Deep Learning from Sleep Big Data Predicts Cardiovascular Disease Risk

CSE6250 2019 Spring

Team 12: Yimei Tang, Xiwen Cheng, Zhichen Guo

Introduction

About ⅓ of the Americans suffer from sleep disorder. Not only does sleep disorder decrease work productivity, lower quality of life and increase motor vehicle accidents, but poor quality of sleep also predisposes risks for numerous comorbid human diseases including but not limited to cardiovascular diseases (CVDs), hypertension (HTN), diabetes mellitus (DM), alzheimer's disease (AD) and cancers. Sleep disorder in U.S. has an estimated $150 billion annual economic cost and a compelling medical need has not been met. Sleep data present rich metric features; however, the utilization of deep learning techniques to model them has been limited thus far. Current challenge lies in development of tool set for engineering these features into a robust prediction model for disease risks. Here, we show a trained convolutional neural network (CNN) model to robustly predict CVD risk from sleep data features derived from the Sleep Heart Health Study dataset (SHHS)[6-9]. This model has a model to predict sleep disorder-associated CVD risk.

Sleep data has rich and unique metric features such as spectral and band signatures, where stages of sleep are determined by physicians with sophisticated domain knowledge. These stages, associated with different physiological characteristic, are previously used to diagnose sleep disorders in heuristic manner. Big data and deep learning have made it possible to explore the digital features of sleep data.

DataSets

The SHHS dataset from sleep data

The Sleep Heart Health Study (SHHS) dataset is a multi-center cohort study carried out by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing.[6-9] The dataset collected sleep data from 5600 patients in two clinical visits with an interim visit or phone call follow-up between. During these two visits, PSG measurement was performed to collect the sleep data. The index date is set at the 1st visit. The CVD feature data was collected between these two visits. The data was first through sanity check, features extraction and selection. We used Cardiovascular Disease (CVD) death (cvd_death) as the target variable in this study by using CNN model. This is a classification machine learning problem.

