# TOWARDS USE OF ELECTRONIC HEALTH RECORDS: CANCER CLASSIFICATION

Siyu Liao
Department of Computer Science
Graduate Center of The City University of New York
365 5th Avenue
New York, NY 10016
sliao2@gradcenter.cuny.edu

Jiehao Xiao
Department of Computer Science
Graduate Center of The City University of New York
365 5th Avenue
New York, NY 10016
jhxiao12@gmail.com

Yi Xie
Electrical Engineering Department
The City College of New York
160 Convent Avenue
New York,, NY 10031
yxie000@citymail.cuny.edu

Feng Gu
Department of Computer Science
College of Staten Island
2800 Victory Boulevard
Staten Island, NY 10314
Feng.Gu@csi.cuny.edu

## ABSTRACT

The electronic health records (EHR) are generating increasing quantities of data like never before. These extensive amount and large varieties of data provide valuable information and bring opportunities for researchers to study, thus to help clinical practice for clinicians and related decision-making for organizational managers, such as disease diagnosis and healthcare policy making. Cancer diagnosis is very important to patients because finding and treating cancer at an early stage can save lives. In this paper, we choose cancer classification as an example to demonstrate the usage of the EHR data. The data are collected and provided by clinics from private practices in New York City. We apply random forest method for cancer diagnosis. The experimental results of our preliminary work show its effectiveness, 68.5% accuracy in classifying 15 different types of cancers. Moreover, five of them are highly identified and classified, reaching the accuracy of 92.84%.

**Keywords:** electronic health records, cancer diagnosis, classification, medical, random forest .

## 1 INTRODUCTION

The wide adoption of electronic health records (EHR) provides improved clinical practice. EHR systems collect patients' health information including diagnosis, lab tests, and medication, etc. The extensive use of these EHR data on one hand provides clinicians with convenience to store and quickly query patients' clinical records. On the other hand, these proliferate data can be used in clinical and translational research. Researchers use EHR data to identify patterns for some diseases and provide useful information for clinicians in their practice. Moreover, other organizations, such as healthcare service providers, can analyze abundant EHR records to facilitate policy making. Although many challenges still exist due to high dimensionality, temporality, sparsity, irregularity, and bias of EHR data (Cheng et al. 2016), a lot of

*SpringSim-MSM 2017, April 23-26, Virginia Beach, VA, USA*

research scholars make great efforts to effectively utilize the EHR data in various aspects. These include risk prediction of patients, personalized treatment recommendation, and patient similarity evaluation (Wang et al. 2014; Zhang et al. 2014; Wang, Sun, and Ebadollahi 2012).

Data representation of EHR is a key step to their effective usage. The main data representation methods include vector based representation (Wang et al. 2014), tensor based representation (Ho et al. 2014), sequence based representation (Gotz, Wang, and Perer 2014), and temporal matrix based representation (Zhou et al. 2014). Vector based representation and tensor based representation define each patient as a vector and a tensor with different modes respectively without considering the temporal features. Sequence based representation and temporal matrix based representation take temporality into consideration and construct a sequence and a matrix for each patient respectively. However, the temporal dimension increases the complexity of the data, resulting in the difficult pattern identification. The choice of data representation is important to meet the needs of different applications.

Many different approaches are adopted for the solution to various problems. To solve the classification problem, the used algorithms include *K* nearest neighbors (Ruiz 1986), support vector machines (Suykens and Vandewalle 1999), decision trees (Quinlan 1986), random forest (Liaw and Wiener 2002) and Naive Bayes (Rish 2001). In terms of deep learning methods, the widely used algorithm is convolutional neural networks (CNN), a neural network algorithm with multiple layers with filters applied to local features (Oquab et al. 2014). CNN is one of the classic deep learning models and has great potentials to be used in deep learning of EHR data with rich temporal information (Cheng et al. 2016). But deep learning methods require huge amount of data compared with traditional methods.

EHR data can be used for disease diagnosis, such as cancer classification. In 2012, cancer caused around 14.6 percent of human death (Stewart et al. 2016). Therefore, its diagnosis and early treatment is very critical to patients' health and their improved survival. The chance of potential therapy can be reduced if there are any diagnosis delays or missed opportunities. EHR data can facilitate detecting potential delays in cancer diagnosis (Murphy et al. 2014). This motivates us to apply machine learning techniques into cancer diagnosis to demonstrate the usage of EHR data. Towards this goal, we represent the EHR data from clinics in New York City as vectors and apply machine learning algorithms for cancer diagnosis. Through the example, we aim to show that EHR can be efficiently used in various occasions.

