HANDLING THE MISSING DATA PROBLEM IN ELECTRONIC HEALTH RECORDS FOR CANCER PREDICTION

Xudong Zhang
Jiehao Xiao
Department of Computer Science
The Graduate Center, City University of New York
365 5th Ave
New York, NY 10016
xzhang5@gradcenter.cuny.edu
jxiao@gradcenter.cuny.edu

Ning Yu
Department of Computer Science
The College at Brockport, State University of New York
350 New Campus Dr
Brockport, NY 14420
nyu@brockport.edu

Yifei Gong
Department of Computer Science
CUNY Graduate Center
365 5th Ave
New York, NY 10016
Yifeigong2014@gmail.com

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

ABSTRACT

Electronic health records (EHRs) are the records containing the patients' clinic information and medical records. The EHRs have been widely used in disease diagnosis and therapy due to the numerous and valuable medical information in them. However, the missing data problem of EHRs hinders the usage of EHRs. Replacing the missing data with mean values is an approach of data imputation. But, that method weakens the feature importance of missing data. In this study, we use the expectation-maximization (EM) algorithm to impute the EHRs containing missing data for cancers. The EM algorithm iteratively estimates the missing data in EHRs until the maximum likelihood is obtained. An artificial neural network and several machine learning models including logistic regression, support vector machine, and random forests are used to evaluate the effectiveness of data imputation in EHRs. The experimental results show that the prediction accuracies of cancers by using artificial neural network, logistic regression, support vector machine, and random forests on the EHRs imputed by EM algorithm are higher than those by mean values, which indicates the EM algorithm is able to provide accurate estimations in data imputation of EHRs.

Keywords: missing data, electronic health records (EHRs), machine learning, neural network, data imputation.

SpringSim 2020, May 19-21, George Mason University, Fairfax Campus, VA, USA;

1 INTRODUCTION

The EHRs are clinic records containing the patients' information, including the ages, genders, diagnoses, medications, allergies, treatment plans, radiology images, lab test results, and medical history. The medical information and treatment histories extracted from EHRs are highly related to patients' health. EHRs can be used by medical providers and researchers to study patients' medical history and provide suitable diagnose and treatment.

Many diseases have long incubation periods and development periods. Analyzing the EHRs is helpful in providing timely treatment and therapy. Recently, cancer is one of the deadliest diseases all over the world. In 2017, 1,688,780 persons were diagnosed to have various types of cancers in the United States (Siegel et al. 2017). There is no doubt that the patients' survival rates can be improved if the cancer can be diagnosed and treated in the early stage. The EHRs of patients with cancers in the early stages contain relative medical information and those information can be used to diagnose cancers.

The artificial neural network and machine learning models, such as the logistic regression, support vector machine, and random forests, have good convergence in classification problems and they have been applied in the prediction of cancers (Kourou et al. 2015). Although different classification methods have been adopted in cancer diagnosis and prediction, there still exist some challenges. One challenge of using classification models on EHRs in disease prediction and diagnosis is the missing data problem because neither the artificial neural network nor the machine learning models are able to handle the data set with missing data. The medical information extracted from EHRs usually have missing data problems due to various reasons, including testing equipment availability, disease progression and testing condition. Usually, the medical information related to patients' health, is recorded by medical providers for several years, which is inevitable to have missing data problems in EHRs. As a result, it is impossible for the machine learning classifiers to deal with the EHRs with missing data.

