, H. Nissan, et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 1, pp. 18, 2018.

[9] G.M.S. Cees, W. Marcel, W.M.S. Arnold, "Early versus late fusion in semantic video analysis," Multimedia'05, pp. 399-402, Nov. 6, 2005.