Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database

Chen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee

Abstract—**Electronic medical claims (EMCs) can be used to accurately predict the occurrence of a variety of diseases, which can contribute to precise medical interventions. While there is a growing interest in the application of machine learning (ML) techniques to address clinical problems, the use of deep-learning in healthcare have just gained attention recently. Deep learning, such as deep neural network (DNN), has achieved impressive results in the areas of speech recognition, computer vision, and natural language processing in recent years. However, deep learning is often difficult to comprehend due to the complexities in its framework. Furthermore, this method has not yet been demonstrated to achieve a better performance comparing to other conventional ML algorithms in disease prediction tasks using EMCs. In this study, we utilize a large population-based EMC database of around 800,000 patients to compare DNN with three other ML approaches for predicting 5-year stroke occurrence. The result shows that DNN and gradient boosting decision tree (GBDT) can result in similarly high prediction accuracies that are better compared to logistic regression (LR) and support vector machine (SVM) approaches. Meanwhile, DNN achieves optimal results by using lesser amounts of patient data when comparing to GBDT method.**

I. Introduction

Making accurate prediction of disease occurrence can be of great clinical value for healthcare professionals. A highly effective data-driven predictive algorithm is desired to increase the efficiency of disease prevention and improve patient outcomes through early detection and treatment. Machine learning (ML) techniques are a set of powerful algorithms capable of modeling complex and hidden relationship between a multitude of clinical variables and the desired clinical outcome from data without stringent statistical assumptions. Further, the electronic medical claims (EMCs) database presents itself as a valuable data source due to its *large-scaled* and *longitudinal* nature of data collection process along with its *variety* in the recorded patients' health-related information. It is, hence, intuitively appealing to apply ML techniques to develop disease prediction from EMCs. However, unlike the steady growth in the application of ML methods in other industries, the utilization of ML approach in the medical records database appears only recently [1,2]. The EMCs usually cover a variety of health-care data, and the type of data varies in its structure; additionally, the complexities in handling the EMC data also result from its implicit inclusion of temporal information. These characteristics of EMCs make the systematic utilization of ML techniques challenging.

Recently, a branch of ML techniques based on deep learning approach, such as deep neural network (DNN), has achieved impressive and sometimes, breakthrough, results across a variety of artificial intelligence tasks. The approach of deep learning is inspired by the ability of human brain to abstract high-level representations from low-level sensory stimuli; these multi-leveled representations can be casted mathematically as multi-layered neural networks, and only recently, it is being able to be trained via layer-wise back-propagation to obtain tractable optimization [3]. These techniques are currently the state-of-art in the areas of speech recognition, computer vision, and natural language processing [4]. In terms of health-care applications, it has also been successfully used to perform automatic recognition of diabetic retinopathy in a very recent study [5]. As volumes of data grow in healthcare systems, such technique has also been applied to solve several other health-related problems, such as prediction of heart-failure [6] and osteoporosis [7]. Despite these works, there remains a resistance to accept deep learning widely as a clinical decision support with its inherent difficulties in obtaining interpretable analyses due to the complexities of the framework. Furthermore, it is also unclear whether such a technique do outperform other conventional ML algorithms, such as logistic regression (LR) and support vector machine (SVM), on prediction tasks using EMCs.

While there exist a vast amount of ML-based techniques across numerous fields and applications in the past decades, few limited works, if any, have systematically applied and/or analyzed DNN and other conventional ML approaches for EMC database in disease prediction tasks. In this study, we will explore whether such a large-scaled EMC data routinely collected for the purpose of health insurance claims is sufficient for deriving predictive analytics in stroke occurrence prediction using DNN and other ML methods (GBDT, LR, and SVM). This study details the framework for ML-based prediction tasks using EMC data, and we demonstrate that DNN is indeed capable of obtaining promising recognition accuracy on a separated testing dataset as compared to SVM and LR.

II. Methods

A. Database and study population

The dataset for this study is extracted from the National Health Insurance Research Database (NHIRD). The National Health Insurance program, which was implemented in Taiwan

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

CYH, WCC, PTL is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan.