Imputation on EHRs with missing information is necessary for the machine learning models to produce reasonable classification and prediction results. The quality of data set is critical for machine learning models, and any wrong or biased data will result in low classification and prediction accuracies. Although, some machine learning models have been successfully used on EHRs imputed with mean values (Liao, Xiao, and Gu 2017; Zhang, Xiao, and Gu 2019). Using the mean values to impute the missing data weakened or eliminated the feature importance because it was hard for the machine learning classifiers to differentiate different cases by using the same mean values. As a result, the classification and prediction accuracies were lowered because of those same values. Based on that motivation, we propose to use expectation–maximization (EM) algorithm to impute the missing data in EHRs. The EM algorithm is an iterative reasoning to find the estimates of maximum likelihood or maximum a posterior in statistical models (Dempster, Laird, and Rubin 1977). There are two basic steps, the expectation (E) step and the maximization (M) step. In the E steps, the EM algorithm creates the expectation of the log-likelihood using the current parameter estimates. In the M steps, the EM algorithm maximizes the log-likelihood obtained in the E steps and updates the estimated parameter. Those E steps and M steps executed alternatively, until the stopping rule is satisfied. We use artificial neural network, logistic regression, support vector machine, and random forests to classify the cancer data set imputed by mean values and EM algorithm to evaluate the proposed method and show the application of EHRs on cancer prediction effectively.

The rest of this paper is organized as follows. Section 2 presents the related work on applications of EHRs, EM algorithm, and machine learning techniques. Section 3 introduces the background of the EM algorithm in data imputation. Section 4 describes the data imputation for EHRs. Section 5 presents the experimental result and discussion. Section 6 concludes this study and points out the future work.

2 RELATED WORK

The EHRs have been widely used in disease diagnosis and treatment due to the tremendous information in the records. Castro et al. (2013) introduced a cross sectional study using the electrocardiographic, pre-

scribing, and clinical data extracted from EHRs to reveal the relationship between the antidepressant dose and QTc. The experimental results indicated that the pharmacovigilance studies using the medical data from EHRs were a useful approach of identifying potential risks associated with treatments. Murphy et al. (2014) proposed to use data mining algorithms to extract retrospective data from EHRs. The data mining algorithms are able to identify all patient records with specific demographics. The experimental results implied that the EHR-based triggers can be successfully used to identify the patient records, which lacked follow-up of abnormal clinical findings for cancers. Singh et al. (2010) evaluated 587 patients of primary lung cancer from the medical data extracted from EHRs. Two physicians independently reviewed the medical records and the disagreements were resolved by consensus. The study indicated that the preventable delays in lung cancer diagnosis appeared from failure to complete the diagnosis in a timely manner. The potential solutions to the delays were the EHR-based strategies to improve recognition of abnormal imaging and track patients with suspected cancers. Jensen et al. (2012) pointed out the EHRs had potentials to establish the new patient-stratification principles and reveal the unknown disease correlations. Integrating EHR data with genetic data was able to provide a understanding of genotype–phenotype relationships. They found the EHRs could improve the medical research and clinical care. Although the EHRs have much potential in helping and improving the disease diagnosis. The missing data problem in EHRs needs to be figured out to provide reasonable and complete medical information.

One approach to impute the missing data in EHRs is to use the mean values. Liao, Xiao, and Gu (2017) and Zhang, Xiao, and Gu (2019) used the mean value to impute the missing data in EHRs. They used machine learning models on the medical data extracted from EHRs, to classify different cancers. The experimental results indicated that the machine learning models can predict the cancers with high accuracy by using the imputed medical data. However, imputing the data with mean values lowers the feature importance and prediction accuracy. The EM algorithm iteratively reasons the estimates of maximum likelihood in statistical models, and it had been successfully used in data imputation. Schneider (2001) proposed a regularized EM algorithm, which was applicable to several sets of climate data. The regularized EM algorithm iteratively analyzed the linear regressions of variables with missing values. The experimental results revealed that the EM algorithm could provide more accurate estimates of the missing values. Catellier et al. (2001) proposed to use the EM algorithm to implement missing value imputation in summarizing the accelerometer activity. The experimental results indicated that the EM algorithm is able to impute the missing data while measuring physical activity by accelerometry.

