Comparing Association Rules and Deep Neural Networks on Medical Data

by
Ian Fund

A thesis submitted to the Department of Computer Science,

College of Natural Sciences and Mathematics

in partial fulfillment of the requirements for the degree of

Master of Science
in Computer Science

Chair of Committee: Carlos Ordonez

Committee Member: Ricardo Vilalta

Committee Member: Hulin Wu

University of Houston
December 2019

ABSTRACT

Deep neural networks are today's most popular tool for building predictive models across various different disciplines. A decade ago, the most popular predictive modeling technique was association rule mining. In this work, we carefully compare these two techniques in an effort to identify a more effective model with which to predict heart disease, a multi-prediction problem. Both techniques require significant knowledge, manual tuning, and experimentation to determine optimal parameters. Our goal was to build a predictive model that is at least as good as the best association rules across our entire data set. Promising results were obtained for some examples, while others still remain unclear. Making predictive models with medical data continues to be a challenging problem to solve that requires more attention from the scientific community.

TABLE OF CONTENTS

LIST OF TABLES

1 Introduction

Ten to fifteen years ago, association rules were the premiere data mining and modeling technique [3, 27, 28, 2]. With the continuous increase and availability of computing power, neural networks and deep learning have effectively become the focus of most modeling and research questions. This work aims to compare disease prediction using the historically popular method of association rules and the currently popular method of neural networks.

Association rules are a powerful data mining technique that is capable of identifying every pattern present in a set of data. While association rules are a useful tool, they also present several issues that have limited their use cases. First, the number of rules generated from a large set of data can be enormous, with many (if not most) of the rules having too little support to be applicable to the population. It is also likely that many of the rules that are identified are only capturing artifacts and aberrations in the data that do not accurately describe the target concept. Next, rules that appear meaningful may have very low support, due to the low sample population that can be accurately modeled using said rule. The ratio of said support for each rule is assigned a cutoff that is usually representative of the target population and concept that we want to model [28]. Lastly, only patterns are generated; while useful, patterns are not as applicable to a population as a predictive model would be.

Deep Neural networks (DNN) have taken over as the premier predictive modeling technique. In addition to this, data sets from many fields are currently being modeled by neural networks. While patterns generated from association rules may be useful, a model could be more easily applicable to the general population. Neural networks have also been shown to successfully resolve non-linear data sets, which is of particular interest to our work; medical data has been shown to most often be non-linear in nature. Being able to successfully model the entire population of interest would therefore provide a significant advantage over association rules.

This work aims to compare patterns discovered from association rules and predictive models generated by neural networks. The comparison will be made on predicting stenosis of the four major arteries in the heart; otherwise known as heart disease. Heart disease, medically known as cardiovascular disease, is the leading cause of death in the United States. In 2019, heart disease is predicted to occur in over one million people, of which 720,000 are new cases. [6]

2 Definitions

This section provides mathematical definitions for association rules and deep neural networks. Additionally, definitions for different types of activation functions and layers are provided.

2.1 Medical Data Set

Consider a medical data set containing n records $S = s_1, s_2, ...s_n$ with categorical and numerical attributes. Each attribute, A_i, is treated as categorical or numeric. For example, if S has attributes $A_1, A_2, ...A_p$, where p is the number of attributes, then A_i is either categorical or numeric [27].

Certain transformations on the data set are necessary for use in both association rule and neural network algorithms. For association rules, numeric attributes are binned at certain cutoff points. Each time a cutoff is applied, the data becomes a new bin. Categorical attributes are transformed by assigning an item to each categorical value. For neural networks numeric values can be used as is. Categorical values must be transformed to dummy variables. Dummy variables are binary attributes reporting whether or not the attribute is present [27].

Input for association rules is a file with n columns and m rows, where n is the number of attributes and m is the number of patient records. Output is a file with the number of rules discovered from the data. Each entry is an association rule, that consists of antecedent and consequent, support value, confidence value, and lift. These definitions are given below.

Input for the neural network is a file with n columns and m rows, where n is the number of attributes and m is the number of patient records. Output is the probability an entry belongs to a certain class. Additional output is the accuracy score for correctly predicting severe heart disease in that specific artery.

2.2 Association Rules

We will use a standard definition of association rules [1, 30, 31]. Each attribute is mapped from the raw data to a value based on the raw data. For example, in our case, ages 0-40 becomes item number 1. This is the input format for the association rule algorithm.

