Predicting Cardiac Disease With Deep Learning

Taylor Archibald, Corey Woodfield, Jesse Robinson, and Benjamin Bay
December 2017
CS 478–Machine Learning
Brigham Young University

Abstract

In this project, we build a classifier to predict whether or not a given person has had coronary heart disease in his or her lifetime. We do so by using the formidable NHANES patient health data set, cleaning this data extensively for our purposes, and training a deep neural network on it. This paper discusses our data processing methods, our investigation of various machine learning models, our chosen deep learning process, and our preliminary results. Our model achieves success while leaving room for further improvements. We discuss possibilities for future expansion of this research, such as optimizing DNN hyperparameters and classifying on additional diseases.

1 Introduction

The purpose of this project was to identify individuals with an elevated risk of coronary disease. With heart disease being the leading cause of death in America accounting for over half a million deaths annually [MEMBERS *et al.*, 2014], being able to efficiently determine the potential diagnoses of coronary disease could be invaluable in preventing thousands of deaths.

Though heart disease diagnoses can often be determined by a set of highly specific blood tests, there is presently no reliable way of diagnosing heart disease using metrics routinely measured by medical professionals. Thus, our report covers our attempts to pick out these features that are both relevant to heart disease and routinely obtained by medical professionals, and use these features in a machine learning algorithm to predict those with an elevated risk of having heart disease. Our data comprises features in the National Health and Nutrition Examination Survey (NHANES) [Damico, 2013].

We explain the difficulties in transforming the disjointed, modular NHANES data into a usable dataset, in addition to troubles encountered identifying features and implementing a deep neural network.

1.1 Data Source

The source of our data is the National Health and Nutrition Examination Survey (NHANES). This survey is conducted by the National Center for Health Statistics (NCHS) and was first conducted in 1971, assessing the health and nutritional status of citizens of the United States. The NHANES data consists of three distinct components:

1. Laboratory tests, including a comprehensive analysis of the individual's blood and urine and a disease profile.
2. Comprehensive health survey, including questions regarding the individual's demographic, socioeconomic status, diet, immunization history, physical activity, occupation, drug and tobacco use, disease history, and other health-related questions.
3. Results from physical examination, which include oral, optic, dental, and other physiological tests and measurements.

We obtained 14 years of NHANES survey data from the public NHANES website [Damico, 2013]. The data comprises over 1,000 different component files, broken up by topic and survey year, which are hosted in various locations on the website. We wrote a webscraping script in Python to traverse the NCHS website and download the files as they were identified.

1.2 Data Processing

In its native state, the size and dimensionality of this data made both the selection and compilation of features difficult. To get a sense of its scope, consider that NHANES carries nearly 10,000 unique features, and over 330,000,000 feature-observations ranging from details of socioeconomic status to dental records. We made detailed processing scripts were in order to:

- convert the native binary XPT files to CSV,
- create translations for and decode file and variables names,
- produce a summary containing statistics for each variable, including the range, mean, median, and percent missing,
- join all CSV files together using the appropriate respondent identifier,

- analyze the data for consistency, identify continuous variables, handle anomalous values, handle missing values, and handle duplicated data, among other preprocessing procedures, and
- combine redundant variables and account for inconsistent naming conventions between surveys.

We generated summary files for each individual data file, and compiled these summaries into a master summary, which we reviewed and identified potentially useful features for our analysis.

Data Limitations

At this point, we identified several concerns regarding our data. One major concern was how frequently features were missing that we had otherwise intended to use in our model. For instance, many laboratory measures, including most of the bloodwork ones, were only populated about 30% of the time. Moreover, it was not immediately obvious from the documentation how these were sampled.

Another issue was we did not have ideal time coverage for every variable. Hence, we had to balance utilizing as many years of data as we could to maximize the number of observations, while simultaneously considering how often variables were missing. Moreover, a nontrivial amount of work was put into pairing and adapting variables that conveyed some the information, but were recorded differently–for instance, depending on the year the survey was taken, ages were recorded in months, years, or both months and years (depending on the person's age). All of these discrepancies had to be rectified for each feature we wanted to use.

Since the NHANES data set are massive and subject to human error, it was not too surprising that many data fields contained erroneous values or were frequently missing. We filled in missing continuous variable fields with the feature mean, and created a unique *unknown* class for missing categorical variable fields. We deleted all other rows with erroneous data or missing labels.