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

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

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

since 1995, covers about 99% of the island's population. The National Health Research Institute (NHRI) has established a systemic sampling of patient data resulting in the available NHIRD. The database contains de-identified EMC data from over 900,000 patients treated in the hospitals and clinics from 2000 to 2008 across the nation. The details of the NHIRD have also been previously described [8,9]. In brief, these random samples of patients have been confirmed by the NHRI to be representative of the general population in Taiwan. The NHRI has further made data available in an anonymous format with extra precaution to protect the privacy of individuals. The database has already been used for several important epidemiological and medical researches [10,11]. Our study is approved by the Institutional Review Board of Taichung Veterans General Hospital.

In this work, to explore the feasibility and effectiveness of DNN and other ML methods developed for EMCs, we design a task to predict the 5-year stroke occurrence using the outpatient department database. Patients aged 0 to 99 years are identified from the database in 2003. Patients are not eligible for enrollment if they had any types of stroke (International Classification of Diseases, Tenth Revision, Clinical Modification, [ICD-10-CM] code: I60~I69) for the duration of 2000-2003. We utilize data from the outpatient department within past three years before enrollment to generate features (Section B). We further remove patients that have inadequate numbers of available clinical measurements. Following this exclusion criteria, our final dataset includes a total number of 840,487 patients. In order to develop and evaluate the model, these data are further assigned into development (90% patients for training sets, and 5% for parameter tuning) and testing datasets (5% of patients). The outcome event is defined as any ischemic stroke recorded (ICD-10-CM code: I63) in the hospital discharge diagnoses in the inpatient database. Down-sampling is performed to guarantee an almost identical class distribution between stroke and non-stroke cases.

B. *EMC Feature engineering and selection*

While this study focuses on stroke prediction, we establish a feature engineering method that is generalizable to derive other diseases' predictive analytics using EMCs. We first gather the following measurements from the record of an individual patient at the enrollment time:

- Demographic measurements: Sex and age.
- Continuous and ordinal measurements: A total of 11 continuous and ordinal measurements are presented in the dataset, including diagnostic fee, treatment fee, medicine service fee, insurance fee, self-payment fee, total health-service fee, individual medicine fee, total medicine fee, total days of prescriptions, total amount of prescriptions, and total fee of prescriptions.
- Categorical measurements: These categorical variables cover statuses or diagnoses assigned by doctors and/or the insurance bureau. We use 5 categorical measurements and map them into 221 binary values.
- Medication use measurements: A list of relevant outpatient medications is classified by the ATC codes and mapped into binary values (429 in total).

TABLE 1. A TOTAL OF 7,932 FEATURES EXTRACTED

Measurement Dimension	Temporal Dimension	No of Features
Demographics: sex and age		2
Continuous and ordinal: diagnostic fee, treatment fee, medicine service fee, insurance fee, self-payment fee, total health-service fee, individual medicine fee, total medicine fee, total days of prescriptions, total amount of prescriptions, and total fee of prescriptions (11 in total)	In past 0.25 year In past 0.5 year In past 1 year In past 2 years In past 3 years (5 in total)	55 mean values, and 55 standard deviation values
Categorical: 5 measurements map into 221 binary values		1,105
Medication use: 429 in total		2,145
Disease diagnosis: 914 in total		4,570

- Disease diagnosis measurements: a list of relevant outpatient diagnoses is classified by the ICD-10-CM codes and mapped into binary values (914 in total). In the NHIRD database, the diagnoses of diseases were coded by using ICD-9-CM code. We convert the ICD-9-CM code to ICD-10-CM code by using the code-converting sheet suggested and provided by the National Health Insurance Bureau.