Let $\mathcal{T} = \{i_i, i_2, ..., i_m\}$, where i is an attribute in the data set. Let D be a set of n transactions such that $D = T_1, T_2, ..., T_n$ where $T_i \subseteq \mathcal{T}$. Let D be a set of n transactions such that $D = T_1, T_2, ..., T_n$, where $T_i \subseteq \mathcal{T}$ is a set of items, $\mathcal{T}$. A subset of $\mathcal{T}$ containing k items is called a k-itemset. Let X and Y be two itemsets such that $X \subset \mathcal{T}$, $Y \subset \mathcal{T}$ and $X \cap Y = \emptyset$. An association rule is an implication denoted by $X \Rightarrow Y$, where X is called the antecedent and Y is called the consequent.

The following is the definition of association rule metrics. Given an itemset X, support $s(X)$ is defined as the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. Consider the probability of $X, P(X)$ in D and $P(Y|X)$ the conditional probability of appearance of Y given X. $P(X)$ can be estimated as $P(X) = s(X)$. The support of a rule $X \Rightarrow Y$ is defined as $s(X \Rightarrow Y) = s(X \Rightarrow Y)/s(X)$. Confidence can be used to estimate $P(Y|X)$:P(Y| X) = $P(X$ Y) / (P(X)P(Y)) = c(X $\Rightarrow Y$). A probabilistic interpretation of support and confidence is discussed in [19]. A third metric, called lift, [4] is defined as $l(X \Rightarrow Y) = P(XY)/(P(X)P(Y)) = c(X \Rightarrow Y)/s(Y)$. Lift values greater than 1 provide strong evidence that X and Y depend on each other. Lower lift values indicate X depends on the absence of Y or vice versa. A lift value close to 1 indicates X and Y are independent.

A mining association rule is defined as finding the set of all rules $(X \Rightarrow Y)$ such that $s(X \Rightarrow Y \geq \psi$ and $c(X \Rightarrow Y) \geq \alpha$, given a support threshold ψ and a confidence threshold α. A k-itemset X such that $s(X) \geq \psi$ is called frequent. [27]

2.3 Deep Neural Network Definition

Let $\mathcal{T} = \{x_1, x_2, ..., x_m\}$, where x is an attribute in the data set. Data must be transformed from its raw state before it can be used by the DNN. For numeric values, the data can be fed to the DNN as is. If x_m is a discrete variable, its value must be transformed to a dummy variable based on the number of possible values. In our case, discrete variables were all binary and categorized as 1 or 0. Additionally, data is scaled before input into the DNN. Z-score was use to make this transformation. Output for DNN is based on the class being predicted. The goal was to predict the probability of a person having severe stenosis in the target artery. Since there are four target arteries, a DNN will be created for each artery.

A learning paradigm amenable to testing the feasibility of knowledge transfer is that of neural networks. A neural network is capable of expressing flexible decision boundaries over the input space [17]; it is a nonlinear statistical model that applies to both regression and classification. In particular, for a neural network with one hidden layer, each output node computes the following function:

$$g_k(X = \mathbf{x}) = \mathbf{f}\left(\sum_{\mathbf{l}} \mathbf{w_{kl}}\ \mathbf{f}\left(\sum_{\mathbf{i}} \mathbf{w_{li}} \mathbf{x_i} + \mathbf{w_{l0}}\right) + \mathbf{w_{k0}}\right)$$

where $\mathbf{x}$ is the input feature vector, $f(\cdot)$ is a nonlinear (e.g., tanh, ReLU) function, and x_i is a component of vector $\mathbf{x}$. Index i runs along the components of vector $\mathbf{x}$, index l runs along the number of intermediate functions (i.e., nonlinear transformations of the input features), and index k refers to the kth output node. The output is a nonlinear transformation of the intermediate functions. The learning process is limited to

finding appropriate values for all weights $\{w\}$. The concepts described below are equally valid for *deep neural networks* [17] where there is more than just one hidden layer between the input and output nodes.

Several activation functions are popularly used in neural networks. The sigmoid function is foundational to logistic regression and became an activation function that is considered when using neural networks. It ranges from 0 to 1, and [22] defines

$$g(x) = \frac{1}{1 + e^{-x}}$$

The hyperbolic tangent function (tanh) is defined as the ratio of hyperbolic sine and hyperbolic cosine. Its values range from -1 to 1 [22]

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$$

The Rectified linear unit function (ReLU) is arguably the most popular activation function used in neural networks today: [26]

$$ReLU = \max(0, x)$$

Leaky ReLU is a slightly modified version of ReLU that has a slight slope when $x < 0$ In this example, k is a small value to used reduce the slope [22].

$$h(x) = \begin{cases} x, & \text{if } x > 0 \\ kx, & \text{otherwise} \end{cases}$$

Dropout is a technique used to prevent over fitting and improve learning performance. It works by temporarily dropping a random neuron from the network after every epoch during the training process. The probability can be manually tuned, but its default value, 0.5, is close to the optimal solution in many instances [41].

3 Contrasting Association Rules and Deep Neural Networks

This section provides more details on the limitations of association rules and the important parameters for neural networks.