Hence, though the data are extremely feature rich, we were not able to fully evaluate every potentially relevant variable as we had initially hoped. We ultimately were able to process some 58 features we thought are routinely recorded by medical professionals and are potentially relevant for identifying those at risk of heart disease. Table 1 shows these final chosen features for each patient's data point.

Another minor setback we discovered early on was there were no diagnoses made during the survey–the only indication we had regarding who suffered from coronary heart disease were self-reported answers to the question "Ever been told you have coronary heart disease?" While we feel this is still a fairly good metric of actual coronary heart disease, potential weaknesses include that it was self-reported, there is no uniform diagnosis standard, and some respondents may have coronary heart disease, but have never been diagnosed. However, others in the literature have relied on these metrics with good success[Lee and Giraud-Carrier, 2013].

Balancing Data

To account for the imbalance among label classes (i.e. there are many more respondents without heart disease than with it), we employed an oversampling technique. One way to think about this is, we effectively duplicated positive heart disease observations until we had roughly the same number of positive and negative samples in our training dataset. This was necessary to ensure our model did not arrive at the trivial solution–people rarely have heart disease, so even a classifier that always guessed "negative" would achieve a comfortable-looking 95% accuracy.

Feature Selection – Decision Tree

One of the primary tasks early on was to evaluate which features were most important. In a very early attempt to analyze which features were most important, our group ran a custom implementation of ID3 to see what kind of a decision tree would be built given varying inputs. For handling unknown data, this implementation would either count the unknown data as its own nominal class or as the mean continuous value, depending on whether its feature was nominal or continuous.

This implementation was run on a number of different groups from the data, including: varying diastolic and systolic reads of blood pressure; different levels of protein and metals within the blood; features that don't require biological labs or testing, such as age, gender, and annual family income; and other measures of the body that could be performed with common measuring equipment, including height, weight, and waist circumference.

When the aforementioned ID3 algorithm was run on the data of these grouped features, the resulting structures of the decision trees (especially their root features) were recorded and randomly sampled in order to gather a list of some of the most critical features for providing information gain on whether or not a patient had been diagnosed with coronary heart disease. This initial list included the measurements in Table 2.

One relatively surprising feature within this list was folic acid, especially since this feature was initially grouped with Direct HDL-cholesterol for its run through our ID3. This means, despite sources more commonly urging to test for cholesterol levels in a patient at risk for coronary heart disease [Torpy *et al.*, 2009], folic acid—which has a comparable cost for blood testing analysis [Walgreens Introduces...,][Folate...,]—could provide a greater information gain on whether that patient is truly at risk.

The ID3 algorithm was then run on this list of the most important features, and the structure of the decision tree produced was again inspected and analyzed. One of the more noteworthy findings using this approach was that the root node of the

Table 1: Chosen Features

Year
Ever been told you have asthma?
Doctor ever said you were overweight?
Ever told you had a stroke?
Ever told you had emphysema?
Ever told you had chronic bronchitis?
Ever told you had any liver condition?
Ever told you had cancer or malignancy?
Pulse regular or irregular?
Pulse type
Gender
Ever told you had a thyroid problem?
Close relative had heart attack?
Breathing problem require oxygen?
Problem taking deep breath?
Ever been told you have psoriasis?
Ever been told you have celiac disease?
Are you on a gluten free diet?
Doctor told you to lose weight?
Doctor told you to exercise?
Doctor told you to reduce salt?
Doctor told you to reduce fat?
Are you now controlling or losing weight?
Are you now increasing exercise?
Are you now reducing salt in diet?
Are you now reducing fat in diet?
Ever been told you have jaundice?
Systolic Blood pres (mm)
Diastolic Blood pres (mm)
Sagittal Abdominal Diameter (cm)
C reactive protein (mg/dL)
Lead (ug/dL)
Cadmium (ug/L)
Mercury total (ug/L)
Mercury Inorganic (ug/L)
Total cholesterol ($mmol/L$)
60 sec pulse 30 sec pulse 2
Weight (kg)
Standing Height (cm)
Body Mass Index (kg/m^2)
Waist Circumference (cm)
Age at Screening Adjudicated
Age in Months
Annual Family Income
Direct HDL Cholesterol (mg/dL)
Folic acid serum ($nmol/L$)
Mercury ethyl (ug/L)
Mercury methyl (ug/L)
Blood selenium (ug/L)
Blood manganese (ug/L)
Average Sagittal Abdominal Diameter (cm)
Increased fatigue
High cholesterol
5,10-Methenyl-tethrofolic acid ($nmol/L$)
5-Formyl-tetrahydrofolic acid ($nmol/L$)
hours watch TV or videos past 30 days
of hours use computer past 30 days
Ever told you had coronary heart disease?