Some researches and studies have been focusing on using machine learning models to handling EHRs. Wu, Roy, and Stewart (2010) compared the performance of different machine learning models, including logistic regression, boosting, and support vector machine in heart failure prediction by using the medical data from EHRs. The experimental results showed that the selection of logistic regression model based on Bayesian criterion provided the most parsimonious model. Boosting with variable importance threshold is able to provide similar performance. Support vector machine had the poorest performance due to the imbalanced data. The machine learning model can predict the heart failure more than 6 months before clinical diagnosis. Carroll et al. (2011) proposed to use the support vector machine model to identify Rheumatoid Arthritis cases by using EHR. The SVM model was trained on medical data from EHRs and the prediction accuracy is 0.94. The experiment revealed that the SVM model was capable to classify different disease cases from EHRs. Garg et al. (2016) used machine learning models to identify rare diseases from a large number of potential diseases. The experimental results indicated that the machine learning models can identify the cardiac amyloidosis from EHRs by using an ensemble machine learning classifier. Liao, Xiao, and Gu (2017) used the random forests model on EHRs in cancer prediction. The prediction accuracy was 68.5% if classify 15 different cancers. A higher prediction accuracy, 92.84%, can be achieved if only 5 of them were classified. Zhang, Xiao, and Gu (2019) used support vector machine model to classify and prediction various cancers by using the medical test results from EHRs. The prediction accuracy was 86.2% when classifying 10 types of cancers with 1,000 pieces of medical records, and 97.33% when classifying 3 types of cancers with 1,200 pieces of medical records.

Although many machine learning methods were successfully used in disease prediction based on the medical records. The missing data problem in EHRs needs to be solved to improve the prediction accuracy. Therefore, we propose to apply EM algorithm to impute the missing data in EHRs and evaluate the imputation by classifying and predicting different cancers with different machine learning models.

3 EXPECTATION–MAXIMIZATION ALGORITHM IN HANDLING MISSING DATA

The EM algorithm was originally introduced to iteratively compute the maximum-likelihood estimations from the incomplete data sets. There are two basic steps in the EM algorithm. They are the estimation steps and the maximization steps. In the estimation steps, the expectation of the log-likelihood using the current parameter estimates is created, as shown in Equation (1). In Equation (1), Y is the set of observation containing missing data; Z is the hidden variable; θ is the parameter of distribution of Y; θ^i is the estimated θ in step i; E represent the expectation.

$$E(\theta) = E\left[\log\left(Y\middle|\theta^i\right)\right] = E\left[\log \sum_z \left(Y, Z\middle|\theta^i\right)\right] \tag{1}$$

In the M steps, the EM algorithm maximizes the log-likelihood expectation obtained in the E steps and updates the estimated parameter, as shown in Equation (2).

$$\theta^{i+1} = arg \max_{\theta} E(\theta) \tag{2}$$

The estimation and maximization steps alternatively executed until the maximum likelihood is obtained. A simple example about handling the missing data problem is shown as follows. There is a Bayesian network with only two nodes, X and Y. X is a hidden node, Y is a observable node and the state can be either 1 or 0. Y is dependent on X as shown in Figure 1.

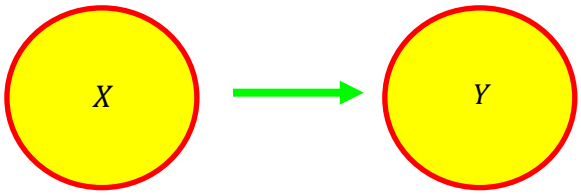

Figure 1: Two nodes, X and Y, in Bayesian network, Y is dependent on X.

Some state relationships between node X and Y with missing data are shown in Table 1. For example, in the second row of Table 1, the state of node Y is 0 when the state of node X is 1. The state of X is missing in the fourth row.

Table 1: States of X and Y.

X	Y
1	0
1	1
	0
0	0

Initially, the probability of random variable X and the likelihood are assumed as shown in Table 2 and Table 3.

Table 2: Initial probabilities of X.

$P(X)$	
$P(X=0)$	$P(X=1)$
0.3	0.7

Table 3: Conditional probability of $P(Y|X)$.

