# A Random Under-sampled Deep Architecture with Medical Event Embedding: Highly Imbalanced Rare Disease Classification with EHR Data

**Yan Hu**[1], **Feng Chen**[2], **Yong Cai**[3], **and Yilian Yuan**[4]
[12]Department of Computer Science, State University of New York at Albany, Albany, NY, US
[34]Department of Advanced Analytic, IQVIA, Plymouth Meeting, PA, US

**Abstract**— *Rare diseases has very low prevalence rate. Electronic Health Records (EHRs), capture and integrate longitudinal patient health information, which includes patient demographics, medications, laboratory test results, etc. Rare disease classification with EHR data, which leaves us two tasks, imbalanced dataset classification and data analytic with EHRs. In this paper, we present a random under-sampled convolutional neural network (RUS-CNN) with medical event embedding to solve these two problems. In order to effectively and precisely extract the medical events that differentiate positive and negative patients, we also incorporate variances sorting with RUS-CNN (vRUS-CNN). Finally, our proposed approach is validated on a real world dataset. Experimental output demonstrates our methods achieve the best results compared to other 3 state-of-the-art baseline techniques.*

**Keywords:** EHRs, imbalanced classification, random under-sample, deep framework, medical event embedding

## 1. Introduction

Electronic Health Records (EHRs), capture and integrate longitudinal patient health information related to all aspects of care systematically. Included in this information are patient demographics, medications, laboratory test results, radiology imaging reports, clinical notes, etc.[1]. The EHRs have the ability to generate a complete record of a clinical patient encounter, as well as supporting other care-related activities, including evidence-based decision support, quality management, and outcomes reporting [2].

As the major carrier for conducting Data-Driven Healthcare (DDH), which is defined as usage of available medical big data to provide the best and most personalized care [3], there has been a remarkable upsurge in adoption of EHRs over the past several years. In particular, the patient data contained in EHR systems has been used for medical concept extraction [4], patient trajectory modeling [5], disease inference [6] [7], etc.

A rare disease, also known as orphan disease, has very low prevalence rate that affects only a small percentage of population. The Rare Disease Act 2002 defines rare disease to be less than 200,000 patients, or 1 in 1,500 [8]. Due to this nature, predicting or classifying rare diseases from an extremely imbalanced dataset is a challenging task. This problem is called imbalanced learning in data mining field [9].

Imbalanced rare disease classification with EHR data, which leaves us two main tasks to be resolved, imbalanced datasets classification and data analytic with patient EHRs.

A data set is class-imbalanced if one class contains significantly more samples than the other. For many disease categories, the imbalance rate ranges between 0.01-29.1% , which is the percent of the data samples that belong to the active class [21]. In such cases, it is hard to create an appropriate testing and training data sets, given that most classifiers are built with the assumption that the test data is drawn from the same distribution as the training data [11].

There are lots of challenges to do data analytic with patient Electronic Health Records (EHRs), to list a few:

- **High Dimension**. There are 14,025 different diagnosis codes and 3,824 procedure codes in terms of International Classification of Diseases $9^{th}$ Version (ICD-9) in patient EHRs. And, these codes interact with each other [12].
- **Sequentiality**. Longitudinal time stamped Patient EHRs reveal important information over time. The temporal EHRs may tell patient disease conditions, especially in chronic diseases development [13].
- **Sparsity**. One major challenge or working with EHR is sparsity, as shown in Fig.1 and Table I. For example, patients will have EHRs recorded only if they paid visits to clinical facilities. But, the patients are more likely to visit clinical sites when they are severely sick and/or need intensive monitoring. Other reasons, such as recording mistake, patient relocation, etc. lead to data missing as well [14].

In this paper, we select identifying Hereditary Angioedema (HAE) disease as our empirical application. HAE is a very rare and potentially life-threatening genetic condition that occurs in about 1 in 10,000 to 1 in 50,000 people [15], [16], [17]. Because of the low prevalence, physicians have rarely encountered HAE patients. On top of that, HAE attacks also resemble other forms of angioedema. These two facts make HAE hard to diagnose, which may cause late

Table 1: List of ICD-9 codes and Definition assigned to a specific patient from 1/1/2010 to 7/31/2015

| Event No. | ICD-9 code and Definition |
|---|---|
| 1 | 49390: Asthma NOS |
| 2 | 7862: Cough |
| 3 | 2512: Hypoglycemia NOS |
| 4 | 28959: Spleen disease NEC |
| 5 | 7802: Syncope and collapse |
| ... | ... |
| 90 | 3004: Dysthymic disorder |
| 91 | 4019: Hypertension NOS |