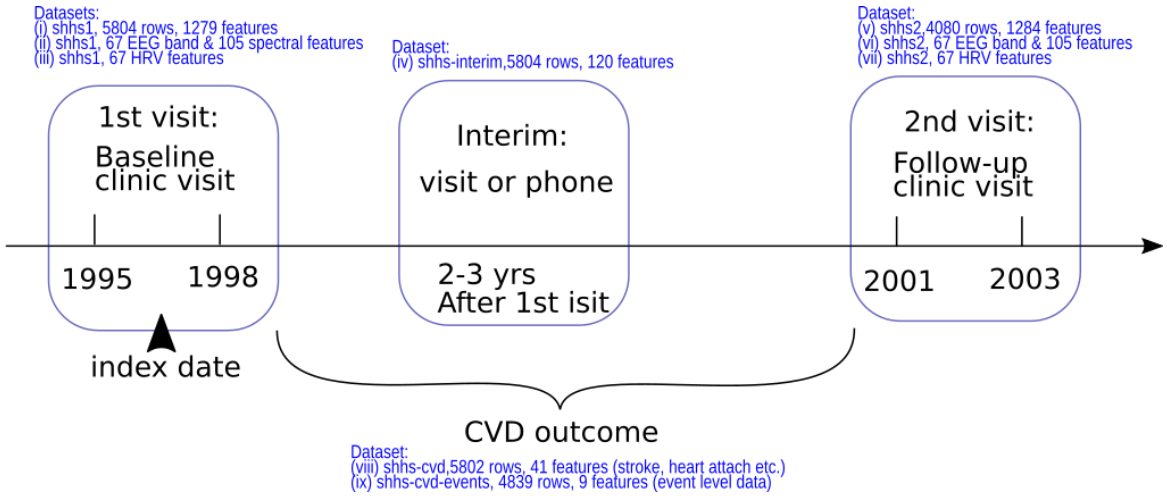

Figure 1| The SHHS data sets

Data Preprocessing & Insights

Feature Selection Strategy One - AHI

The variables considered will be apnea/hypopnea index (AHI), defined as the number of apneas and hypopneas per sleep hour, identified if at least a 3% desaturation or an arousal occurred in association with a change in breathing. CVD is associated with increased AHI in subjects with baseline AHI >5, but not in subjects with AHI <5[12]. We selected 13 features of AHI, namely, ahi_a0h3, ahi_a0h4, ahi_a0h3a, ahi_a0h4a, ahi_c0h3, ahi_c0h4, ahi_c0h3a, ahi_c0h4a, ahi_o0h3, ahi_o0h4, ahi_o0h3a, ahi_o0h4a and oahi. We fill null values of these features with mean values. Based on a decision tree classification algorithm, implemented with the help of scikit-learn libraries, we successfully predict patients with high CVD risk and the accuracy is about 75%.

Feature Selection Strategy Two- AHI

We first get rid of all the features that have more than 5% blank values since we believe more than 5% incomplete data will greatly reduce the credibility of the imputed data. After filtering out these low-quality features, we have 632 features. Due to the way when the data was collected, most of the categorical data have "float" type which is incorrect. Therefore, we first correct those mis-classified features by making these fake "numerical" data to categorical data. After this step, we have 418 numerical features and 213 categorical features. Then, we use the median to impute the missing data for numerical data and the mode to impute the missing data for categorical data.

Since the target variable is binary, we calculate point biserial correlation coefficients(between -1 and 1, positive number means positive correlation and 0 means no correlation) between each numerical feature and the target variable to get the top 40 relevant features. The top three important numerical feature are age_s1(0.246), age_s2(0.245) and FEV1(-0.149). age_s1 stands for "Age at Sleep Heart Health Study Visit One" means the older the patient is, the more likely that he will die from CVD. FEV1 means how much air a person can exhale during a forced breath in the first second, the less the person can exhale, the more likely that he will die from CVD.

For the categorical features, we use chi-squared test(result is a p-value for the null hypothesis that these two categorical variables are independent) to get the top 40 relevant features. The top three important categorical feature are timeremp,rdi0p and rdi2p. timeremp represents Percent Time in rapid eye movement sleep (REM). rdi0p represents Overall Respiratory Disturbance Index (RDI) all oxygen desaturations and rdi2p represents Overall Respiratory Disturbance Index (RDI) at >=2% oxygen desaturation.

Feature Selection Final Strategy

Since the dataset is imbalanced, as we have 198 patients that died from CVD and 3882 that are not. When evaluating CNN model results, we use recall and precision as the metrics because accuracy will give us false sense of the model achieving high accuracy. Since the cost of failing to identify patients that have high risk of CVD, we should not use accuracy metric but instead we should use recall and precision metrics since the cost of erroring on side of caution is much smaller than the cost of having false negative.

After using CNN models to run on above mentioned two sets of features, we found the results far from satisfying. The accuracy score is high but the recall and precious scores are extremely low. Upon further research on the two sets of features, we found out actually they have many features in common, for example: age_s1,timeremp and ahi_o0h3. Therefore, we use the sklearn.neural_network.MLPClassifier to pick the features that have higher contribution to the recall score of the classifier and the final result is much better than using either feature list to train the CNN. The final feature list has 30 features down from 80 features from two seperate lists originally.

Method

The convolutional neural network

We built a convolutional neural network using two convolutional layers. The first layer has 8 filters of kernel size of 5 and stride set at 1. The second layer has 32 filters with kernel size of 5 and stride set at 1. Both layers are followed consecutively by a rectified linear unit (ReLU) activation and a max pooling layer with the size as well as stride of 2. There are two fully connected layers, one builds 256 hidden units followed by ReLU activation and the other one builds 2 units as the output layer. The model was implemented using *PyTorch* for tensor computation and deep neural network learning and with the help from Python packages including Pandas, SciPy/Numpy, Scikit-learn libraries. We used cross-entropy criterion to train and optimize our neural network.