Table 2: Six possible top features as suggested by decision tree learning

Label	Type
C-Reactive Protein	Continuous
Systolic: Blood pres (1st rdg) mm Hg	Continuous
Body Mass Index	Continuous
Waist Circumference	Continuous
Age at Screening	Discrete
Folic Acid, serum	Continuous

tree was BMI, suggesting that BMI could be one of the most critical features to assess to determine one's risk for coronary heart disease.

Further examination and random sampling of the tree also showed a trend depending on a participant's BMI (recorded as units of kg/m**2). For those with a measurement below twenty or above thirty, the next feature's class was most often folic acid. For those with a BMI between twenty or thirty, however, the next feature's class was consistently waist circumference. This finding would mean, if our ID3 implementation was accurate, that a patient concerned with their risk of coronary heart disease should first inspect their own BMI. If this measurement is between twenty or thirty—which roughly correlates with classifications of "normal" and "overweight"—then the next most important feature to check is their waist circumference. However, if the BMI is higher or lower than this range—roughly correlating with classifications of "underweight" or "obese"—then a folic acid blood test should be their next step. This is likely because if BMI is between 20 and 30, waist circumference can be used with the BMI to help determine the patient's overall health, however, if they are already severely under- or overweight, the information we would've gained from waist circumference can instead be inferred just from BMI.

Of course, these findings were rough analyses based upon random sampling of the decision tree produced by our ID3 algorithm. These conclusions could certainly be further refined, as they depend on which features were inspected and how they were grouped. Many features had large stretches of "unknown" records, and thus a feature's unknown data—or lack thereof—could have easily biased its calculated information gain in our decision tree.

Thus, we recognize that these findings are by no means conclusive, and with initial testing within our deep neural network, we quickly found that our accuracy was disappointing with just the six features found above. We consequently further processed our data, refining our pool of potential features to further improve our model.

1.3 Feature Selection – Principal Component Analysis

To further help determine which features were relevant and necessary for inclusion in the data for the neural network, we used Principal Component Analysis (PCA) [Wold *et al.*, 1987]. This effectively compressed our data from 58 features to 21. We retained enough information in the principal components to account for 90% of the variance, and we allowed up to 5 features from the original feature space to make up each feature in the new feature space. We eliminated any feature with an eigenvalue less than 1.0. We examined the features that resulted from PCA, and though not always easily interpretable, they were nevertheless informative. From a brief inspection, there seems to be a correlation between time spent on the computer or watching TV and having asthma, and being on a gluten free diet seems to correlate to having jaundice, and the likelihood of having a stroke.

Figures 1, 2, and 3 show the results of running our principal components through our DNN. We explore details of the DNN in section 2.2.

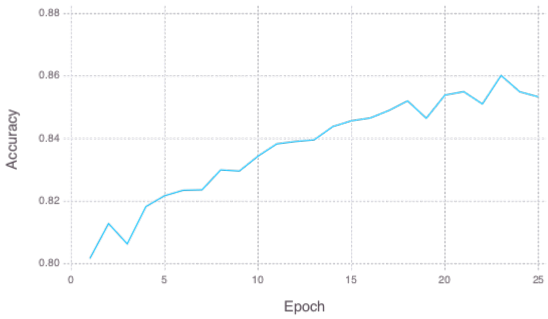

Figure 1: Accuracy over 25 epochs of DNN training on principal components.

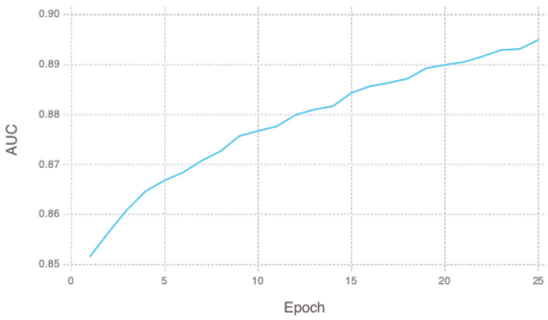

Figure 2: Area under curve over 25 epochs of DNN training on principal components.