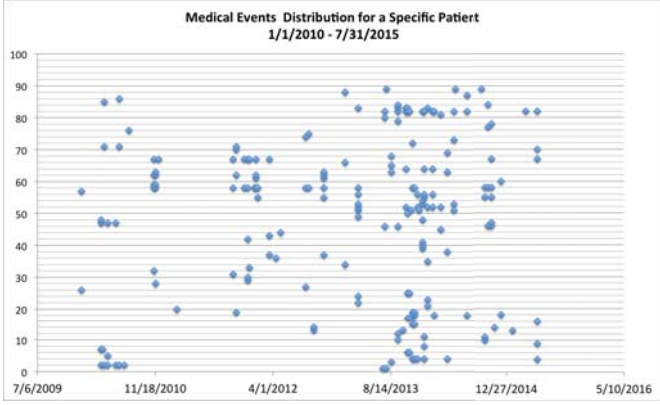


Fig. 1: An illustration of a patient EHRs. The horizontal axis represents the date, here from 1/1/2010 to 7/31/2015. The vertical axis corresponds to different diagnosis in terms of ICD-9 codes. A blue diamond indicates the specific medical event is diagnosed for this patient at the corresponding day.

or missed diagnosis, giving rise to incorrect treatment or unnecessary surgical intervention[18]. Given the low patient number and high cost of bringing new product into the market, it is essential for the pharmaceutical companies to develop budget friendly and efficient marketing approaches. Our objective in this application is to identify HAE patients effectively and efficiently using patient historical prescription and diagnosis information.

## 2. Related Work

As mentioned in the previous section, there are two tasks in this work, imbalanced datasets classification and data analytic with patient EHRs. Until the last few years, there are quite a few works on both problems.

### 2.1 Imbalanced Data Classification

Research on the imbalance data classification problem is critical in both data mining and machine learning. Two observations account for this point: (1) the class imbalance problem is pervasive in a large number of domains of great importance in data mining community; and (2) traditional classiïňĄcation learning systems are demonstrated to be inadequate while dealing with the class imbalance problem.

#### 2.1.1 Examples of Application Domains

The significant diïňČculty of the imbalance classification problem and its frequent occurrence in practical applications have attracted a lot of research interest. The following examples illustrate some cases.

- **Medical Diagnosis**. Data mining techniques applied on clinical databases, attempt to discover relationships and patterns among clinical and pathological data to understand the progression and features of certain diseases [19] . In the clinical databases, disease cases are fairly rare as compared with the normal populations. The discovered knowledge can be used for early diagnosis [20] [21] [22].
- **Network Intrusion Detection**. One of the problems faced in the development of Network Intrusion Detection Systems (NIDS) is that the datasets used in the construction of classifiers are typically imbalanced. This is because the classification categories do not have relatively equal representation in the datasets [23] [24].
- **Modern Manufacturing Plants**. For example, an alarm system is set up when ïňĆaw patterns are detected.The number of available defective cases is signiïňĄcantly fewer than that of ordinary procedures, e.g.Boeing assembly line [25].

#### 2.1.2 Popular Classifier Algorithms

Among the traditional statistical methods or standard machine learning algorithms that deal with imbalanced data, which are biased toward much larger negative classes. There are several classifier learning techniques dealing with imbalanced data, like oversampling, under-sampling, boosting, bagging and repeated random sub-sampling [26] [27]. In this study, we implement Random Under-Sampling (RUS) to build an artificial balanced dataset, upon which classifiers will be applied.

Most popular classiïňĄcation learning systems are reported to be inadequate when encountering the class imbalance problem. For instance, decision trees, there is a high probability that some branches that predict the small classes are removed and the new leaf node is labeled with a dominant class; the multi-layer perceptron (MLP), the subsequent rate of decrease of net error for the small class was very low; support vector machines (SVMs), which simply learn to classify everything as the prevalent class in order to make the margin the largest and the error the minimum; Bayesian network, the samples of the small classes are most likely misclassiïňĄed while the learned networks are inferred for classiïňĄcation; the associative classiïňĄcation approaches, because small classes are unlikely to be found since the combination of items characterizing the small classes occur

too seldom to pass certain significance tests for detecting association patterns [28].