3.1 Limitations of Association Rules

Association rule generation has the advantage of being exhaustive in the way of rule discovery. Given that the data set for this experiment was small, association rules have a distinct advantage over neural networks. Association rules will find every pattern given a data set. On the other hand, neural networks work best with enormous data sets. However, there are several disadvantages which has lead to the decline in popularity for association rules.

- running time
- support
- confidence
- antecedent/consequent constraint

These are perhaps the most limiting components of association rule discovery. With no constraints applied, the run time of this algorithm drastically increases, hours or days depending on the size of the data set. Support is the parameter that most greatly affects run time. With support low enough, one would be discovering rules that apply to so few people the information gained would not generalize to a population. In addition, run time greatly increases. Confidence has a different effect. As one changes confidence values, the number of patterns is changed. If one has a lower minimum confidence value then the number of rules discovered dramatically increases. The concern in this case is that rules discovered are less meaningful if one is less confident in how they apply to the population. In our case, limiting the items that appear in the antecedent and consequent is of the utmost importance. We only care about rules where severe artery stenosis is the consequent. Conversely, artery items should not appear in the antecedent. However, there cannot be too few items in the antecedent because then one sacrifices predictive power and medical usefulness. In a perfect world, we would discover rules that have 50% support and 90% confidence. Unfortunately, those rules are difficult to come by. Support is the biggest limitation because using it to discover rules is a balancing act: if support is too low we discover too many rules, most of which will not generalize to the population at large. On the contrary, if we set support too high we won't discover any rules at all. Confidence is the next limitation. As previously mentioned, we are after rules with high confidence. Based on the opinion of domain experts, the minimum confidence was set at 70%. This number is based on balancing identifying sick and health patients. We increased the size of the antecedent from four items, in previous experiments, to six items for this research. It is possible that six items is too many and would be too specific for medical usefulness. On the contrary, rules with more items would have more predictive power due to their specificity.

Support minimum: $\psi = 0.02$. Confidence minimum: $\alpha = 0.70$. Lift minimum = 1.

- excessive number of rules generated
- constraints necessary
- designed to overfit
- no predictive model generated
- similar patterns are disconnected from one another
- minimum support and confidence necessary, but difficult to define
- negation must be considered
- item groups can be used to reduce complexity

The use of additional constraints is necessary in order to reduce the number of rules generated and to make run time reasonable. Due to the fact that we have more computing power now than we did ten years ago, fewer constraints were implemented. Negation was not considered for this project because we are more interested in what makes people sick as opposed to what makes them healthy. Additionally, negation was unnecessary because we discarded patients with boarder-line disease which condensed output to only one item for each artery. Grouping was also not necessary for this project. Another concern with association rules is overlap of rules. Similar items appear in many rules, but have different meaning and interpretation when considered as an overall rule.

3.2 Parameters of Neural Network

When utilizing DNNs to model data, it is important to realize the impact that the various hyperparameters can have on the performance of the machine. The number of layers and number of neurons per hidden layer are among the most important hyperparameters in any classification or regression problem [13, 25]. Although the performance of a machine is somewhat predictable based on these hyperparameters, it is important to realize that each separate data set is likely to have unique optimal hyperparameters. This makes hyper parameter optimization an important topic in machine learning, with various optimization techniques being implemented by the scientific community [34, 20, 7]. In order to optimize neural network performance for comparison to association rules, we elected to use various hyperparameters to address this topic. Secondary parameters include the learning rate and activation function. While these are important, they seem to have less impact on performance in our case.

Overfitting is an issue in machine learning where the model is over trained on the training data to the point where accuracy on the training data is 100%. Once this occurs, the model does not perform as well on the testing data because it has become far too specific. The goal is to learn as much and as long as possible before a model overfits and then scale back to a point where the testing accuracy does not suffer at the expense of the training accuracy.

3.3 Examples

3.3.1 Association Rule Example

A simple example of an association rule is be presented below. Consider an example data set with five records and four attributes, that is illustrated in Table 1. If the association rule algorithm were to be applied to the data set, one of the rules generated would be:

$$\{200 <= CHOL < 250, SMOKE = Y\} \Rightarrow \{70 <= LAD < 100\}$$

AGE	CHOL	SMOKE	LAD
65	253	Y	72
42	258	Y	71
56	186	N	51
47	251	Y	46
51	132	N	36

Table 1: Sample Data for Association Rule Example

In the above rule, support $\psi = 0.4$, confidence $\alpha = 0.67$, and lift = 1.86. This example is a rule that describes the following: A patient that is between 60 and 100 years old, has a certain region of the heart that is considered to have defect, has cholesterol between 200 and 250, and has an LAD artery that is blocked between 70% and 100%. This particular antecedent (left side) appeared in 40% of the population. Of that 40%, 67% had the artery blockage. If someone has the attributes in the antecedent, that person is 1.1 times more likely to have the item in the consequent present as well. Support is calculated as $\psi = \frac{2}{5}$, where the numerator is the number of items that have both antecedent and consequent and the denominator is the total sample. Confidence is calculated as $\alpha = \frac{2}{3}$, where the numerator is the number of records that satisfies both antecedent and consequent and the denominator is the number that only satisfies the antecedent. Lift is the confidence divided by the fraction of items containing the consequent, $\frac{\frac{2}{3}}{0.6*0.6}$.