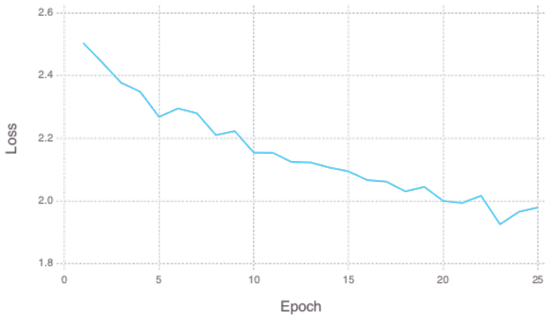

Figure 3: Loss over 25 epochs of DNN training on principal components.

1.4 Deep Learning

Once we collected our preliminary data, we developed a deep neural network using the TensorFlow machine learning library [Abadi *et al.*, 2016]. This network was designed to classify and predict disease diagnoses based on the features we had identified. This was done by taking a vector of patient data as input, utilizing several hidden fully-connected layers, and outputting a probability for the likelihood of having coronary heart disease.

Topology

We used 4 fully-connected hidden layers holding [100, 75, 50, 25] nodes, respectively.

Activation Function

We used the rectified linear unit [Nair and Hinton, 2010] as our activation function. This was done based primarily on recommendations from top researchers in the deep learning community.

Normalization

For normalization and in order to reduce single node independence, we used dropout [Srivastava *et al.*, 2014] between each layer, settling on $p = 0.5$. This means that for each hidden layer in each epoch, 50% random-selected nodes were temporarily deactivated during training. This forced our model's learning to spread throughout most or all nodes of each layer.

1.5 Metrics of Success

Among many possible ways of measuring our model's success, we chose 3 primary ones: classification accuracy, receiver operating characteristic (*ROC*), and area under curve (*AUC*).

Classification accuracy refers to the total percentage of correct classifications in a given model evaluation.

While accuracy is a common metric for success in machine learning, this problem requires more advanced success metrics. This is because our data has about 20 survey participants without cornary heart disease for each afflicted person. Hence, a naive classifier could score 95 % accuracy and appear proficient by always guessing "healthy". *ROC* and *AUC* are industry-standard metrics for measuring precision and recall, which help solve this problem caused by data sparsity.

ROC plots the true positive rate (TPR) against the false positive rate (FPR) as given in Figure 5. Given a threshold parameter T and probability densities $f_1(x)$ for positive classes and $f_0(x)$ for negatives, TPR is defined as

$$TPR(T) = \int_T^\infty f_1(x)dx, \tag{1}$$

and FPR is defined as

$$FPR(T) = \int_T^\infty f_0(x)dx. \tag{2}$$

AUC refers to the area under the ROC curve, or the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It may be formally represented as

$$AUC = \int_{-\infty}^\infty TPR(T)(-FPR'(T))dT. \tag{3}$$

2 Results

Using the aforementioned hyperparameters on our cleaned data set, we were able to successfully train a coronary heart disease classifier.

Figure 4 shows accuracy gradually increasing with training time. Note that the increase here was not as steady as observed in PCA in Figure 1.

Our ROC curve after 15 epochs of training is displayed in Figure 5. Its integral covers a healthy, large area. This area increased with training; AUC change is shown in Figure 6. During early epochs it increased quickly, then converged around 0.84. One way to interpret the ROC curve is, if we want to identify 95% of heart disease cases, we can do so, but suffer a 65% false positive rate. This obviously is not the ideal classifier–however, it is a significant achievement above randomly always guessing positive, which would yield a roughly 95% false postive rate to achieve the same true positive rate.

We also include results for loss in Figure 7, which while not of the greatest importance, do help show the effects of normalization on the training process.

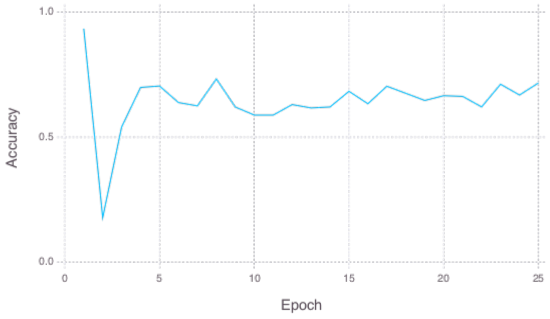

Figure 4: Accuracy over 25 epochs of DNN training.