Model training, evaluation and test

The cleaned data set will be randomly split into training and test data sets with a ratio of 8:2. The model is trained through 10 epochs with cross-entropy criterion. The accuracy metrics were plotted for assessing the model fitting. Parameters were tuned to avoid overfitting and underfitting. To evaluate the cost of fitting, we also plotted the confusion matrix. The optimal parameters were determined in the best trained model. The trained model was used for prediction on the test data set. And the same metrics were evaluated to assess the model performance.

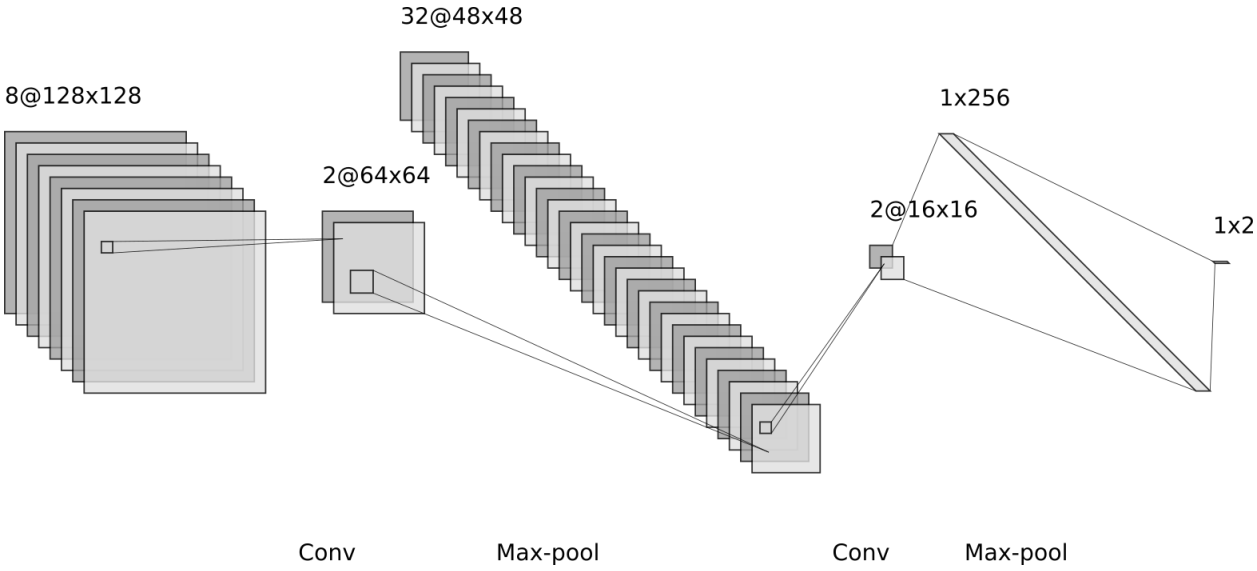

Figure 2| The CNN model

Result

CNN model built on the selected features

As described in material method, we train CNN model through 10 epochs with cross-entropy criterion on selected 30 features. We successfully predict the mortality of patients and the accuracy is about 97%. The precision is about 98% and the recall is about 85% and performance of the CNN model is very well (Figure 3, 4). The training process is going well and the model is not overfitting and underfitting with number of epochs increasing.

Discussion

Sleep data and CVD death relationship

After evaluating three CNN models using different sets of features, we found out that some features are much more relevant than other features in classifying the CVD death. For example, in our final strategy of feature selection, we have these following features that are crucial in classifying CVD death.

1. age_s1: Age at Sleep Heart Health Study Visit One. This means older people are more likely to die from CVD given all other conditions are same.

2. remepop: Sleep Time (Rapid eye movement sleep (REM) in minutes. This means that how many minutes the patient has in REM phase has relationship with the risk of the patient will die from CVD.