In order to generate the final feature vector that can capture both the relevant clinical measurements and temporal information as input to the ML algorithm, we further utilize the time stamp of these measurements. In total, we extract 7,932 features (clinical variables) from the dataset. These features can be abstracted as combinations of two dimensions derived from the EMCs (Table 1): the measurement dimension and the temporal dimension. In this paper, the temporal dimension that we use covers 5 time periods (0.25 year, 0.5 year, 1 year, 2 years, and 3 years). For continuous and ordinal measurements, we calculate the mean value and the standard deviation over the selected time period. For the categorical measurements, the total sum over the selected time period is used. In the medication use measurements, we compute the total number of specific medication classes recorded during these time periods. In the disease diagnosis measurements, we use the total number of times that a specific diagnoses is made during these time periods.

We additionally perform feature selection to reduce and identify the most discriminative features using GBDT as a preprocessing feature selection method. We perform this calculation over the training dataset for 18 times (each time with around 5% of patients' data). In the end, we use a total of 2,007 important attributes out of 7,932.

C. *Prediction Algorithms based on DNN and other ML*

In this study, our aim is to compare DNN with other ML algorithms in deriving stroke occurrence prediction using EMCs. DNN can automatically learn feature relationships computed from the EMCs at multiple levels of abstraction [3]. The architecture of our DNN used is composed of three fully connected hidden layers. The number of neurons per hidden layer is equal to the dimension of input data, and hyperbolic tangent is used as the activation function. During the training process, the parameters of the DNN are randomly initialized. For each batch of training data, parameters of the DNN are modified gradually to decrease the cross entropy loss function.

TABLE 2. COMPARISONS OF UAR, SENSITIVITY, SPECIFICITY AND PREDICTIVE ACCURACY FOR DNN, GBDT, LA AND SVM MODELS

Model	UAR	Sensitivity	Specificity	Accuracy
DNN	0.858	0.845	0.871	0.873
GBDT	0.860	0.856	0.865	0.868
LR	0.841	0.820	0.864	0.866
SVM	0.824	0.813	0.837	0.839

The optimization algorithm used to train the network here is based on stochastic gradient descent [12]. In order to speed up the training process, we apply a simple normalization approach by scaling the feature values to a range between 0 and 1. The DNN is implemented using the Keras (2015, GitHub) toolbox.

In this work, we also utilize other ML-based classifiers, such as GBDT, LR and SVM (all implemented in python 2.7 using the scikit-learn version 0.18.0 packages). GBDT is trained using 100 boosted trees and binomial loss function. L^2-regularization is used with the strength set at 1.0 for LR method. For SVM, we use the linear kernel suitable for high dimensionality of our feature space. The tuning dataset (around 5% of total patients) is used to adjust all the hyper-parameters in these algorithms.

We apply these ML-based methods to predict patients' 5-year stroke occurrence based on their EMC-derived features in the past 3 years. Aside from the accuracy, we additionally report the unweighted average recall (UAR) to be used as a measure of performance of these ML algorithms, due to the imbalance class distribution (stroke vs. non-stroke) in the separated test set. The UAR is defined as: UAR = (A+B)/2 = (Sensitivity+ Specificity)/2

A = No of accurately predicted stroke / No of true stroke

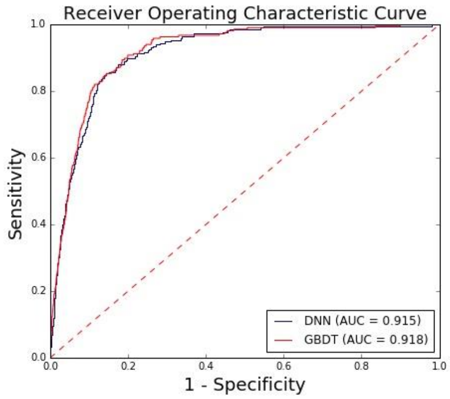
B = No of accurately predicted non-stroke / No of total non-stroke

D. *Subsampling experiments*

To understand the impact on the amounts of training information needed for the predictive model, we further perform two additional subsampling experiments:

- Reducing training data amount: To understand the effect of reducing the number of patients' data in training set, our training dataset is first divided into 9 separated parts (each sub-dataset includes around 80,000 patients' data). We iteratively add these sub-datasets into the training process of ML algorithms and examine the performance of theses ML models on a separated testing set.
- Reducing temporal information in the features: To understand the effect of reducing temporal related information, different time period is used when deriving features. In the standard predictive model, we use information of EMCs within the past 3 years before recruitment to generate the predictive features. In this experiment, we reduce the information to only past 2 years, followed by 1 year and 0.5 year. A new model is trained for each time period and its performance is measured on a separated testing set.