Table 2: Statistical Table of the Dataset

| Rare Disease Name | HAE |
|---|---|
| Total no. of Patients | 247,833 |
| no. of Positive Patients | 1,233 |
| no. of Unknown Patients | 246,600 |
| Total no. of unique medical events (Dx + Rx) | 58,030 |
| Total no. of unique Dx (ICD-9 code) | 13,582 |
| Total no. of unique Rx (National Drug Code NDC) | 44,448 |

Table 3: Performance Metrics

| Metric | Definition |
|---|---|
| Recall | TP/(TP+FN) |
| Precision | TP/(TP+FP) |
| True Positive Rate (TPR) | TP/(TP+FN) |
| False Positive Rate | (FPR) FP/(FP+TN) |

## 2.2 Analytic with EHRs

EHRs collect diagnosis, medication, lab tests, procedures, etc. One key aspect to the success of applying EHRs is extracting effective features, which is usually referred to as electronic phenotyping in medical informatics [29]. Recently some computational models have been proposed for EHR based electronic phenotyping, e.g., a matrix based method [30] and a tensor based algorithm [31]. Popular application with EHRs, such as [32] presented risk prediction with patient EHRs with a multi-linear sparse logistic regression. [33] proposed personalized treatment recommendation via a similarity based approach.

Instead of analyzing rich EHR data, based on traditional machine learning and statistical techniques such as logistic regression, support vector machines (SVM), and random forests [34]. Deep learning techniques play important roles in many domains through deep hierarchical feature construction and capturing long-range dependencies in EHRs data in an effective manner[35]. with the increasingly vast amount of patient data, there has also seen an increase in the number of publications applying deep learning to EHR data for clinical informatics tasks [36] [38] [39] [40] which achieved better performance than traditional methods and require less time-consuming pre-processing and feature engineering.

# 3. Data Source and Methodology

## 3.1 Data Source

As we select Hereditary Angioedema disease as an empirical use case in our work. The EHRs data has been extracted from IQVIA (formerly Quintiles and IMS Health, Inc.) longitudinal prescription (Rx) and diagnosis (Dx) medical claims data. The Rx data is derived from electronic data collected from pharmacies, payers, software providers and transactional clearinghouses. This information represents activities that take place during the prescription transaction and contains information regarding the product, provider, payer and geography. Additionally, prescription information can be linked to office based claims data to obtain patient diagnosis information. The Rx data covers up to 88% for the retail channel, 48% for traditional mail order, and 40% for specialty mail order. The Dx data is electronic medical claims from office-based individual professionals, ambulatory, and general health care sites per year including patient level diagnosis and procedure information. The information represents nearly 65% of all electronically filed medical claims in the US.

All data is anonymous at the patient level and HIPAA compliant to protect patient privacy.

## 3.2 Proposed Approach

We build our model as shown in Fig. 2. Since HAE disease is difficult to identify, many patients with this condition donâĂŹt have a positive diagnosis ICD-9 code associate with them. Our objective was to build a predictive model to find such patients using their past EHRs, including prescription records and diagnosis histories.

### 3.2.1 Medical Event Embedding

Extracted EHRs are represeted by temporal matrix. This approach represents the patient EHRs as temporal matrices with one dimension corresponding to date and the other dimension corresponding to medical events. Specifically, we model the EHR record as an longitudinal binary event matrix, where the horizontal dimension corresponds to the time stamps and vertical dimension corresponds to the event values. The $(i, j)$-th entry of an EHR matrix is 1 if the $i$-th event is observed at the $j$-th specific date for the corresponding patient.

In order to get the contextual embedding of each medical event, we concatenate all medical events in each patient's EHRs according to the happening date (if events recorded in the same date, we regardless of the order). Then, we will obtain a "paragraph" describing the historical records of him/her. Word2Vec technique is then applied to derive word representations, which trains a two-layer neural network from a text corpus to map each word into a vector space encoding the word contextual correlations [41]. The similarities, cosine distance in our setting, evaluated in embedded vector space reflecting the contextual associations (e.g., words A and B with high similarity suggests they tend to appear in the same context).