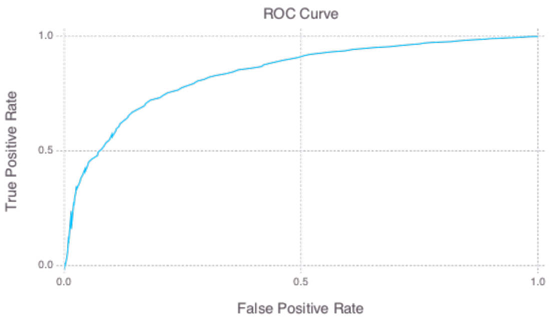

Figure 5: ROC curve. Shows true positive rate with regards to false positive rate. Note that the integral of this curve indicates success in true positive classification, as shown in Figure 6.

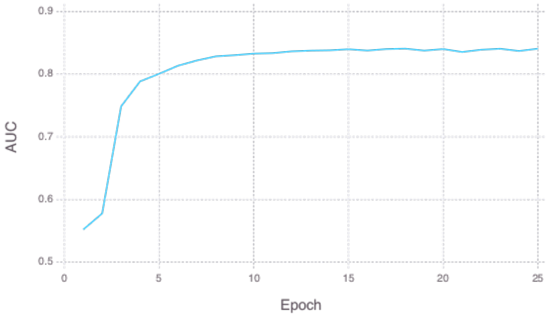

Figure 6: Area under curve over 25 epochs of DNN training. After balancing data, this offers the best indication of our success for our final classification model because of its emphasis on positives.

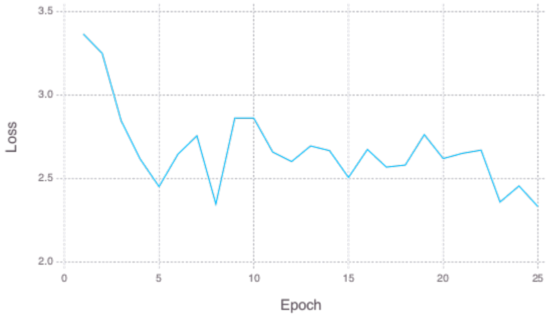

Figure 7: Loss over 25 epochs of DNN training.

3 Future Work

This paper represents an initial effort towards solving a complex problem. In the future, we hope to complete our processing and analysis of the NHANES data to ensure we are not leaving any features on the proverbial table. Similarly, we would like to expand this analysis to beyond just heart disease–instead, have it return a vector of probabilities for any number of possible health conditions (e.g., asthma, obesity, diabetes, cancer, stroke, bronchitis). This kind of tool could ensure doctors make accurate diagnoses for rare diseases or otherwise difficult-to-diagnose diseases, as the model finds subtle patterns in the data that humans might miss. We would also further refine this model, and explore other potential classifiers and feature selection techniques in greater depth.

If we were to repeat this project, we would also recalibrate our approach to the problem. We initially strived to use as much as the data we had available to us as was possible. However, reviewing and processing these data as thoroughly as we did meant we had less time to work on refining the model.

References

[Abadi *et al.*, 2016] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[Damico, 2013] Anthony Damico. National health and nutrition examination survey (nhanes), 2013.

[Folate...,] Folate (folic acid) serum test. https://www.walkinlab.com/folate-folicacid-serumtest.html. Accessed: 2017-12-13.

[Lee and Giraud-Carrier, 2013] Jun Won Lee and Christophe Giraud-Carrier. Results on mining nhanes data: A case study in evidence-based medicine. *Comput. Biol. Med.*, 43(5):493–503, June 2013.

[MEMBERS *et al.*, 2014] WRITING GROUP MEMBERS, Alan S Go, Dariush Mozaffarian, Véronique L Roger, Emelia J Benjamin, Jarett D Berry, Michael J Blaha, Shifan Dai, Earl S Ford, Caroline S Fox, et al. Heart disease and stroke statistics—2014 update: a report from the american heart association. *circulation*, 129(3):e28, 2014.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[Torpy *et al.*, 2009] Janet M Torpy, Alison E Burke, and Richard M Glass. Coronary heart disease risk factors. *Jama*, 302(21):2388–2388, 2009.

[Walgreens Introduces...,] Walgreens introduces daily testing for cholesterol, blood glucose and a1c at more than 1,400 stores in 33 states and washington, d.c. http://news.walgreens.com/press-releases/community-news/walgreens-introduces-daily-testing-for-cholesterol-blood-glucose-htm. Accessed: 2017-12-13.

[Wold *et al.*, 1987] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.