Figure 1. The receiver operating characteristic curve and AUC for the predictive performance of DNN and GBDT

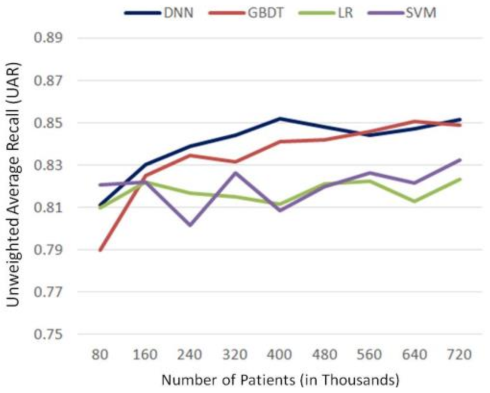


III. RESULTS

In this study, the development set includes 798,611 patients (including 756,556 in the training sets and 42,055 in the tuning sets). The testing set consists of 41,876 patients. In the follow-up 5-year period, 4,726 patients in the development set and 218 patients in the testing set had stroke events. A total of 7,932 features are generated from the datasets. After feature selection with the gradient boosting classifier, a total of 2,007 features are used in the training of the algorithms. Table 2 summarizes the overall results. The stroke prediction performance of DNN is compared with GBDT, LR and SVM. DNN and GBDT show higher UARs, sensitivities, and specificities while LR and SVM showed lower ones. DNN also has the highest predictive accuracy, followed by the GBDT, LR and SVM methods. The DNN and GBDT both achieve better classification UAR and accuracies in this study setting. These results highlight the importance of ML algorithms selection when analyzing EMCs. For a clear view of the overall predictive performance of DNN and GBDT, we plot the receiver operating characteristic curve (Figure 1). The DNN and GBDT algorithms achieve similar area under curves (AUCs) of 0.915 (95% confidence interval [CI], 0.900-0.931) and 0.918 (95% CI, 0.902-0.934). These findings indicate that DNN and GBDT can achieve similar predictive results while LR and SVM both show a less effective performance.

In the first subsampling experiment, we compare DNN and other ML methods with different training data amount. Multiple predictive models are trained with varying numbers of patients included. We aim at to determine whether the size of the training dataset would influence the performance of the algorithms. These results are obtained by testing these predictive models over the testing set. Figure 2 shows that the UAR of the model increases as we increase the training data amount in DNN and GBDT. The effects on the dataset size plateaus at around 320,000 and 560,000 patients for DNN and GBDT, respectively. Statistically significant difference (p value is 0.001 from McNemar's test) is noted between prediction result from DNN and GBDT at the dataset size of 400,000 patients. Furthermore, we observe that the performances of the model seem to be less stable (jumps abruptly up and down) as we increase the number of patients for models of LR and SVM. Moreover, the modeling power of both DNN and GBDT surpass both LR and SVM when increasing the number of patients over 240,000 patients.

Figure 2. Performance of DNN, GBDT, LR and SVM models with different training data amount



In the second experiment, we want to explore the impact of time period included in the feature extraction process on the performance of these predictive models. As showed in figure 3, we see that by increasing the number of time periods included in the feature extraction process (i.e., inclusions of an individual's longer temporal clinical information), the performance increases in DNN and other ML models. The performances of DNN and GBDT models show a slight upward improvement; meanwhile, the performances of LR and SVM models yields a relatively greater improvements as we increase the time period for training. This result suggests the inclusion of longer past history of clinical information may help in developing the stroke occurrence predictive model.