3.3.2 Deep Neural Network Example

An example of the DNN is made using the same mock data set as above, illustrated in Table 2. The output for DNN is the probability that a certain input belongs to an output class. Here we create probabilities that would correspond to $LAD >= 70\%$. According to the DNN in this example, if one is 65 years old, has cholesterol of 253, and smoke, the probability of having greater than 70% stenosis of LAD is 62%. It is worth noting that these probabilities are independent of overall model accuracy. Chance is defined as the ratio of the largest class in the classification problem. In this case, that is $\frac{2}{5}$ or 40%. The goal of the model is to have higher than chance predictive power. For this example, we want the model to be predicting more than 40% accuracy to show that learning has occurred.

AGE	CHOL	SMOKE	LAD	DNN Output
65	253	Y	72	0.62
68	258	Y	71	0.71
56	186	N	51	0.38
55	251	Y	46	0.47
58	132	N	36	0.15

Table 2: Sample Data and Output for DNN

4 Experiments

This section discusses the processes used for experimentation. This includes hardware and software information as well as specific parameters for each technique used in our experiments. Results and comparisons are discussed as well.

4.1 Hardware and Software

Experiments were conducted on a Ubuntu 18.04 machine with a quad core Intel(R) Pentium(R) CPU @ 1.60GHz and 8GB of DDR3 1600 MHz memory.

Association rules were generated from an updated C++ program, approximately 5000 lines of C++ code.

Approximately 500 lines of python code were used to create the DNNs used in this project. The major python libraries used were: keras, tensorflow, scikit-learn, pandas, and numpy.

4.2 Medical Data Set Description

The data set used to validate the two approaches was derived from 655 cardiovascular disease patients. Each patient had 25 attributes associated with his data that are considered relevant to his condition by the

collaborating physician. Four additional parameters were created by the researchers using data from the main set (more details below). Parameters in the set include risk factors (age, cholesterol, sex, hypertension, diabetes, hyperloipodemia, and smoking habits), historical information (family history of heart disease, claudication or pain caused by reduced blood flow, previous angina or chest pain caused by reduced blood flow to the heart, previous stroke, and previous cardiac surgery), heart region images, and carotid artery stenosis. Four additional parameters were created for experimentation. Severe disease of each artery was created based on the stenosis present (50 +% for LM, 70 +% for all other arteries). The continuous variables used were: age, LM, LAD, LCX, RCA, AL, AS, SA, SI, IS, IL, LI, LA, AP, and cholesterol. The binary variables used were: sex, hypertension, diabetes, hyperloipodemia, family history of heart disease, smoker, claudication, previous angina, previous stroke, and previous heart surgery. Originally, the binary variables were "m" or "f" for sex and "y" or "n" for all the others. Slight modifications were necessary for the neural network input. The "sex" attribute was transformed to the male attribute, 1 for male and 0 for female. Remaining discrete attributes were transformed from 1 for yes and 0 for no. As previously mentioned, dummy variables are necessary for the neural network to work with. This is why these transformations were performed. Additionally, we created binary variables for the neural network to train and test on. These variables were based on the discrete values of the arteries (LM, LAD, LCX, RCA). Each artery became its own attribute for severe disease, defined as 70+% for LAD, LCX, and RCA and 50% for LM, with 1 if severe disease is present and 0 otherwise. Descriptions of all attributes used is given in Table 3.

Abbreviation	**Definition**	**Abbreviation**	**Definition**	**Abbreviation**	**Definition**	**Abbreviation**	**Definition**
Age	Patient Age	**SA**	Septo-Anterior	**HTA**	Hypertension	**PSTROKE**	Previous Stroke
LM	Left Main	**SI**	Septo-Inferior	**DIAB**	Diabetes	**PCARSUR**	Previous heart surgery
LAD	Left Anterior Descending	**IS**	Infero-Septal	**HYPLPD**	Hyperlipidemia	**CHOL**	Cholesterol
LCX	Left Circumflex	**IL**	Infero-Lateral	**FHCAD**	Family History of Heart Disease		
RCA	Right Coronary	**LI**	Latero-Inferior	**SMOKE**	Smokes		
AL	Antero-Lateral	**AP**	Apical	**CLAUDI**	Claudication		
AS	Antero-Septal	**Sex**	Sex	**PANGIO**	Previous angina		

