Comparing Association Rules and Deep Neural Networks for Heart Disease Prediction

Carlos Ordonez
Department of Computer Science
University of Houston
USA

Ian Fund
Department of Computer Science
University of Houston
USA

Ladjel Bellatreche
LIAS/ISAE
ENSMA
France

***Abstract*—Two decades ago, the most popular data mining technique were association rules (ARs). Nowadays deep neural networks (DNNs) are the most popular mechanism for building predictive models. On the other hand, medical data sets, despite being generally small in size (low volume), they are challenging for predictive models due to diverse attribute content (high variety) and variables with low redundancy (high variability). In this work we compare these two analytic techniques to identify effective models to predict heart disease, a multi-target prediction problem. Both techniques require expertise, manual tuning, and iterative experimentation to determine optimal parameters. Our goal is to build a DNN model that is at least as good as the best ARs. There exist two Big Data challenges: risks factors combined with imaging attributes produce a large number of hidden patterns and the number of association rules reaches millions, without using search constraints at low support (frequency) values. Preliminary experiments on a real data set show discovered rules have high predictive accuracy and they provide a highly accurate, but highly specific, profile of sick patients. Despite careful data pre-processing and hyper-parameter tuning DNNs are slightly more accurate than association rules, but more generalizable. Therefore, both techniques can complement each other.**

***Index Terms*—Pattern, Deep learning, Frequent Itemset, Data mining, Classification**

I. Introduction

Association rules are a powerful data mining technique that is capable of discovering every frequent pattern present in a set of data, meeting minimum probability constraints. From a machine learning perspective, association rules identify simple, intuitive, specific predictive patterns. About 20 years ago, association rules (ARs) [12] were the premiere data mining technique, with thousands of papers being published since then. But trends have changed dramatically. With the continuous rise and availability of multicore CPUs, GPUs and larger RAM, deep neural networks (deep learning) [1], [8], [15], [25] have effectively become the main predictive model for machine learning and data science nowadays. Deep Neural Networks (DNNs) have taken over as the premier predictive modeling technique due to their high accuracy [1], [11], surpassing most previous machine learning models. Data sets from diverse fields, including medicine, are currently being analyzed by deep neural networks. In addition, their ability to cope with noise, high dimensionality and raw data have made them the preferred technique to analyze text and images. On the other hand, medical data sets (from a hospital in our case), despite being generally small in size, they are challenging for predictive models due to their high variability (many hidden patterns), their diverse attribute content (high variety), and having variables with low redundancy (high variability). In other words, they do represent a Big Data problem, despite lacking one V: Volume.

Based on the motivation above, this work compares the historically popular technique of association rules and the currently most popular predictive method of deep neural networks for heart disease prediction. Our goal is predicting stenosis (plaque buildup) of the four major arteries in the heart. This work is not the first to study ARs or DNNs for heart disease prediction, but previous work has not compared both techniques with each other. Previus research has neither considered heart disease as a multi-target prediction problem.

We study the problem from a comprehensive analytic perspective. We explain how the input data set is pre-processed for each technique. We compare the strengths and weaknesses of each tecnique for our medical problem. Finally, we compare both techniques on a real medical data set, tuning their parameters to increase accuracy and reduce processing time (without compromising accuracy). To round up our study, we also compare DNNs against other supervised machine learning models, widely used today: Support Vector Machines and Logistic Regression. We should emphasize logistic regression remains a workhorse in medicine given its explainability and robustness, despite being slightly less accurate than SVMs or DNNs.

II. Background and Definitions

This section provides mathematical definitions for association rules and deep neural networks that will be used throughout the paper. Each subsection can be skipped by a reader familiar with each technique.

A. Input Data Set

Consider a raw data set S containing n records $S = \{x_1, x_2, .., x_n\}$ with categorical, numerical and image attributes, to be transformed into the data set D defined below. More precisely, if S has p attributes $A_1, A_2, ...A_p$, then A_j is either categorical or numeric.

The input for association rules is a data set D with n transactions, coming from a discretization of S (explained

below). The output is a set of discovered rules meeting several thresholds. Each association rule consists of an antecedent and a consequent and has three associated metrics: support, confidence, and lift, defined below.

In contrast, for DNNs the input is a matrix. The input for the neural network is a matrix with d columns and n rows, where d is the so-called dimensionality and n is the number of patient records. The output is a probability of disease per artery per patient, as well as the accuracy score for correctly predicting severe heart disease in that specific artery.

Transformations on the raw data set are necessary to use it in both association rule and deep neural network algorithms. In association rules numeric attributes are binned at medically-recommended chosen cutoff points. Categorical attributes are transformed by assigning an item to each categorical value. For neural networks numeric values of predictive attributes can be used as is. On the other hand, categorical values must be transformed to dummy binary variables, reporting whether or not each attribute value is present. For both techniques target variables (arteries) must be transformed into binary variables to distinguish healthy and sick patients.

B. Association Rules (ARs)

The standard definition of association rules [12] is as follows. Let $\mathcal{I} = \{i_1, i_2, ..., i_m\}$, where i_j is an item (intuitively the subscript of a binray dimension), coming from a transformation on an attribute of S. Let $D = \{T_1, T_2, ..., T_n\}$ be a set of n transactions, coming from discretizing S, where $T_i \subseteq \mathcal{I}$. A subset of $\mathcal{I}$ containing k items is called a k-itemset. We can now define a predictive rule whose quality can be measured as explained below. Let X and Y be two itemsets such that $X \subset \mathcal{I}, Y \subset \mathcal{I}$ and $X \cap Y = \emptyset$. An association rule is a predictive pattern denoted by $X \Rightarrow Y$, where X is called the antecedent and Y is called the consequent.