### 3.2.2 Variances Sorting for Feature Selection

All the extracted patients EHRs are represented by a binary matrix with same size. Then, we count the frequency

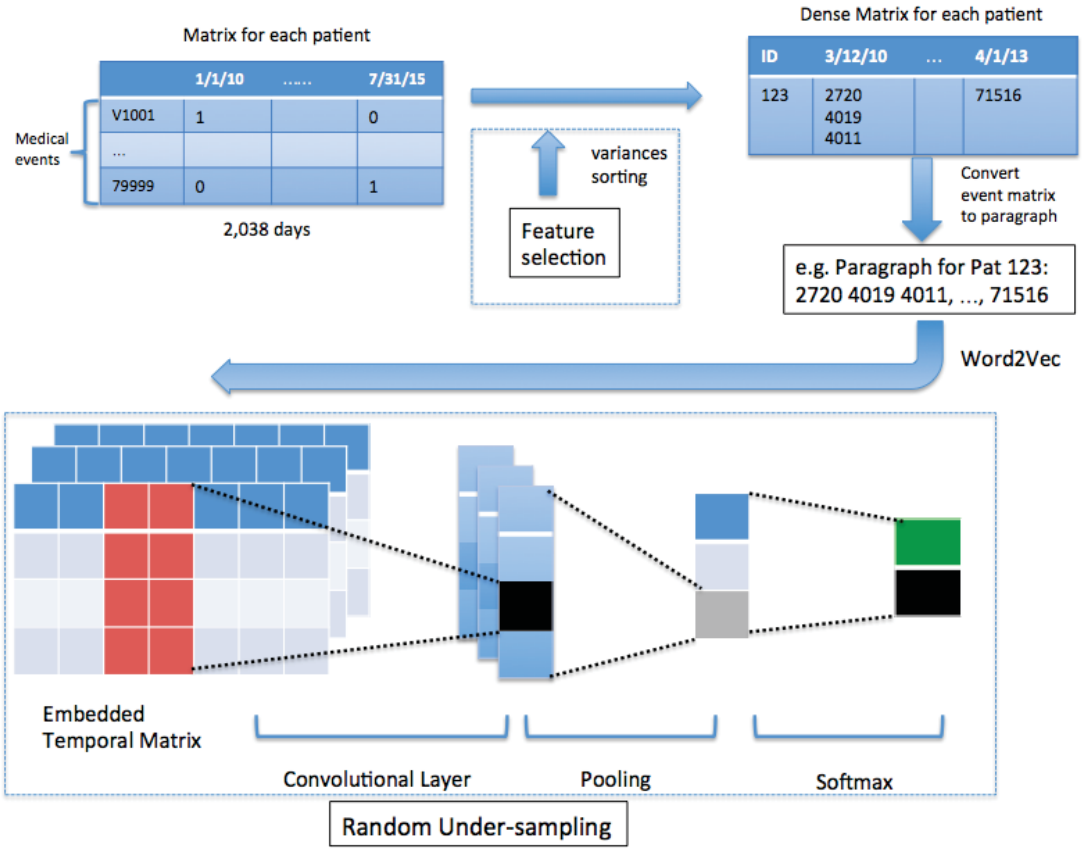


Fig. 2: The overall framework of our proposed random under-sampled convolutional neural network (RUS-CNN) with medical event embedding. (1) Patients EHRs are represented by a binary matrix with same size. (2) Optional variance sorting technique is used to do feature selection. (3) Convert binary matrix to dense matrix representation, in order to get paragraph expression for each patient. The "ID" is patient ID, and "Pat" represents patient. (4) Word2Vec is used to convert medical events into vectors. (5) By summing up vectors to generate embedded temporal matrix as input for CNN. (6)By applying Random Under-sampling to build an artificially balanced dataset. (7) Deep leaning approach CNN for disease classification.

of each event for each patient, and calculate variances for each event, sort the variances from largest to smallest.

### 3.2.3 Temporal Patient Representation

After the medical event embedding step, We expect that the medical event's vector representations learned by Skip-gram model [42] of Word2Vec will support clinically meaningful vector additions. Then, we can simply convert all medical concepts in a patient $p$ EHRs to medical concept vector representation. By summing all those vectors, we will obtain a single representation vector for patient's entire medical history.

At the same time, we want to keep the temporal information for each vector representation. So, we utilize a temporal representation: the records of each patient $p$ is represented as a matrix $M$ with dimension $dT_p$, where $d$ is the fix embedding dimension and $T_p$ is the total number of visit patient $p$ has. A single representation vector of one visit is obtained by summing all the medical vectors in that visit. Usually, $T_p$ varies from patient to patient.