3. timeremp: Percent Time in rapid eye movement sleep. This means the percentable of time that the patient has in REM has relationship with the risk of the patient will die from CVD.

4. The final result of CNN shows that the recall score is 0.85 and the precision score is 0.98 for the test set. This means for every true case the model predicts (the patient will die from CVD), it is 98% probable that this the patient is indeed die from CVD.

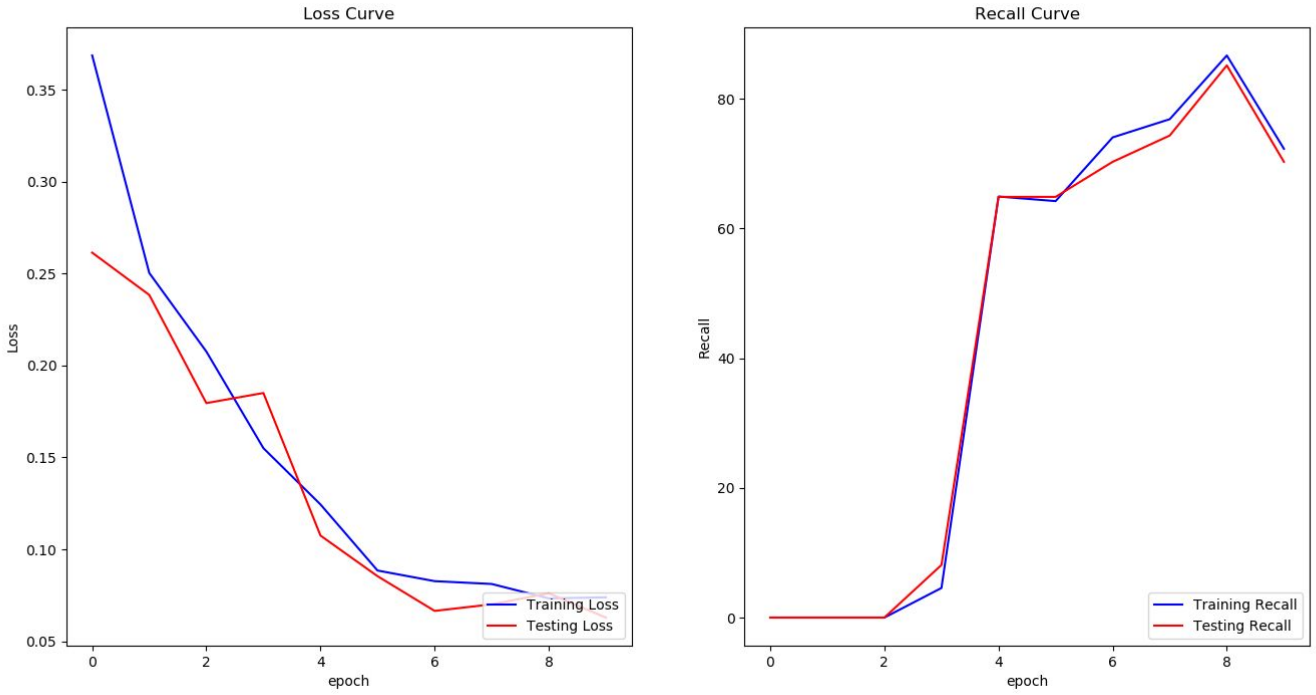

Figure 3| CNN recall learning curve.