	$P(Y\|X)$	
	$Y=0$	$Y=1$
$X=0$	0.2	0.8
$X=1$	0.6	0.4

In the estimation step, the probability of $P(X|Y=0)$ is calculated for the missing data in Table 1 by using Equation 3. Then, we obtain $P(X=0|Y=0)=0.125$ and $P(X=1|Y=0)=0.875$.

$$P(X|Y=0)=\frac{P(Y=0|X)P(X)}{\sum P(Y=0|X)P(X)} \tag{3}$$

In the maximization step, Table 2 and Table 3 need to be updated from those obtained probabilities. $P(X=0)=\frac{1.125}{4}=0.28$ and $P(X=1)=\frac{2.825}{4}=0.72$. $P(Y=0|X=0)=1$, $P(Y=1|X=0)=0$, $P(Y=0|X=1)=0.65$, and $P(Y=1|X=1)=0.35$. Also, the likelihood $P(XY)$ is calculated.

In the following estimation step, the $P(X=0|Y=0)$ and $P(X=1|Y=0)$ are calculated based on the obtained probabilities from the last maximization step. The estimation steps and maximization steps alternatively executed, until the maximal likelihood is obtained. The missing state of node X is 0, if $P(X=0|Y=0)$ is larger than $P(X=1|Y=0)$, otherwise it is 1.

In order to evaluate the data imputation by using EM algorithm, we impute the EHRs with mean values and EM algorithm, and create two corresponding data sets. Those two data sets will be used by artificial neural network and machine learning models to evaluate the data imputations.

4 IMPUTATION OF MISSING DATA

The data set we used in this study is extracted from the EHRs collected from the clinics in New York City over the past 20 years. The EHRs contains the patient's vital signs, genders and ages, laboratory testing results and corresponding cancer types. There exist missing data in the EHRs, especially the laboratory testing results and vital signs because it cannot guarantee all the patients have the same laboratory tests and vital sign measurements over the past 20 years.

Firstly, the laboratory results and vital measurements of patients with cancers are extracted from the EHRs to be the original data set. The nonnumerical results and measurements in the data set are converted to numerical data. The cancer types are encoded by the ICD-9 or ICD-10 codes in EHRs, which cannot be classified by machine learning classifiers. Thus, the ICD-9 or ICD-10 codes are also converted to numerical data. The laboratory results, vital measurements, genders and ages are considered to be the features, and the numerical cancer types are considered to be the cases (labels) in the classification task. The features and cases are stored in vectors. Each vector is a data sample containing some features and a case.

Secondly, we check the data missing percentages for each feature in the data set. The EM algorithm imputes the missing data based on the existent data, and the percentages of existent data might influence the imputation accuracy of EM algorithm. There are totally 24 types of cancers with 28 features in the data set. Figure 2 shows the relationship between the threshold of non-missing data percentage and the number of features in the EHRs. In Figure 2, the horizontal axis represents the thresholds, which are percentages of non-missing data in features. For example, 0.7 means the features with non-missing data larger or equal to 70% of whole data. The vertical axis shows the number of features. Figure 2 shows that all the features have at least 60% non-missing data and only 6 features have more than 80% non-missing data.

Finally, the EM algorithm is applied on the data set to impute the missing information. We selected the cancers with corresponding samples more than 100 to be the data sets for the following classification task. This is because less or imbalanced data sets influence the classification and prediction accuracies.