Table 3: Attribute Definitions

Two additional changes were made to the patient data: missing data were filled in and heart region images were modified. Missing patient data were substituted in one of two ways. For binary variables (smoking, previous cardiac surgery, etc.) the mode was taken for each sex and then applied to the missing patient's data. For continuous variables (cholesterol, age, or heart region images) the average was taken for each sex and then applied to the missing patient's data. Originally, the data ranged from -1 to 1 for heart region images. Regions with no defect were labeled from -1 to 0.2. However, no ranges between -1 and 0 were ever used. To prevent the neural network from assigning lower weights to the images because of the large difference in value, the -1 images were replaced with 0. All data was then scaled using standard normalization to properly distribute the potential predictive power of each attribute.

4.3 Parameters

4.3.1 Association Rule Parameter Settings and Constraints

The first step was to divide the data into separate ranges for healthy and unhealthy. The four major arteries (LM, LAD, LCX, and RCA) were divided into two ranges, no disease or severe disease. Typically, arteries are considered healthy if stenosis is under 50%, moderate disease if stenosis is between 50% and 70%, and severe if greater than 70%. The exception is LM, which is healthy under 30%, moderate between 30% and 50%, and severe above 50%. These cutoff points come from popular cardiology practices. For these experiments, only severe disease cutoffs were considered.

The nine heart region images (AL, IL, IS, AS, SI, SA, LI, LA, AP) were divided into two categories. Healthy regions with no defect ranged from 0 to 0.2. Regions with defects were grouped as greater than or equal to 0.2. Cholesterol was cutoff at three different values. Between 0 and 200 was considered healthy, between 200 and 250 was considered warning, and over 250 was considered high. We grouped age into three different ranges: 0-40 years old, 40-60 years old, and 60-100 years old.

In an effort to reduce the number of association rules generated the following constraints were applied. The maximum size of the antecedent was set to six items.The heart arteries were restricted to only appearing in the consequent of the rule. All other attributes were set to only appear in the antecedent. Rules that applied to approximately fewer than 33 patients in the data set(minimum support, $\psi = 2\%$) would not be considered. Rules with high support are preferred due to the higher possibility of generalizing the rules to the population. Based on expert opinion, confidence, α, was set to 70%. Medically speaking, rules under 70% confidence are not useful in application [31, 32, 31, 29] . Minimum lift was set to 1.0. Maximum run time for association rules was set to thirty minutes.

4.3.2 Association Rules Results

With the aforementioned constraints applied, we were left with 2,634 association rules. Of these rules, 2,549 were for LAD, 3 for LCX, and 82 for RCA. Values for minimum, maximum, mean, and standard deviation are located in Tables 5, 6, and 7. Of note, the maximum support for a rule was only 13% of the population of the data set. Some of the most statistically interesting rules are included in the Table 4. At first glance, the distribution of rules stands out. A vast majority of the rules are for LAD, 96.77% of all rules. No rules for LM were discovered. This is likely due to a small number of patients in the data set with severe disease of LM. LCX had only three rules; all of which were only at a support level of 2%. Finally, RCA had 82 rules discovered.

Interestingly, several rules had a confidence value of 100% with 2-3% support. This means that these rules applied to each patient who had the items in the antecedent present. Although support is low, ranging from approximately 13 and 20 people, it is still significant to find a rule that applies to each affected person. Another interesting finding is that each rule had at least one of the nine heart region images in the antecedent. Previous research had grouped all images together in an effort to reduce complexity and run time. Now that we have included each image as its own attribute we can get a more accurate picture of which regions of the heart are impacting heart disease. Many rules had several heart images, but fewer had several of our other attributes. The non-image attributes help contribute to a clearer picture of who is likely at risk. For example, one rule with several non-image attributes was age between 60 and 100, defective SI region of the heart, male, hyperloipodemia, and smokers indicates RCA between 70 and 100, with 2% support and 88% confidence.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**
{0.2<=SA<1.1,0.2<=LI<1.1,200<=CHOL<250}	{70<=LAD<100}	0.03	1	3.2
{0.2<=AS<1.1,0.2<=IL<1.1,0.2<=AP<1.1,SEX=F}	{70<=LAD<100}	0.03	1	3.2
{0.2<=SA<1.1,0.2<=LI<1.1,SEX=M,200<=CHOL<250}	{70<=LAD<100}	0.02	1	3.2
{0.2<=AS<1.1,0.2<=IL<1.1,0.2<=AP<1.1,SEX=F,PSTROKE=y}	{70<=LAD<100}	0.02	1	3.2
{0.2<=AS<1.1,0.2<=AP<1.1}	{70<=LAD<100}	0.13	0.72	2.3
{60<=AGE<100,0.2<=AS<1.1}	{70<=LAD<100}	0.11	0.7	2.2
{0.2<=IL<1.1,0.2<=LI<1.1,HYPLPD=y,200<=CHOL<250}	{70<=LCX<100}	0.02	0.78	2.9
{40<=AGE<60,0.2<=IL<1.1,0.2<=LI<1.1,DIAB=y}	{70<=LCX<100}	0.02	0.72	2.7
{60<=AGE<100,0.2<=IL<1.1,0.2<=LI<1.1,HYPLPD=y}	{70<=RCA<100}	0.08	0.7	2.2
{60<=AGE<100,0.2<=IL<1.1,DIAB=y,HYPLPD=y}	{70<=RCA<100}	0.02	0.89	2.8
{60<=AGE<100,0.2<=SI<1.1,SEX=M,HYPLPD=y,SMOKE=y}	{70<=RCA<100}	0.02	0.88	2.8