The rest of this paper is organized as follows. Section 2 briefly discusses existing work in EHR data and their usage and related methods. Section 3 introduces the used method in this paper and section 4 presents experiments settings and the experimental results. Section 5 draws conclusions and points out the future work.

## 2 RELATED WORK

Electronic health records have received much attention recently since they contain valuable information of patients. A lot of research has been done to study and translate those data. Choi et al. (2016) developed a predictive model Doctor AI to predict clinical events using recurrent neural networks. The model was tested by longitudinal time stamped EHR data to show its effectiveness. Henao et al. (2015) presented a multi-modality architecture for EHR analysis based on Poisson Factor Analysis. EHR of over 240,000 were used to examine the model and better mortality and morbidity predictions were achieved. Miotto et al. (2016) designed an unsupervised deep feature learning method to derive a patient representation from EHR data, thus to facilitate clinic predictive modeling. The method was evaluated by EHR data of 76,214 patients. Gary et al. (2016) used machine learning approaches to identify rare disease patients from EHR data. They targeted cardiac amyloidosis identification with the help of experienced cardiologists.

To effectively utilize EHR data, we need to represent the data using various data structures. This is called electronic phenotyping. Wang et al. (2014) represented each patient as a vector with linear combination of raw medical events and their related coefficients were decided by optimization methods. However, those events didn't have time stamps. Ho et al. (2014) constructed a tensor for each patient with different

modes, each of which is corresponded to a specific type of medical entity. They explored the interactions among medical entities. It didn't consider the temporal factor either. Gotz, Wang, and Perer (2014) represented each patient as a sequence with time stamped events. It returned a large number of patterns and therefore, it is hard to identify useful patterns. Zhou et al. (2014) defined a patient as a two dimensional matrix with time and medical events to detect shift invariant patterns using EHR data. However, it needed to enumerate all the possible values and had low performance. Therefore, vectors and tensors are simple, but don't have temporal dimension. Sequences and temporal matrices incorporate time, but suffer from performance. Due to its simplicity, our example of EHR data usage in this paper will be based on vector representation.

Based on the data presentation of EHR, numerous approaches and methods are applied to analyze EHR data of patients. Gary et al. (2016) developed a bootstrap machine learning approach to identify rare diseases from EHR data. In the paper, they systematically compared various approaches including *K* nearest neighbor, support vector machines, naive Bayes, decision trees, random forest, and Adaboost. Among them, random forest achieved a high tenfold cross-validation F1 score of 0.98. Cheng et al. (2016) used deep learning methods for risk prediction from EHR data. Convolutional neutrual networks (CNN) was applied to apply EHR data of patients. Their deep learning framework showed the effectiveness by quality and quantity evaluations. Although other approaches, such as regression models, were utilized, machine learning and deep learning methods gain popularity and we will use random forest in our example of usage of EHR data in this paper.

## 3 A CASE STUDY FOR USE OF EHR DATA: CANCER CLASSIFICATION

In this section, we present a case of cancer classification using EHR data. A classification problem is categorizing a new observation based on a training set which contains observations with known membership. Thus, cancer diagnosis can be seen as a classification problem in the machine learning community (Guyon et al. 2002). In cancer classification, the existing EHR data are treated as the training set. When given new EHR without diagnosis results, cancer diagnosis can be achieved by solving a classification problem. The new EHR data without diagnosis results are a testing set. There are two main steps in classification, feature extraction (also called phenotyping or data presentation) and data analytics. In feature extraction, the clinically relevant information from raw EHR data is selected for efficient use of future steps. In data analytics, based on the data presentation in feature extraction, various algorithms are applied to analyze the extracted data to achieve classification. The patients can be represented by vectors, tensors, sequences, or matrices for data analytics. Due to its simplicity and effectiveness, we adopt vector based representation, in which each patient is defined by a vector. We use random forest to analyze those vectors from EHR data. Random forest has two main steps, constructing decision trees and classifying the test case based on individual trees. We present them below.