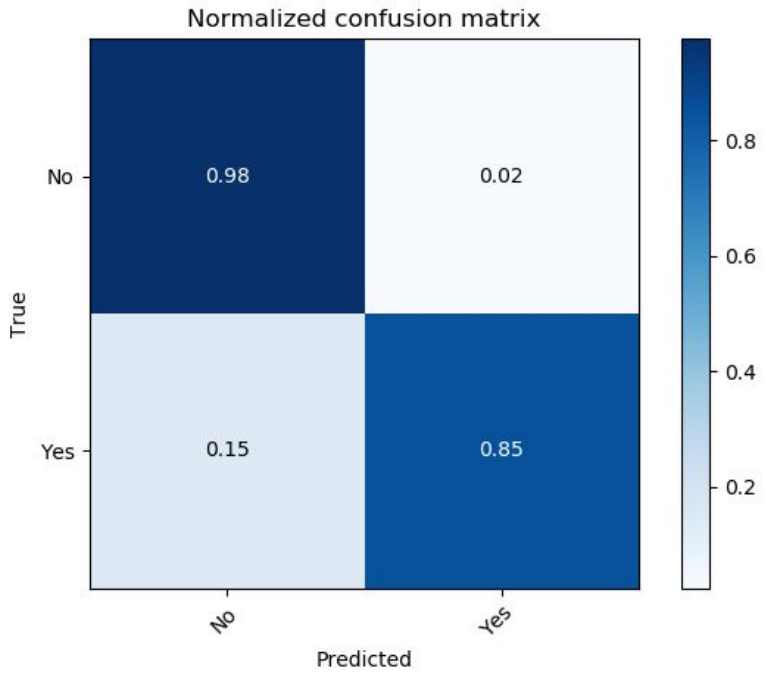

Figure 4| Confusion matrix for CNN model

Sleep data can predict disease

Latest work[1] has shown sleep disorder as an early indicator for accumulation of β-amyloid (Aβ) and tau proteins, two important pathological markers in preclinical AD, even in cognitively healthy senior adults. This suggests prediction of relevant sleep disorder at early stage may directly benefit diagnostics of preclinical disease. Sleep polysomnography (PSG) is the golden standard to diagnose sleep disorder and analysis of PSG heavily relies on domain knowledge from well-trained physicians; it remains a gap to interpret pathological sleep disorder from PSG in an unbiased and systematic manner. Currently, data analysis approaches have been demonstrated a robust and reliable way to extract sleep stage features from the PSG signals. For example, Tsinalis et al. [2] developed automatic sleep stage scoring algorithm using convolutional neural networks (CNNs) on single-channel electroencephalography (EEG). Besides, Biswal et al. [3] also successfully developed an algorithm to classify 5 different sleep stages N1, N2, N3, REM and Wake from EEG signal using recurrent neural network. Additionally, Zhao et al. [4] demonstrated it feasible to extract sleep-specific invariant features from RF signals. Together, these progression in the field suggest data analysis can automate PSG signal phenotyping and sleep data can be sampled at a big-data scale from non-invasive devices. Thus, a new avenue has been opened for big-data approach to build a sophisticated predictive model to detect disease-related sleep disorder in real-time in immediate future.

Neural network model best suits the characteristics of sleep data

Previous works on analysis of SHHS dataset has shown that a number of key sleep-related variables in the SHHS dataset are not normally distributed or highly skewed [7]. This makes generalized linear model (GLM) based logistic regression analysis is not quite suitable due to violation of statistic assumption. Here we successfully built and trained a convolutional neural network nonlinear regression model which is able to robustly classify and predict cardiovascular disease-associated clinical features from sleep data features. To further tune-up the model, we could consider the nature of the CVD disease where the cost of false negative may be higher than false positive, and we consider that in model evaluation accordingly.

Improvement of the model toward a marketable medical device

Lately, the Food and Drug Administration (FDA) has granted de novo classification to market clinical support software applications under the FD&C act Section 513(f)(2) after a Not Substantially Equivalent (NSE) determination in response to a 510(k) submission. Alternatively, a software-based medical device can also file de novo classification application to FDA under section 513(a)(1) given no legally marketed device upon which to base a determination of substantial equivalence comparison without initial submission to 510(k) [11]. That means machine learning software can be marketed as a novel medical device if it supersedes the physician decision performance or per se is first of class on the market. The neural network on sleep dataset has reliable prediction power for CVD risk, as first of class or the superseded among others, can be developed into a novel medical device to provide critical and essential clinical support to the market.