Table 4: Interesting Rules

	Whole Set	**LAD**	**LCX**	**RCA**
Min	0.02	0.02	0.02	0.02
Max	0.13	0.13	0.02	0.08
Mean	0.0388	0.0392	0.0200	0.0284
Std	0.0170	0.0170	0	0.0130

Table 5: Support Values

	Whole Set	**LAD**	**LCX**	**RCA**
Min	0.70	0.70	0.72	0.70
Max	1.00	1.00	0.78	0.89
Mean	0.7779	0.7790	0.7400	0.7417
Std	0.0558	0.0557	0.0346	0.0444

Table 6: Confidence Values

	Whole Set	LAD	LCX	RCA
Min	2.2	2.2	2.7	2.2
Max	3.2	3.2	2.9	2.8
Mean	2.4866	2.4899	2.7667	2.3732
Std	0.1820	0.1818	0.1155	0.1441

Table 7: Lift Values

Of the 2,634 rules, 1,851 of them are between 2% and 4% support. This means roughly 70% of rules discovered apply to approximately 13 and 26 people. With support as low as this, it is difficult to apply these findings to the population at large. Another observation on support is its relationship with confidence. As support goes up confidence goes down. Therefore, rules that have very high confidence were only found on a small percentage of the data set.

4.3.3 Neural Network Parameter Settings and Constraints.

Many iterations of experimentation occurred in an effort to find the model that had the highest predictive capabilities. In order to accomplish this, hyperparameters for the neural network were changed for each experimental iteration (results presented in the chart below). Performance is most affected by the number of hidden layers of the neural network or the number of neurons in each respective layer. The types of activation functions were varied as well. Popular activation functions that were tested were: rectified linear units (ReLU), hyperbolic tangent (tanh), and Leaky ReLU, defined in Section 2.3. We also attempted to use Bayesian optimization [20] to select which hyperparameters perform the best. Maximum run time for neural networks was set to thirty minutes. The main constraint from the data set applied to neural networks is that the four arteries are not included in the data set, and the transformed binary artery disease values are the target variables.

4.3.4 Neural Network Results

As a sanity check, we applied some classic machine learning algorithms on the data to ensure better-than-chance predictions, and to ensure that pursuing the models with deep neural networks was the correct choice. Results are presented in Table 8. Logistic regression and support vector machines were applied to the data (transformed from raw data for use by the neural network). Results with prediction accuracies are included in the chart below, as well as an example from a deep neural network. Default parameters were used for both logistic regression and support vector machines.

	Logistic	**SVM**	**DNN**
LAD	71.8	74.8	77.9
LCX	77.1	77.9	79.4
RCA	67.2	67.2	71
LM	93.1	94.7	93.1

Table 8: Machine Learning Model Comparisons

Next, we conducted experiments with various neural network layouts. Results are presented in Table 9. Each item in the square brackets represents a layer of the network. The number is how many neurons are in that layer. The type of activation function is next to that.