EHR data can contain both categorical and continuous features, such as pain-level and globulin test. This makes it difficult to directly apply machine learning techniques. However, decision tree (Quinlan 1986) turns out to be a powerful classification algorithm with many good properties. It is suitable for multi-class tasks, able to handle mixture of categorical and continuous features, and able to measure importance of different features. Therefore, it is a popular choice for cancer diagnosis tasks. A decision tree is like a flow chart as shown in Figure 1. In the figure, a node (display in ellipse) means testing on a feature and a branch represents the output. The leaf node (display in square) has no branches and stands for a certain diagnosis result. From the figure, we see that there exist three features, four outputs, and three diagnosis results. The path from the root to a leaf node means the classification strategy, or diagnosis strategy. The basic idea of determining branches is to make nodes within a branch have high purity, in other words, be of the same class as much as possible. This could be done according to some mathematical heuristics, for example, information gain.

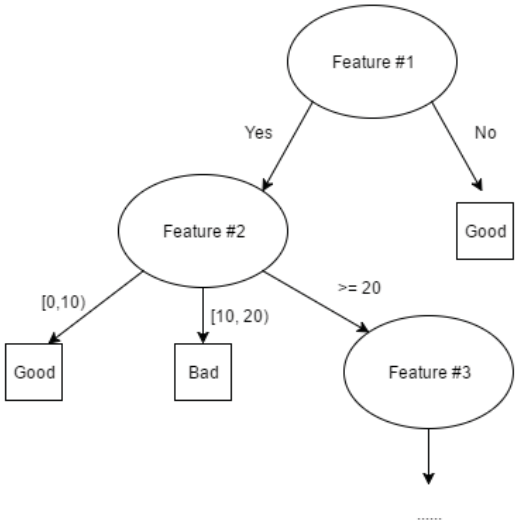


Figure 1: An Example of Decision Tree.

Let $D$ represent the set of current observations, or samples, and $A$ is the feature used to classify these observations. There are $Y$ classes for these observations. The ratio of each class over $D$ is $P_k$, where $k \in \{1,2,...,Y\}$. The information entropy $Ent(D)$ is defined below.

$$Ent(D) = -\sum_{k=1}^{Y} P_k \log_2 P_k .$$

The lower the information entropy, the higher purity for current classification. Information gain is based on the idea of making the entropy of each branch as small as possible, so the last branch would have high purity. Thus, similar cases are well classified and represented in the tree. This in turns means gaining much information entropy at current branch. Let feature $A$ have $V$ possible divisions to generate branches, i.e., $\{A^1, A^2,..., A^V\}$. The choice of these division is based on the information gain $Gain(D, A)$, which is defined as follows.

$$Gain(D, A) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v),$$

where $D^v$ represents observations with value $A^v$ on the feature $A$. The greater information gain, the less entropy left for each division. As a result, it is assumed to have high purity for the classification. Features at the top of the tree contribute to a larger fraction of observations than those in the bottom. The expected fraction can be used as an estimate of the importance of the feature.

Based on constructed decision trees, random forest ensembles a set of them during training and takes the mode of classification results of these trees for testing. The training algorithm for random forest is bagging (Breiman 1996), which uniformly samples from the training set with replacement. Each sample will be equally likely sampled and put into a bag to train a decision tree. Thus, the algorithm can form a forest with multiple decision trees. Moreover, during training, random selection of features is introduced to construct tree with controlled variance (Amit and Geman 1997), in which instead of evaluating all features for branching, only part of features are selected for evaluation. In random forest, the feature importance can be calculated by the average of feature importance from those trees.

## 4 CASE STUDY EVALUATION AND RESULTS

### 4.1 Case Study Evaluation

In our case study, we processed the raw EHR data using vector based presentation and applied random forest to the processed data for cancer classification. The used EHR data are from some clinics in New York City which consist of prescriptions, diagnosis, laboratory tests and vital-signs. Each of them is stored separately as a table where each row represents a record and each column refers to an event. Each record has the patient ID and the visiting time. For cancer diagnosis, we take tables of diagnosis, laboratory tests and vital signs for evaluating the performance. We assume that the patient has the laboratory tests, vital-signs and the diagnosis on the same date. Thus, these tables can be combined based on the patient ID and their visiting time, which is often referred as an join operation in database. After the join, the patient ID and visiting time columns are dropped because we classify only according to clinical features which are seen as irrelevant to the date. Moreover, since there are 45 different cancers in the dataset, the joined table that contains other diseases will be filtered by these cancers that are encoded in ICD-9 or ICD-10. As a result, 9,052 cancer records are left where each has 522 features.