### 3.2.4 Deep Architecture: CNN

The basic CNN model architecture as shown in Fig. 2, where each event matrix of length $t$ is represented $X$ and $X \in \mathcal{R}^{dt}$ . Let $x_i \in R^d$ be the dimensional event vector corresponding to the $i_{th}$ event items. In general, let $x_{i:i+j}$ refer to the concatenation of items $x_i$, $x_{i+1}$, Âů, $x_{i+j}$ . A one-side convolution operation involves a filter $w \in R^{dh}$, which is applied to a window of $h$ event features to produce a new feature. For example, a feature $f_i$ is generated from a window of events $x_{i:i+h1}$ by $f_i = f(w \times x_{i:i+h1} + b)$, where $b \in R$ is a bias term and f is a non-linear function such as rectification (ReLU), tangent (Tanh). This filter is applied to each possible window of features in the event matrix $x_{1:h}, x_{2:h+1}, ..., x_{nh+1:n}$ to produce a feature map

$f = [f_1, f_2, ..., f_{nh+1}]$, where $f \in \mathcal{R}^{nh+1}$. We then apply a mean pooling over the feature map and take the average value $\hat{f} = maxf$. The idea is to capture the most important feature one with the highest value for each feature map. This pooling scheme naturally can deal with variable time stamp lengths of the EHR records. The final layer is a connected layer with dense connections and a softmax classifier.

# 4. Experiment and Evaluation

## 4.1 Sample selection

Those patients who with HAE diagnosis (ICD-9 CODE = 277.6), and at least one HAE treatment (prescription or procedure) during 1/1/2012 to 7/31/2015 are selected as positive HAE patients. The earliest date of HAE diagnosis or HAE treatment is marked as patientâĂŹs index date. Then the lookback period is defined as from earliest diagnosis or prescription date from 1/1/2010 till the day before index date. By applying this rule, we have 1,233 HAE patients with variable lengths of lookback periods in the sample data.

Then, we randomly select 200 times of non-HAE patients with similar lookback period for each HAE patients. For example, if an HAE patient has a lookback period from 2/10/2010 to 7/5/2015, then a matched non-HAE patient will be selected if he/she had clinical activity (any activity of prescription, procedure or diagnosis) from any day in February 2010 to any day in July 2015. The lookback period for this non-HAE patient is counted from earliest clinical activity date in February 2010 to latest clinical activity date in July 2015.

Non-HAE patients set starts from an empty set, then 200 matched non-HAE patients for a given HAE patient are selected to the non-HAE sample. WE repeat this approach for each 1,233 HAE patient until we get 246,600 distinctive non-HAE patients.This process has been done in a greedy manner, and the final dataset configurations are shown in Table 2.

## 4.2 Optional Feature Selection

As listed in Table II, there are 58,030 Rx and Dx events in our dataset. By applying basic statistical learning and medical knowledge, we find certain common medical events, like "ICD-9 code 786.2: Cough", will not affect the final results. Thus, we apply variances sorting to select significant features. In our experiment, we select top 5%, 2,901 (58,030 * 0.05) events that have the largest variances from both HAE and non-HAE datasets. Since there are overlapped events, the size of the subset of events is 4,019. For example, the top ranking events "2776: Other deficiencies of circulating enzymes", "9951: Angioneurotic edema, not elsewhere classified", "4770: Allergic rhinitis due to pollen" are highly related to HAE symptoms.

## 4.3 Imbalanced Classification

In this paper we implement the Random-Under-Sampling (RUS) (majority) approach to deal with imbalanced data. By randomly under-sample the majority class, here is non-HAE patients, and combine them with the minority class, in our case is HAE patients, we build an artificial balanced dataset. Then, different algorithms will be applied to this "balanced" dataset. We repeat this process for several iterations. The final model is an aggregation of models over all iterations.

Three traditional machine learning algorithms, Logistic Regression (LR), LASSO and Random Forest (RF) along with Random-under-sampling are our baseline methods in this work, to compare with deep learning method CNN.

## 4.4 Evaluation

Our dataset is highly imbalanced, the ratio between HAE and non-HAE patients is 1/200, if a classifier simply ignores the HAE patients and predicts all the patients as non-HAE, it may obtain an accuracy of 99.5%. So, we use Precision at various Recall levels for model performance comparisons in this paper, which is defined in Table 3.

### 4.4.1 5-Fold Cross Validation

For validation purpose, data is split into 80% for training and 20% for testing.