	LAD	**LCX**	**RCA**	**LM**
Description of Network				
[50,50,50,50] LeakyReLU	78.8	75.6	71.4	92.2
[5,5,5,5] ReLU	75.6	72.8	69.1	92.2
[50,50,50,50,50] ReLU	77.9	77.4	71.4	92.2
[20,20,20,20] ReLU	75.1	77.4	72.8	92.2
[50,50,50,50,5] ReLU	78.3	75.1	71.9	92.2
[50,50,50,50] tanh	77	77.4	71.4	92.2
[50,50,50,50] linear	76	76.5	71.9	92.2
[50,50,50,50] LeakyReLU	78.8	77.9	67.9	93.1
[50,50,50,50] LeakyReLU, dropout 0.5	77.9	79.4	71	93.1

Table 9: Deep Neural Network Accuracies

The best results were obtained from a network with four layers of fifty neurons per layer, an activation function of LeakyReLU and dropout of 0.5. It is worth noting that the artery LM had higher prediction values than the other arteries, as expected. From a medical standpoint, LM is the base from which the other arteries branch out of. Because of this, severe disease in LM is rarer than the other three arteries. Due to its rarity, our data set is lacking in positive examples. The high predictive values from the neural network is due to the chance of predicting no disease in LM is particularly high. We attempted to apply Bayesian optimization to search for the best performing hyperparameters. Bayesian optimization was applied on two different networks as a trail with fifty iterations. Results showed lower accuracy than deep neural networks. In addition, the run time was approximately between forty and seventy-five minutes. Fifty iterations is low for Bayesian Optimization; we would have preferred over one thousand iterations. However, with run time so long at low iterations, we decided not to continue with Bayesian optimization.

4.3.5 Association Rule and Deep Neural Network Comparison

In comparing these two methods, it is difficult to say that either is clearly superior. Some instances show association rules to have higher accuracy (confidence) than neural networks. However, rules with higher

confidence have lower support. This relationship makes it difficult to generalize rules to the population at large. This relationship between support and confidence is one of the main issues with association rules: as one goes up the other goes down. One of the rules for LCX had comparable confidence, but had the minimum support value of 2%. From a medical perspective, discovering rules and having high predictability of LCX was most surprising. Medically, it has been difficult to predict. Expanding on the medical standpoint, RCA and LAD were the arteries we expected to learn the most about. However, RCA had the lowest performance in the neural network: 71%. The mean confidence for rules containing RCA was 74.17%. The max confidence was 89%, but again has the minimum support of 2%. In this case, it is less clear which model is better overall. However, neural networks do have the advantage of being a widely applicable predictive model. Run times for each method are described in Table 10.

LAD	**LCX**	**RCA**	**LM**	**AR**
6m 59.88s	7m 6.45s	7m 10.11s	7m 5.68s	4m 11.78s

Table 10: Run Times in Minutes and Seconds

5 Related Work

This section reviews previous works on association rules and deep neural networks in the medical field.

Data mining on medical data presents unique challenges [38]. Potential issues include fragmented data collection, stricter privacy concerns, rich attribute types (image, numeric, categorical, missing information), complex hierarchies behind attributes and an already rich and complex knowledge base. Research that discusses how computer programs can be used to diagnose heart disease are explored in [16, 24, 23]. Association rules have been used to help infection detection and monitoring [8, 9], to understand what drugs are co-prescribed with antacids [10], to discover frequent patterns in gene data [5, 11], to understand interaction between proteins [33] and to detect common risk factors in pediatric diseases [14]. Association rules have also been extended with fuzzy sets and alternative metrics to improve their application [12]. Association rules were proposed in the seminal paper [1]. Mining association rules combining numeric and categorical attributes are studied in [37]. [18] optimizes an algorithm that incorporates constraints into the association rule mining process. In [40], algorithms are proposed to include constraints that exclude or include certain items in the association rule.

Popular activation functions for neural networks, as well as their performance are discussed in [22, 26]. The basic structure of neural networks and deep neural networks is discussed in [17]. Neural networks are successfully applied to medical data in [15, 39]. Deep neural networks are applied to the problem of heart

disease in [36, 42, 35]. Association rules and neural networks are used in tandem on a medical data set in [21].

6 Conclusion

In this section we review experimental results, discuss issues with the research, and provide some options for future research.

In this work, we compared the predictive powers of association rules and neural networks. Results indicated a distinct advantage in some cases and less clear results in others. Overall, deep neural networks are preferred due to consistently high accuracy scores and a widely applicable model, whereas the association rules cannot keep up due to only finding small patterns within a data set. For our experiments, the arteries LAD and LCX had much higher predictive accuracy in the neural network than with the association rules. Further, only three association rules were found with LCX. The deep neural network produced a model able to predict LCX with 79.4% accuracy, which is clearly preferable. RCA had a peak accuracy of 71%. While not as high as the other results, it still has the benefit of being a predictive model.

Higher accuracy was unable to be obtained for a multitude of reasons. Perhaps the largest factor is the data set. Big data sets have the luxury of being able to drop incomplete samples and have a large training and testing set. Medical data sets have a much harder time with this. Medical data is harder to come by and is therefore usually smaller in size. Another issue with data sets is that neural networks perform much better on a balanced data set (data sets that have an even distribution based on target classes). In our case, it should be 50/50.

The biggest consideration for future research would be the data set. Using a large, well balanced data set is preferred for this sort of work. Another consideration would be utilizing Bayesian optimization with a large number of iterations.