A threshold is then set as 100 for both cancers and each column. This is to make sure there are enough records for analyzing each cancer and each feature. More specifically, each row represents the record of a cancer patient on certain date. Cancers with less than 100 records will be filtered out. Similarly, columns with less than 100 records are also filtered out. This results in the data used for further analysis, as shown in Table 1. From the table, we know that there are 8,607 records totally with 79 features on 15 different cancers.

Table 1: Preprocessed EHRs.

| ICD | Cancer Name | Number of Records |
|---|---|---|
| 174.9 | Breast Carcinoma Female | 3,461 |
| 162.9 | NSCLC/Adenocarcinoma | 1,075 |
| 153.9 | Colon Cancer | 701 |
| 151.0 | Gastric Carcinoma | 529 |
| 199.0 | Disseminated Carcinoma | 445 |
| 147.9 | Nasopharyngeal Cancer | 411 |
| 154.1 | Rectal Carcinoma | 409 |
| 151.9 | Gastric Carcinoma | 343 |
| 189.1 | Renal Cell Carcinoma | 223 |
| 150.9 | Esoph Carcinoma | 210 |
| 185 | Prostate Cancer | 207 |
| 180.9 | Cervical Carcinoma | 187 |
| 157.9 | Pancreatic Cancer | 178 |
| 152.9 | Adenocarcinoma of Small bowel | 114 |
| 183.0 | Ovarian Cancer | 114 |

From Table 1, we also notice that the data is highly imbalanced. For example, Female Breast Carcinoma has around 3,000 records while the Ovarian Cancer has only around 100 records. This could result in biased classification results. It means even if all cases are classified as Female Breast Carcinoma, the accuracy of the classifier can still be around 40.21%. But this doesn't require learning anything from the data. Two ways can be used to solve the imbalanced data problem, oversampling and undersampling, which make the data uniformly distributed by generating synthetic data or removing extra data, respectively. In this paper, we adopt the undersampling method so that all training data are from real world and with practical meaning. So we finally narrow down the data to 1,710 records for 15 cancers where each has 114 records.

### 4.2 Experimental Results

We use *K*-fold cross validation for assessing how random forest could perform on new EHR. In other words, our objective is to estimate how well the random forest could perform in practice. All experiments are conducted using the Python Scikit-learn Toolkit (Pedregosa et al. 2011) and averaged over 10 times. In *K*-fold cross validation, the given data set is randomly partitioned into *K* subsets with equal size. Among the *K* subsets, there are *K*-1 subsets used as training data and the remaining one is for testing data. This process will be repeated *K* times to make each subset used exactly once as testing data. Finally *K* times evaluation performance will be averaged as the estimation of performance of the classifier. Based on *K*-fold cross validation, we report our findings below.

We classify 15 different cancers as listed in Table 1 with accuracy of 68.5% with 200 trees. Table 2 shows the top 10 important features in this classification for these 15 cancers. From the table, we see that the glucose is the most important feature in classifying these cancers. Figure 2 shows the confusion matrix for these cancers represented in their ICD codes. In the figure, the horizontal axis means diagnosis generated by the classifier while the vertical axis represents the ground-truth diagnosis of corresponding EHR. The element in the matrix means number of records that are classified as its horizontal axis but is actually diagnosed as its vertical axis. Thus, values in the diagonal mean the number of correctly diagnosed records. Since it is the averaged result, values in confusion matrix are not necessarily integer numbers. From the figure, we know that Adenocarcinoma of Small Bowel (ICD 152.9) has the most correctly classified records.

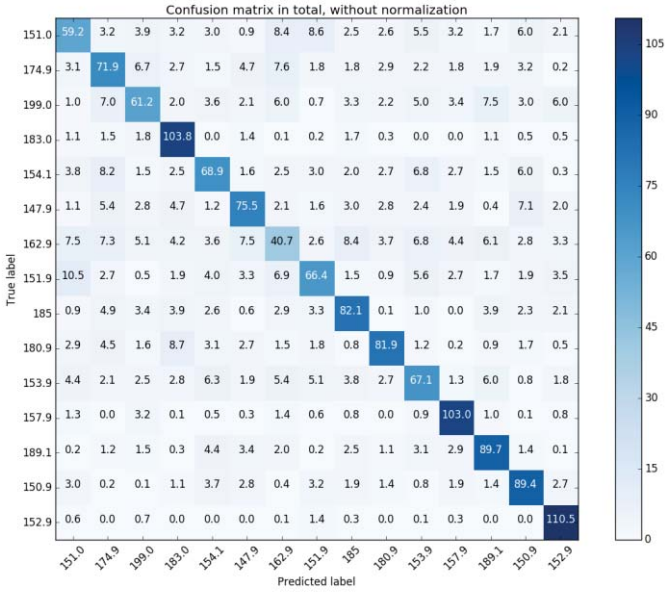


Figure 2: Confusion Matrix for Classifying 15 Cancers.