Step 1: Separate the 1,233 HAEs to a 80% training HAE set, denoted as TrP, and a 20% testing HAE set, denoted as TeP. For the training HAE set, we identify their matched non-HAE set, denoted as TrN. For the testing HAE set, we identify their matched non-HAE set, denoted as TeN. There is no overlap between TrN and TeN.

Step 2: For 5-fold cross validation, we first separate the training data to 5 folds.For each turn, we pick one fold for validation and other folds for training, and train 200 models since we have 200 matched non-HAE patients. The 200 models are then applied to predict the class label of each patient in the testing fold based on majority vote, then overall precision and recall are calculated on the testing fold. After the above five turns, the average precision and recall are returned as the final result of the 5-fold cross validation process.

## 4.5 Results

With the 80% training data, we perform five-fold cross-validation procedure. Summary of the results are listed in Table 4 and Fig.3.

We further validate the model performances by applying the models to the 20% testing data. A summary of the testing results are listed in Table 5 and Fig.4.

From the experimental results, our proposed Random under-sampled CNN (RUS-CNN) and vRUS-CNN frameworks boost precision at various recall levels compared to 3 traditional machine leaning models, which are also applied with random under-sampling.

Table 4: MODEL PERFORMANCE (Training Data)

| Metric | LR | Lasso | RF | RUS-CNN | vRUS-CNN |
|---|---|---|---|---|---|
| Precision (Recall = 5) | 36.22% | 35.01% | 30.24% | 47.62% | 57.42% |
| Precision (Recall = 10) | 29.49% | 26.87% | 32.58% | 48.78% | 55.35% |
| Precision (Recall = 15) | 22.91% | 19.14% | 30.97% | 49.18% | 53.27% |
| Precision (Recall = 20) | 17.72% | 15.66% | 24.75% | 45.40% | 49.75% |
| Precision (Recall = 25) | 14.85% | 13.93% | 20.83% | 40.91% | 44.87% |
| Precision (Recall = 30) | 13.56% | 11.08% | 16.27% | 35.00% | 40.72% |
| Precision (Recall = 35) | 9.77% | 9.81% | 13.34% | 30.13% | 37.43% |
| Precision (Recall = 40) | 6.78% | 7.86% | 9.46% | 25.82% | 33.94% |
| Precision (Recall = 45) | 6.38% | 7.14% | 7.03% | 20.00% | 21.03% |
| Precision (Recall = 50) | 4.23% | 5.75% | 5.21% | 15.16% | 17.51% |

Table 5: MODEL PERFORMANCE (Testing Data)

| Metric | LR | Lasso | RF | RUS-CNN | vRUS-CNN |
|---|---|---|---|---|---|
| Precision (Recall = 5) | 56.77% | 62.48% | 58.49% | 49.95% | 62.49% |
| Precision (Recall = 10) | 47.17% | 39.19% | 40.26% | 49.02% | 61.86% |
| Precision (Recall = 15) | 28.57% | 24.16% | 39.67% | 45.18% | 59.24% |
| Precision (Recall = 20) | 27.09% | 22.93% | 31.92% | 44.25% | 54.29% |
| Precision (Recall = 25) | 22.90% | 17.84% | 23.57% | 41.34% | 46.57% |
| Precision (Recall = 30) | 18.02% | 14.66% | 21.67% | 36.65% | 42.77% |
| Precision (Recall = 35) | 13.52% | 12.57% | 15.48% | 32.70% | 38.48% |
| Precision (Recall = 40) | 10.11% | 10.30% | 12.34% | 26.60% | 28.65% |
| Precision (Recall = 45) | 7.54% | 10.03% | 10.12% | 21.46% | 22.37% |
| Precision (Recall = 50) | 8.70% | 7.31% | 9.14% | 16.97% | 20.83% |