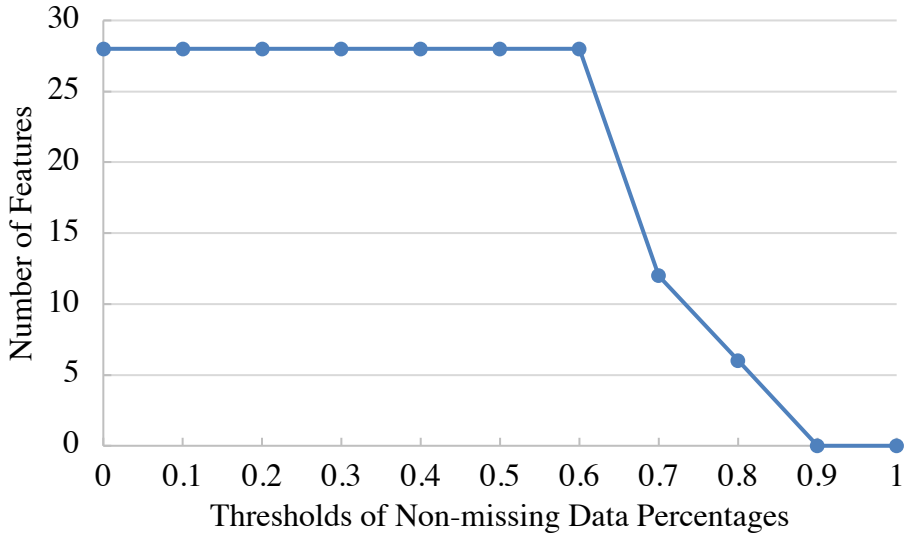

Figure 2: The relationship between the threshold of non-missing data percentage and the number of features in the EHRs.

5 EXPERIMENTS AND RESULTS

In this study, our goal is to validate the effectiveness of EM algorithm in imputing the missing data in EHRs and the cancer prediction by using them. We use an artificial neural network (ANN) model and three types of machine learning models including logistic regression (LR), support vector machine SVM with Radial Basis Function kernels and random forests (RF) to classify and predict different cancers by using the imputed data sets.

In the ANN model, we use 3 layers as the network structure, and they are the input layer, the hidden layer, and the output layer. The number of nodes in input layer is the same as the feature dimension in EHRs. The number of nodes in hidden layer is two times of the number of nodes in input layer. The number of nodes in the output layer is the same as the number of cancer types. Those three layers are fully connected.

In the machine learning models, we use cross-validation with K- $fold$ to select the parameters. The suitable parameters allow the models to obtain the highest prediction accuracies. In the K- $fold$ cross validation, the input data is divided to K portions. Each machine learning model is trained with the training data set and tested with the testing data set for K times. In each time, one classifier uses $K-1$ portions of input data as training data set and the rest one portion as testing data set. Each classifier obtains k testing scores after training and testing for k times. The average testing score is used as the final testing score to evaluate and determine the parameter settings.

The ANN, LR, SVM, and RF are trained and tested with the data set imputed by EM algorithm and mean values to evaluate the effectiveness of EM algorithm on data imputation. The data sets with different data missing percentages are also used to evaluate the influence of data missing percentages on the data imputing accuracy by using EM algorithm. The data set with lower thresholds of non-missing data percentages contains more missing data than that with higher thresholds. Figure 3 shows the testing scores of classifi-

cation by using the data set with different thresholds. The black columns in Figure 3 represent the highest prediction score among the LR, SVM, RF and ANN on the data set imputed by the mean value (the missing data are replaced with the mean of non-missing data for each feature). Only 60%, 70%, and 80% percentages are considered because the original data missing percentages are from 60% to 80%. In Figure 3, the prediction scores with higher thresholds are higher than those with lower thresholds, which indicates the EM algorithm imputes the data with more precise if there exist more non-missing data. Also, the prediction accuracies of models using data set imputed by EM algorithm are higher than that by mean values as shown in Figure 3. That phenomenon indicates the EM algorithm is more effective than imputation with mean values.