Table 2:Top 10 Important Features for Classifying 15 Cancers.

| Feature Name | Importance |
|---|---|
| Glucose | 0.060,883 |
| Creatinine | 0.051,786 |
| Calcium | 0.048,298 |
| blood pressure (high) | 0.045,695 |
| blood pressure (low) | 0.044,525 |
| urea nitrogen | 0.042,171 |
| Potassium | 0.041,529 |
| Temperature | 0.039,963 |
| Sodium | 0.039,839 |
| Chloride | 0.039,221 |

According to the confusion matrix in Figure 2, we focus on top five cancers that are highly classified including Renal Cell Carcinoma, Pancreatic Cancer, Adenocarcinoma of Small Bowel, Esoph Carcinoma and Ovarian. We use the same method to process the EHRs data of these five cancers to obtain 570 records, each of which has 34 features. We achieved 92.84% accuracy for classifying these 5 cancers. The confusion matrix and top 10 important features are shown in Table 3 and Figure 3 respectively. From the table, we see that glucose is still the most important feature. However, compared with the previous result in Table 2, there are other new features ranked top 10, for example, the glomerular filtration rate cal. This feature actually measures the overall index of kidney function, which is related to some of these cancers.

Table 3: Top 10 Important Features for Classifying 5 Cancers.

| Feature Name | Importance |
|---|---|
| glucose | 0.101,137 |
| glomerular filtration rate cal | 0.057,319 |
| calcium | 0.055,394 |
| urea nitrogen | 0.050,696 |
| creatinine | 0.047,268 |
| chloride | 0.045,413 |
| egfr african american | 0.041,210 |
| egfr non afr american | 0.041,037 |
| blood pressure (high) | 0.040,976 |
| sodium | 0.040,101 |

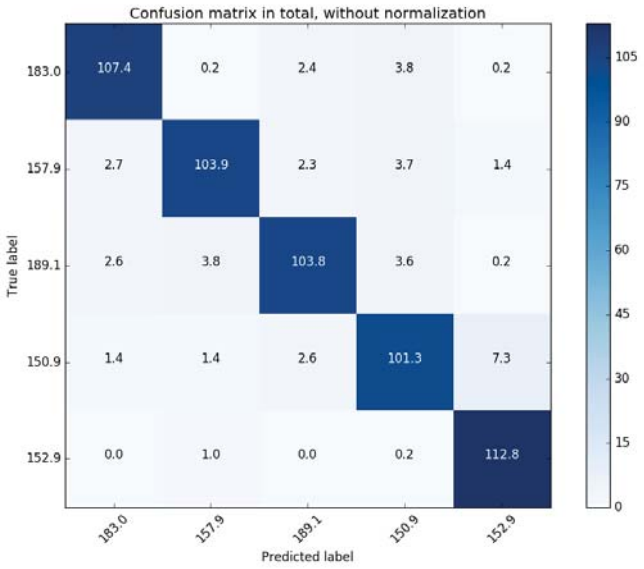


Figure 3: Confusion Matrix for Classifying 5 Cancers.

## 5 CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the potential usage of EHR data and use cancer diagnosis to demonstrate its usage and effectiveness. The designed experiments show that an acceptable accuracy is achieved to evaluate 15 different cancers with 68.5% accuracy. Further, five of them can be identified with accuracy 92.84%. These promising results will provide guidelines for our future exploration to use EHR data, especially in disease diagnosis and healthcare policy making. Our future work will be in the following directions. First, we will further analyze and evaluate the used method using more EHR data. Second, we plan to apply the method to other disease diagnosis. Finally, we will examine the effects of different methods on the diagnosis accuracy, for example Gradient Boosting.

## REFERENCES

Amit, Y., and D. Geman. 1997. "Shape quantization and recognition with randomized trees". *Neural computation* vol. 9 (7), pp. 1545–1588

Breiman, L. 1996. "Bagging predictors". *Machine learning* vol. 24 (2), pp. 123–140.