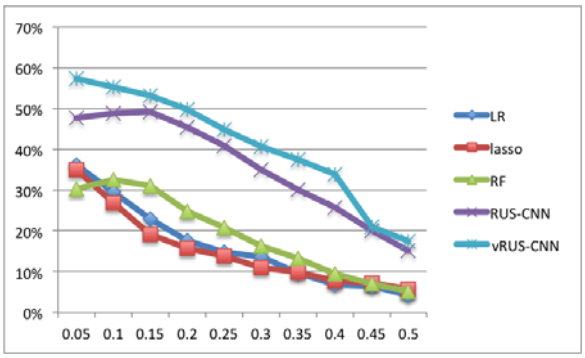


Fig. 3: Five-fold Cross-Validation PR curve for Training data

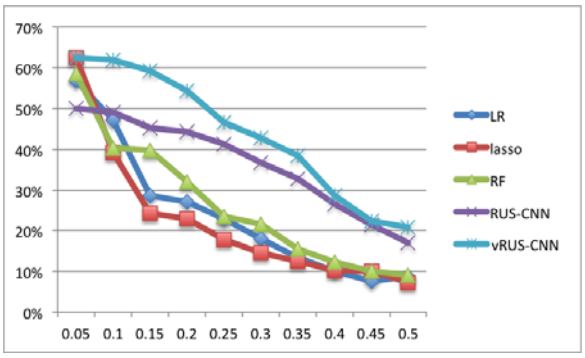


Fig. 4: Five-fold Cross-Validation PR curve for Testing data

## 5. Conclusion

Rare disease identification with EHRs facing two challenging tasks: imbalanced classification and applying EHRs. It is hard to correctly identify true positive rare disease patients out of much larger number of negative patients. In this paper, we investigate the use of Random Under-Sampling strategy to solve the imbalanced instances in the dataset and construct temporal-aware event embedding for deep learning models. Unlike the traditional machine learning methods that require time-consuming pre-processing and feature engineering, we propose an easy approach to transfer the patient historical prescription and diagnosis information in EHRs to embedding representations for deep learning models. In order to gain better predictive accuracy, our proposed Random under-sampled CNN (RUS-CNN) and vRUS-CNN framework work on an empirical rare disease dataset.The experimental results show that our model achieves significantly better results over the 3 state-of-the-art baselines, which enables more accurate rare disease discovery.

# References


[1] P. B. Jensen, L. J. Jensen, and S. Brunak. "Mining electronic health records: towards better research applications and clinical care", Nature reviews. Genetics 13.6 (2012), pp. 395–405.

[2] S. Jaan. "It ainâĂŹt necessarily so: the electronic health record and the unlikely prospect of reducing health care costs", Health Affairs 25.4 (2006), pp. 1079–1085.

[3] L. B. Madsen, "Data-Driven healthcare: how analytics and BI are transforming the industry", John Wiley & Sons, 2014.

[4] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, " A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries", J Am Med Inform Assoc, vol. 18, no. 5 (2011), pp. 601–606.

[5] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, "Predicting Patient âĂŹ s Trajectory of Physiological Data using Temporal Trends in Similar Patients : A System for Near-Term Prognostics", in AMIA Annu Symp Proc (2010), pp. 192–196.

[6] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction", Journal of Biomedical Informatics, vol. 44, no. 5 (2011), pp. 859–868.

[7] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes", Journal of Clinical Epidemiology, vol. 66, no. 4 (2013), pp. 398–407.

[8] "Rare diseases act of 2002 public law 107280 107th congress".

[9] H. He and Y. Ma, Eds., "Imbalanced learning: foundations, algorithms, and applications", John Wiley & Sons, 2013.

[10] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest", BMC medical informatics and decision making 11. 1 (2011), pp. 51.

[11] P. Foster, "Machine learning from imbalanced data sets 101", In Proceedings of the AAAIâĂŹ2000 workshop on imbalanced data sets (2000), pp. 1–3

[12] "International Classification of Diseases, (ICD-10-CM/PCS) Transition-Background". Available: $https : //www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm.$

[13] Choi, Edward, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. "Doctor ai: Predicting clinical events via recurrent neural networks." arXiv preprint arXiv:1511.05942 (2015).

[14] Wang, Fei, Jiayu Zhou, and Jianying Hu. "Densitytransfer: A data driven approach for imputing electronic health records." 2014 22nd International Conference on Pattern Recognition. IEEE, 2014.

[15] A. Bygum, K. E. Andersen, and C. S. Mikkelsen, "Self-administration of intravenous C1-inhibitor therapy for hereditary angioedema and associated quality of life benefits", European Journal of Dermatology 19.2 (2009), PP. 147–151.

[16] M. Cicardi, and A. Agostoni, "Hereditary angioedema", New England Journal of Medicine, vol. 334, no. 25 (1996), pp. 1666âĂŞ-1667.

[17] L. C. Zingale, L. Beltrami, A. Zanichelli, L. Maggioni, E. Pappalardo, B. Cicardi, and M. Cicardi, "Angioedema without urticaria: a large clinical survey", Canadian Medical Association Journal, vol. 175, no. 9 (2006), pp. 1065âĂŞ-1070.