Challenges and perspectives

Imbalance of dataset requires justification of model metrics

Since the dataset is imbalanced, we can't fully rely on accuracy score to evaluate the models and it is easily subject to overfitting or false sense of the model is accurate. We should be highly alert to this kind of situation and choose the right metrics to evaluate the model.

Missing value imputation

Almost 72% of the sshs1 dataset and as high as 91% of the sshs2 observations contain certain missing values. How to impute these missing values is worth further in-depth investigation. Many times, the missing values are generated by deterministic feature yet to be discovered. One can't simply drop them or use simply imputation method. Other sophisticated imputation methods, such as Bayesian way may be further implemented in the CNN model.

Mutual benefits between machine learning and domain expert

Many of the essential digital features we have discovered in our model are not readily transferable to known domain knowledge on the diseases. On the other hand, these novel disease-associated features shall surely

shed a light on the pathophysiological research on the diseases. A multidisciplinary team between the machine learning engineers and the domain experts shall be formed to better understand the disease.

Feature engineering on CVD outcome remains a challenge

Currently, the response variable is univariate to simplify the model. In realty, the clinical outcome of CVD is multivariate and various clinical endpoints can be achieved. A more complicated multivariate model can be derived; however that requires a deep understanding of CVD features, which is currently lack from the domain knowledge. Meanwhile, other feature engineering methods shall be explored, including but not limited to, PCA, randomForest, graphic network analysis. Besides, the time-domain of the data is not fully incorporated in the current analysis. More feature engineering will be beneficial to classify CVD death given the data is semi-temporal. For example, a significant arm of patients visited the clinics from 1995 to 2003. Utilization these time-domain information, for example in RNN (Recurrent Neural Network) model, may discover more essential hidden features linked to CVD from the sleep data sets.

References

1. Carvalho DZ, St Louis EK, Knopman DS, et al. Association of Excessive Daytime Sleepiness With Longitudinal β-Amyloid Accumulation in Elderly Persons Without Dementia. JAMA Neurol. 2018;75(6):672–680. doi:10.1001/jamaneurol.2018.0049

2. O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. arXiv preprint arXiv:1610.01683, 2016.

3. S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. Brandon Westover, M. T. Bianchi, and J. Sun. SLEEPNET:Automated sleep staging system via deep learning. 26 July 2017.

4. M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In International Conference on Machine Learning, pages 4100–4109, 2017.

5. Stephen Harvell, Giulio Borghesi, Olutosin Sonuyi. Single channel EEG sleep stage scoring using a Recurrent Neural Network.

6. Dean DA 2nd, Goldberger AL, Mueller R, Kim M, Rueschman M, Mobley D, Sahoo SS, Jayapandian CP, Cui L, Morrical MG, Surovec S, Zhang GQ, Redline S. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016 May 1;39(5):1151-64. doi: 10.5665/sleep.5774. Review. PubMed PMID: 27070134; PubMed Central PMCID: PMC4835314.

7. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 May 31. doi: 10.1093/jamia/ocy064. [Epub ahead of print] PubMed PMID: 29860441.

8. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. PubMed PMID: 9493915.

9. Redline S, Sanders MH, Lind BK, Quan SF, Iber C, Gottlieb DJ, Bonekat WH, Rapoport DM, Smith PL, Kiley JP. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. Sleep. 1998 Nov 1;21(7):759-67. PubMed PMID: 11300121.

10. Sleep Heart Health Study Data Analysis Tip Sheet. https://sleepdata.org/datasets/shhs/pages/03-data-analysis-tip-sheet.md

11. Device Classification under Section 513(f)(2)(de novo), Food and Drug Administration, 2017

12. Association of Incident Cardiovascular Disease With Progression of Sleep-Disordered Breathing, https://www.acc.org/latest-in-cardiology/journal-scans/2011/03/24/23/39/association-of-incident-cvd-with-progression-of-sdb, Circulation 2011;123:1280-1286