Cheng, Y., F.Wang, P. Zhang, and J. Hu. 2016 "Risk Prediction with Electronic Health Records: A Deep Learning Approach". In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 361–369

Choi, E., A. Schuetz, W. F. Stewart, and J. Sun. 2016. "Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction". *arXiv preprint arXiv:1602.03686*.

Centers for Disease Control and Prevention. 2012. "International Classification of Diseases (ICD)". *National Center for Health Statistics*.

Garg, R., S. Dong, S. Shah, and S. R. Jonnalagadda. 2016. "A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records". *arXiv preprint arXiv:1609.01586*.

Gotz, D., F.Wang, and A. Perer. 2014. "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data". *Journal of biomedical informatics* vol. 48, pp. 148–159.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. "Gene selection for cancer classification using support vector machines". *Machine learning* vol. 46 (1-3), pp. 389–422.

Henao, R., J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin. 2016. "Electronic Health Record Analysis via Deep Poisson Factor Models". *Journal of Machine Learning Research* vol. 17 (186), pp. 1–32

Ho, J. C., J. Ghosh, and J. Sun. 2014. "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization". In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 115–124. ACM.

Hripcsak, G., and D. J. Albers. 2013. "Next-generation phenotyping of electronic health records". *Journal of the American Medical Informatics Association* vol. 20 (1), pp. 117–121.

Liaw, A., and M. Wiener. 2002. "Classification and regression by randomForest". *R news* vol. 2 (3), pp. 18–22.

Miotto, R., L. Li, B. A. Kidd, and J. T. Dudley. 2016. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". *Scientific reports* vol. 6.

Murphy, D. R., A. Laxmisan, B. A. Reis, E. J. Thomas, A. Esquivel, S. N. Forjuoh, R. Parikh, M. M. Khan, and H. Singh. 2014. "Electronic health record-based triggers to detect potential delays in cancer diagnosis". *BMJ quality & safety* vol. 23 (1), pp. 8–16.

Oquab, M., L. Bottou, I. Laptev, and J. Sivic. 2014. "Learning and transferring mid-level image representations using convolutional neural networks". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* vol. 12, pp. 2825–2830.

Quinlan, J. R. 1986. "Induction of decision trees". *Machine learning* vol. 1 (1), pp. 81–106.

Rätsch, G., T. Onoda, and K.-R. Müller. 2001. "Soft margins for AdaBoost". *Machine learning* vol. 42 (3), pp. 287–320.

Rish, I. 2001. "An empirical study of the naive Bayes classifier". In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Volume 3, pp. 41–46. IBM New York.

Ruiz, E. V. 1986. "An algorithm for finding nearest neighbours in (approximately) constant average time". *Pattern Recognition Letters* vol. 4 (3), pp. 145–157.

Stewart, B., C. P. Wild et al. 2016. "World cancer report 2014". *World*.

Suykens, J. A., and J. Vandewalle. 1999. "Least squares support vector machine classifiers". *Neural processing letters* vol. 9 (3), pp. 293–300.

Wang, F., J. Sun, and S. Ebadollahi. 2012. "Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment". *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5 (1), pp. 54–69.

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

Wang, X., F. Wang, J. Hu, and R. Sorrentino. 2014. "Exploring joint disease risk prediction". In *AMIA Annual Symposium Proceedings*, Volume 2014, pp. 1180. American Medical Informatics Association.

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

Zhou, J., F. Wang, J. Hu, and J. Ye, "From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 135–144.

## AUTHOR BIOGRAPHIES

**SIYU LIAO** is a PhD student in the Department of Computer Science at the Graduate Center of The City University of New York. His research interests include machine learning and high performance computing. His email address is sliao2@gradcenter.cuny.edu.

**JIEHAO XIAO** is a PhD student in the Department of Computer Science at the Graduate Center of The City University of New York. His research interests include machine learning and medical simulation. His email address is jxiao@gradcenter.cuny.edu.

**YI XIE** is a PhD student in Electrical Engineering Department at The City College of New York. Her research interests include machine learning and stochastic computing. Her email address is yxie000@citymail.cuny.edu.

**FENG GU** is an Assistant Professor in the Department of Computer Science at College of Staten Island, The City University of New York. He holds a Ph.D. in Computer Science from Georgia State University. His research interests include modeling and simulation, high performance computing, and bioinformatics. His email address is Feng.Gu@csi.cuny.edu.