[18] Dai, Dong, and Shaowen Hua. "Random under-sampling ensemble methods for highly imbalanced rare disease classification". Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016.

[19] Yang, Pengyi, Liang Xu, Bing B. Zhou, Zili Zhang, and Albert Y. Zomaya. "A particle swarm based hybrid system for imbalanced medical data sampling." In BMC genomics, vol. 10, no. 3, p. S34. BioMed Central, 2009.

[20] Li, Der-Chiang, Chiao-Wen Liu, and Susan C. Hu. "A learning method for the class imbalance problem with medical data sets." Computers in biology and medicine 40, no. 5 (2010): 509-518.

[21] Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest." BMC medical informatics and decision making 11, no. 1 (2011): 51.

[22] Khatami, Amin, Morteza Babaie, Hamid R. Tizhoosh, Abbas Khosravi, Thanh Nguyen, and Saeid Nahavandi. "A sequential search-space shrinking using CNN transfer learning and a Radon projection pool for medical image retrieval." Expert Systems with Applications 100 (2018): 224-233.

[23] Zong, Wei, Yang-Wai Chow, and Willy Susilo. "A Two-Stage Classifier Approach for Network Intrusion Detection." In International Conference on Information Security Practice and Experience, pp. 329-340. Springer, Cham, 2018.

[24] Zainal, Anazida, Mohd Aizaini Maarof, and Siti Mariyam Shamsuddin. "Ensemble classifiers for network intrusion detection system." Journal of Information Assurance and Security 4, no. 3 (2009): 217-225.

[25] Dietrich, Andreas, Abe Stephens, and Ingo Wald. "Exploring a boeing 777: Ray tracing large-scale cad data." IEEE Computer Graphics and Applications 27, no. 6 (2007): 36-46.

[26] N. Japkowicz, and S. Stephen, "The class imbalance problem: A systematic study", Intelligent data analysis 6.5 (2002), pp. 429–449.

[27] J. R. Quinlan, "Bagging, boosting, and C4.5", AAAI/IAAI, Vol. 1 (1996).

[28] Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23, no. 04 (2009): 687-719.

[29] J. Pathak, A. N. Kho, and J. C. Denny, âĂIJElectronic health records-driven phenotyping: challenges, recent advances, and perspectives,âĂİ Journal of the American Medical Informatics Association, vol. 20, no. e2, pp. e206âĂŞe211, 2013.

[30] J. Zhou, F. Wang, J. Hu, and J. Ye, âĂIJFrom micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records,âĂİ in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014, pp. 135âĂŞ144.

[31] J. C. Ho, J. Ghosh, and J. Sun, âĂIJMarble: highthroughput phenotyping from electronic health records via sparse nonnegative tensor factorization,âĂİ in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014, pp. 115âĂŞ124.

[32] F.Wang, P. Zhang, B. Qian, X.Wang, and I. Davidson, "Clinical risk prediction with multilinear sparse logistic regression", in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2014), pp. 145–154.

[33] P. Zhang, F.Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: Leveraging patient similarity and drug similarity analytics", AMIA Joint Summits on Translational Science, 2014.

[34] K. P. Murphy, "Machine learning: a probabilistic perspective". MIT press, 2012.

[35] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning". MIT Press, 2016.

[36] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, âĂIJDeep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records.âĂİ Scientific reports, vol. 6, no. April, p. 26094, 2016.

[37] A. N. Jagannatha and H. Yu, âĂIJStructured prediction models for RNN based sequence labeling in clinical text,âĂİ in EMNLP, 2016.

[38] A. Jagannatha and H. Yu, âĂIJBidirectional Recurrent Neural Networks for Medical Event Detection in Electronic Health Records,âĂİ arXiv, pp. 473âĂŞ482, 2016.

[39] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, âĂIJDisease Inference from Health-Related Questions via Sparsely Connected Deep Learning,âĂİ Knowledge and Data Engineering, IEEE Transactions, vol. 27, no. 8, pp. 2107âĂŞ2119, 2015.

[40] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, âĂIJMedical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction,âĂİ arXiv, p. 45, feb 2016.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, âĂIJEfficient estimation of word representations in vector space,âĂİ arXiv preprint arXiv:1301.3781, 2013.

[42] Goldberg, Yoav, and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." arXiv preprint arXiv:1402.3722, 2014.