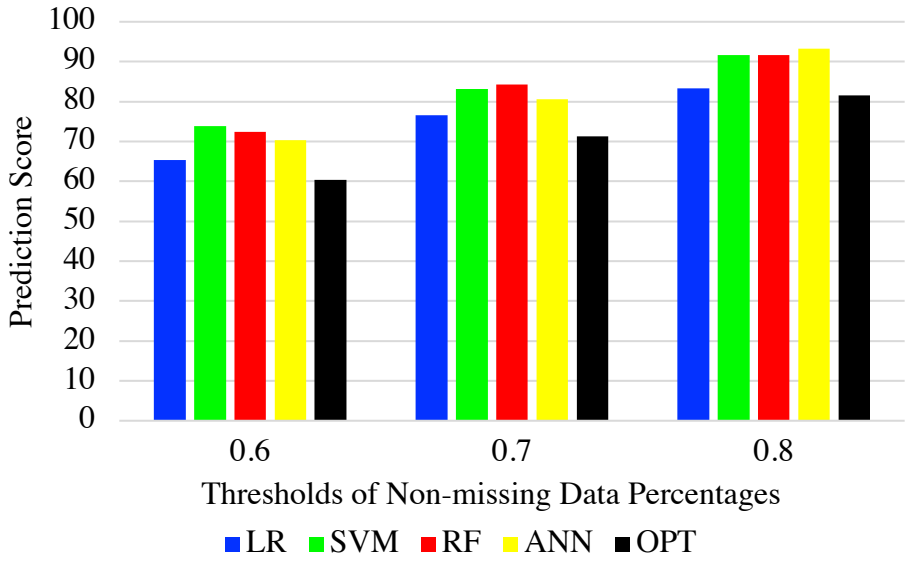

Figure 3: The relationship between the prediction scores of different classifiers and the threshold of non-missing data percentage. The OPT represents the optimal prediction scores among the LR, SVM, RF, and ANN by using the data sets imputed with mean values.

Besides the prediction accuracy, the running time of different models on the same data sets are also compared as shown in Figure 4. The ANN consumes much more time than other models no matter the data missing percentages because the ANN has a fully connected network, which has a higher computation cost, while the LR, SVM, and RF have similar time consumptions.

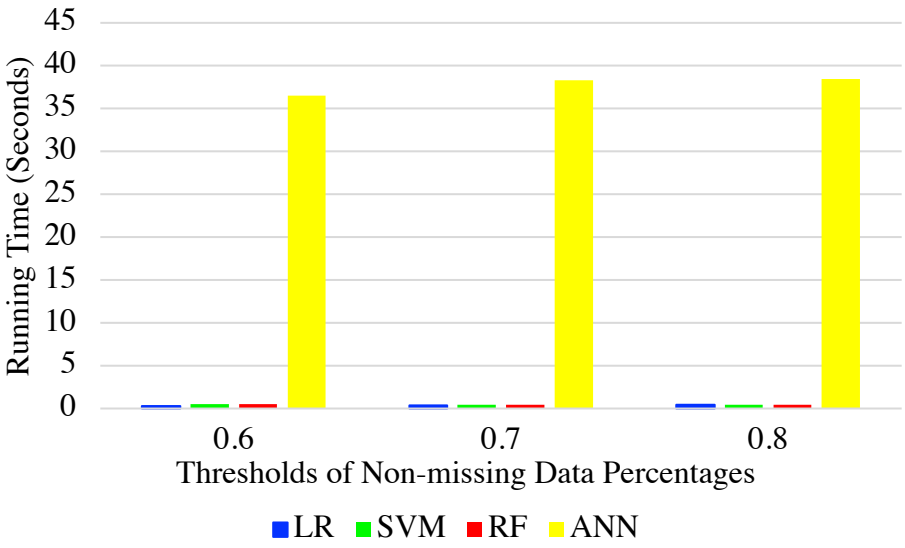

Figure 4: The relationship between the running time of different models and the thresholds of non-missing data percentages.

6 CONCLUSIONS AND FUTURE WORK

In this study, the EM algorithm is applied to impute the data set extracted from EHRs. The logistic regression, support vector machine, random forests and artificial neural network are used to classify the imputed data to verify the effectiveness of EM algorithm in data imputation. The experimental results indicate the data imputed by EM algorithm is more accurate than that by mean value. The cancer prediction accuracy is improved on the missing data imputed by EM algorithm. Our future work will focus on exploring other approaches in data imputation for EHRs and the application of imputed EHRs.

REFERENCES

Boser, B. E., I. M. Guyon and V. N. Vapnik. 1992. "A training algorithm for optimal margin classifiers". In *Proceedings of the fifth annual workshop on Computational learning theory* pp 144-152.

Carroll, R. J., A. E. Eyler and J. C. Denny. 2011. "Naïve electronic health record phenotype identification for rheumatoid arthritis". In *AMIA annual symposium proceedings* pp 189.

Catellier, D.J., P.J. Hannan, D.M. Murray, C.L. Addy, T.L. Conway, S. Yang, and J.C. Rice. 2005. "Imputation of missing data when measuring physical activity by accelerometry". *Medicine and science in sports and exercise*, Vol. 37, p.S555.