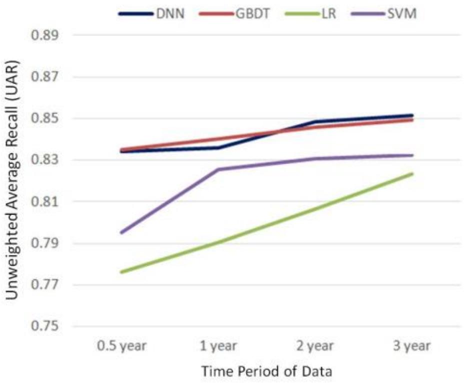
IV. DISCUSSION

In summary, we demonstrate that it is promising to utilize ML-based technique (DNN and others) on a large-scale EMCs to predict stroke with high UAR and accuracy. In this study, an encouraging AUC of 92% is achieved by both the DNN and GBDT algorithm while DNN requires lesser amount of training data. This novel approach in developing automated system for the prediction of stroke occurrence potentially offers several advantages, including consistency of results, high accuracy, and rapidly reporting of predictions. In addition, because these predictive algorithms can have multiple operating points, its sensitivity and specificity can be adjusted to match the clinical requirements. Our results also showed that performances of DNN and GBDT are superior to that of LR and SVM, both in terms of predictive performance as well as predictive stability. This difference may result from the nature of nonlinear modeling power in both DNN and GBDT algorithms. In this work, we demonstrate that DNN can be a promising method to model and extract the implicit correlations among features from EMCs that can handle complex disease prediction tasks.

V. CONCLUSIONS

In this evaluation of applying ML-techniques by using EMCs from outpatient department, algorithms based on DNN and GBDT can achieve high UAR and AUC for prediction of future stroke occurrence. Using longer time periods of EMCs data can help improve predictive power. Meanwhile, DNN can achieve the best performance using smaller training dataset compared with the GBDT method. Further research is necessary to determine the feasibility of applying DNN in the clinical setting and to determine whether the use of DNN could lead to improved clinical care and patients' outcomes – inspiring the use of appropriate algorithms in deriving transformative clinical informatics and bringing an era of ML-based decision support in health-care service.

Figure 3. Performance of DNN, GBDT, LR and SVM models with different temporal information in the features



REFERENCES

[1] O. Arandjelović, "Prediction of health outcomes using big (health) data," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2543-2546, 2015.

[2] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," Springerplus, vol. 5, p. 1410, 2016.

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

[4] H. Du, M. M. Ghassemi, and M. Feng, "The effects of deep network topology on mortality prediction," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2602-2605, 2016.

[5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, vol. 316, pp. 2402-2410, Dec 13 2016.

[6] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, et al., "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2530-2533, 2015.

[7] S. K. Kim, T. K. Yoo, E. Oh, and D. W. Kim, "Osteoporosis risk prediction using machine learning and conventional methods," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 188-191, 2013.

[8] C. Y. Hung, C. H. Lin, W. Loh el, C. T. Ting, and T. J. Wu, "CHADS(2) score, statin therapy, and risks of atrial fibrillation," Am J Med, vol. 126, pp. 133-40, Feb 2013.

[9] C. Y. Hung, Y. C. Hsieh, C. H. Li, J. L. Huang, C. H. Lin, and T. J. Wu, "Age and CHADS2 Score Predict the Effectiveness of Renin-Angiotensin System Blockers on Primary Prevention of Atrial Fibrillation," Sci Rep, vol. 5, p. 11442, Jun 22 2015.

[10] C. Y. Wu, Y. J. Chen, H. J. Ho, Y. C. Hsu, K. N. Kuo, M. S. Wu, et al., "Association between nucleoside analogues and risk of hepatitis B virus-related hepatocellular carcinoma recurrence following liver resection," JAMA, vol. 308, pp. 1906-14, Nov 14 2012.

[11] C. J. Shih, H. Chu, P. W. Chao, Y. J. Lee, S. C. Kuo, S. Y. Li, et al., "Long-term clinical outcome of major adverse cardiac events in survivors of infective endocarditis: a nationwide population-based study," Circulation, vol. 130, pp. 1684-91, Nov 04 2014.

[12] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, et al., "Large scale distributed deep networks," Adv Neural Inf Processing Syst, pp. 1223-1231, 2012.