Bibliography

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216, 1993.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.

[3] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[4] R. Bayardo and R. Agrawal. Mining the most interesting rules. In *ACM KDD Conference*, pages 145–154, 1999.

[5] C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut, and O. Gandrillon. Strong association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genom Biol.*, 3(12), 2002.

[6] E. J. Benjamin, P. Muntner, and M. Sommer Bittencourt. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation*, 139(10):e56–e528, 2019.

[7] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[8] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser. Association rules and data mining in hospital infection control and public health surveillance. *J. Am. Med. Inform. Assoc. (JAMIA)*, 5(4):373–381, 1998.

[9] S. E. Brossette, A. P. Sprague, W. T. Jones, and S. A. Moser. A data mining system for infection control surveillance. *Methods Inf Med.*, 39(4):303–310, 2000.

[10] T. J. Chen, L. F. Chou, and S. J. Hwang. Application of a data mining technique to analyze coprescription patterns for antacids in Taiwan. *Clin Ther*, 25(9):2453–2463, 2003.

[11] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[12] M. Delgado, D. Sanchez, M. J. Martin-Bautista, and M. A. Vila. Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine*, 21(1-3):241–5, 2001.

[13] T. Domhan, J. T. Springenberg, and F. Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[14] S. M. Down and M. Y. Wallace. Mining association rules from a pediatric primary care decision support system. In *Proc of AMIA Symp.*, pages 200–204, 2000.

[15] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[16] H. S. Fraser, W. J. Long, and S. Naimi. Evaluation of a cardiac diagnostic program in a typical clinical setting. *J. Am. Med. Inform. Assoc. (JAMIA)*, 10(4):373–381, 2003.

[17] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[18] J. L. Han. Pushing constraints in templates for mining association rules. In *Florida AI Research Symp*, pages 375–379, 1996.

[19] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, New York, 1st edition, 2001.

[20] K. Kandasamy, W. Neiswanger, J. Schneider, B. Poczos, and E. P. Xing. Neural architecture search with bayesian optimisation and optimal transport. In *Advances in Neural Information Processing Systems*, pages 2016–2025, 2018.

[21] M Karabatak and M. C. Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2):3465–3469, 2009.

[22] B. Karlik and A. V. Olgac. Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122, 2011.

[23] W. J. Long. Medical reasoning using a probabilistic network. *Applied Artifical Intelligence*, 3:367–383, 1989.

[24] W. J. Long, H. S. Fraser, and S. Naimi. Reasoning requirements for diagnosis of heart disease. *Artificial Intelligence in Medicine*, 10(1):5–24, 1997.

[25] I. Loshchilov and F. Hutter. CMA-ES for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*, 2016.

[26] V. Nair and G. 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.

[27] C. Ordonez. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine (TITB)*, 10(2):334–343, 2006.

[28] C. Ordonez. Comparing association rules and decision trees for disease prediction. In *Proc. ACM HIKM Workshop*, pages 17–24, 2006.

[29] C. Ordonez, N. Ezquerra, and C. A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems (KAIS)*, 9(3):259–283, 2006.

[30] C. Ordonez and E. Omiecinski. Discovering association rules based on image content. In *IEEE Advances in Digital Libraries Conference (ADL'99)*, pages 38–49, 1999.

[31] C. Ordonez, E. Omiecinski, L. de Braal, C. Santana, and N. Ezquerra. Mining constrained association rules to predict heart disease. In *IEEE ICDM Conference*, pages 433–440, 2001.

[32] C. Ordonez, C.A. Santana, and L. Braal. Discovering interesting association rules in medical data. In *Proc. ACM SIGMOD Data Mining and Knowledge Discovery Workshop*, pages 78–85, 2000.

[33] T. Oyama, K. Kitano, T. Satou, and T. Ito. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18(5):705–714, 2002.

[34] F. J. Pontes, G. F. Amorim, P. P Balestrassi, A. P. Paiva, and J. R. Ferreira. Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing*, 186:22–34, 2016.

[35] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V McConnell, G. S. Corrado, L. Peng, and D. R. Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

[36] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[37] R. Rastogi and K. Shim. Mining optimized association rules with categorical and numeric attributes. In *Proc. IEEE ICDE Conference*, pages 503–512, 1998.

[38] J. F. Roddick, P. Fule, and W. J. Graco. Exploratory medical knowledge discovery: Experiences and issues. *SIGKDD Explorations*, 5(1):94–99, 2003.

[39] S. Sarraf and G. Tofighi. Classification of Alzheimer's disease using FMRI data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*, 2016.

[40] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proc. ACM KDD Conference*, pages 67–73, 1997.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[42] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 95–102. Springer International Publishing, 2016.