Castro, V.M., C.C. Clements, , S.N. Murphy, V.S. Gainer, M. Fava, J.B. Weilburg, J.L. Erb, S.E. Churchill, , I.S. Kohane, D.V. Iosifescu, and J.W. Smoller. 2013. "QT interval and antidepressant use: a cross sectional study of electronic health records". *Bmj* Vol. 346, pp. 288.

DesRoches, C. M., E. G. Campbell, S. R. Rao, K. Donelan, T. G. Ferris, A. Jha, R. Kaushal, D. E. Levy, S. Rosenbaum and A. E. Shields. 2008. "Electronic health records in ambulatory care—a national survey of physicians". *New England Journal of Medicine* Vol. 359(1), pp. 50-60.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm". Journal *of the Royal Statistical Society: Series B (Methodological)*, Vol. 39(1), pp.1-22.

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* Vol. pp.

Jensen, P. B., L. J. Jensen and S. Brunak. 2012. "Mining electronic health records: towards better research applications and clinical care". *Nature Reviews Genetics* Vol. 13(6), pp. 395.

Kohane, I. S. 2011. "Using electronic health records to drive discovery in disease genomics". *Nature Reviews Genetics* Vol. 12(6), pp. 417.

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis. 2015. "Machine learning applications in cancer prognosis and prediction". *Computational and structural biotechnology journal* Vol. 13, pp. 8-17.

Liao, S., J. Xiao, Y. Xie and F. Gu. 2017. "Towards use of electronic health records: cancer classification". In *Proceedings of the Symposium on Modeling and Simulation in Medicine* pp 4.

Miriovsky, B. J., L. N. Shulman and A. P. Abernethy. 2012. "Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care". *Journal of Clinical Oncology* Vol. 30(34), pp. 4243-4248.

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 Qual Saf* Vol. 23(1), pp. 8-16.

Persell, S. D., J. M. Wright, J. A. Thompson, K. S. Kmetik and D. W. Baker. 2006. "Assessing the validity of national quality measures for coronary artery disease using an electronic health record". *Archives of Internal Medicine* Vol. 166(20), pp. 2272-2277.

Schneider, T.. 2001. “Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. “Journal *of climate* Vol. 14(5), pp.853-871.

Siegel, R. L., K. D. Miller and A. Jemal. 2017. "Cancer statistics, 2017". *CA: a cancer journal for clinicians* Vol. 67(1), pp. 7-30.

Singh, H., K. Hirani, H. Kadiyala, O. Rudomiotov, T. Davis, M. M. Khan and T. L. Wahls. 2010. "Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record–based study". *Journal of Clinical Oncology* Vol. 28(20), pp. 3307.

Weiss, J. C., S. Natarajan, P. L. Peissig, C. A. McCarty and D. Page. 2012. "Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records". *AI Magazine* Vol. 33(4), pp. 33.

Wu, J., J. Roy, and W.F. Stewart. 2010. “Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches”. *Medical care*, pp.S106-S113.

Zhang, X., J. Xiao, and F. Gu. 2019. “Applying support vector machine to electronic health records for cancer classification”. In *Proceedings of the Modeling and Simulation in Medicine Symposium* pp 2.

Zheng, T., W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang and Y. Chen. 2017. "A machine learning-based framework to identify type 2 diabetes through electronic health records". *International journal of medical informatics* Vol. 97 pp. 120-127.

AUTHOR BIOGRAPHIES

XUDONG ZHANG is a PhD candidate in the Department of Computer Science at the Graduate Center of The City University of New York. His research interests include machine learning, bioinformatics, high performance computing, and modeling and simulation. His email address is xzhang5@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.

NING YU is an Assistant Professor in the Department of Computer Science at The College of Brockport, State University of New York. His research interests include bioinformatics, deep learning, and security. His email address is nyu@brockport.edu.

YIFEI GONG 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 agent based modeling and machine learning. His email address is yifeigong2014@gmail.com.

FENG GU is an Associate 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, bioinformatics, and machine learning. His email address is Feng.Gu@csi.cuny.edu.