Association rule have significance metrics, which are defined as follows. For an itemset X, support $s(X)$ is defined as the fraction of transactions $T_i \in D$ s.t. $X \subseteq T_i$. That is, support can be understood as the probability of appearance of the pattern X. Let $P(X)$ be the probability of X in D and let $P(Y|X)$ be the conditional probability of appearance of Y given X. Then $P(X)$ can be estimated as $P(X) = s(X)$. By the same reasoning, the support of a rule $X \Rightarrow Y$ is defined as $s(X \Rightarrow Y) = s(X \cup Y)$. The confidence of the rule $X \Rightarrow Y$ is defined as $c(X \Rightarrow Y) = s(X \cup Y)/s(X)$, which is basically $P(Y|X)$. A third metricm, called lift, is defined as $l(X \Rightarrow Y) = P(X \cup Y)/(P(X)P(Y)) = c(X \Rightarrow Y)/s(Y)$; when lift values are ≥ 1 they provide evidence that X and Y depend on each other. Lift values below 1 indicate X depends on the absence of Y or vice-versa. Lift is used to rank rules and discard redundant rules.

The association rule mining problem 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 s.t. $s(X) \geq \psi$ is called frequent.

C. Deep Neural Networks (DNNs)

We now turn our attention to deep neural networks (a.k.a. DNNs or multi-layer perceptrons (MLPs)) [1], [8], [25]. The raw input data set is S, which must be transformed into a data set X with only numeric features. Numeric attributes can be used "as is". Otherwise, if A_j 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. A Z-score was used to make this transformation. The goal is to predict the probability of a person having high stenosis in a specific artery. Since there are four target variables (arteries), four DNNs are computed (one DNN for each artery). The output for each DNN is the probability of having disease Y/N. Like most current DNN research, we use an ReLU and LeakyReLU activation functions (robust to noise) and neuron dropout (to improve generalization). Further details can be found on popular deep learning papers.

A powerful 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 ; 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:

$$g_k(X = \mathbf{x}) = f(\sum_l w_{kl}\ f(\sum_i w_{li} x_i + w_{l0}) + w_{k0}),$$

where $\mathbf{x}$ is the input feature vector, $f(\cdot)$ is a nonlinear (e.g., sigmoid, tanh, ReLU) function, and x_i is a component of vector $\mathbf{x}$. Subscript 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*, 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's considered when using neural networks. It ranges from 0 to 1, and [18] defines as:

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

The Hyperbolic tangent function (tanh) is defined as the ratio between the hyperbolic sine and the hyperbolic cosine. Its values range from -1 to 1 [18]

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

The Rectified Linear Unit function (ReLU) has become the most popular activation function used in neural networks today: [20]

$$ReLU = max(0, x) \tag{1}$$

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 to reduce the slope [18] .

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

There are also techniques to improve generalization. Dropout is a technique used to prevent overfitting 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 [28].

III. Comparing Both Techniques

This section provides details on pros and cons of association rules (ARs) and deep neural networks (DNNs) in the context of our medical problem.

A. Strengths and Limitations of ARs

Association rules have the advantage of being exhaustive: all patterns above the input thresholds are discovered. A rule with high confidence provides a valuable pattern, but its real predictive value is dictated by its support: the higher, the better. However, ARs do have several limitations and disadvantages. Unfortunately, many high-confidence rules tend to have low support. Running time can take hours: there is tradeoff between support and confidence. Many rules are meaningless from a predictive perspective: they must be filtered. The next section explains how these limitations are solved.

B. Search Constraints to Discover Medically Significant Rules

The most important user-specified parameters are minimum support and minimum confidence. In a perfect world, we would discover rules that have at least 50% support and 90% confidence. Unfortunately, those rules are rare. Support is the parameter that most greatly impacts running time and rule generalization. If support is too low, discovered rules apply to so few people that the prediction does not generalize to the whole population. Equally important, running time greatly increases, in general exponentially. On the other hand, confidence should be relatively high: above 90% is highly desirable, but difficult to get, below 70% has little predictive value as they approach tossing a coin.

We now explain search constraints. Without search constraints, the number of explored and discovered rules can be exponential and therefore the running time of the AR discovery algorithm drastically increases, –hours or days depending on p, the number of attributes and data set size n. The first constraint is limiting k, the number of items that can appear together, resulting in simpler rules, fewer rules, less passes over the data set. The rationale is that a rule with too many items is too specific and rules tend to be redundant. The second constraint is a template for rules, after frequent itemsets are discovered. We only care about rules where the predicted atribute (severe artery stenosis) is the consequent. Conversely, items from predicted attributes should not appear in the antecedent.

The use of additional constraints is necessary in order to reduce number of discovered rules and to make run time reasonable. Due to the fact that we have more computing power now (faster CPUs, more RAM) than we did ten years ago fewer constraints were applied. Finding rules with negation is a major challenge due to the explosion of number of patterns. Negation was not considered 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 items to discard well-known combinations of risk factors was neither necessary. Another limitation of association rules is the overlap of attributes, a form of redundancy. The same items may appear as subsets of the antecedent across many rules, but they have a different meaning and medical interpretation when considered as an overall rule.

C. Strengths and Limitations of DNNs

When utilizing DNNs, it is important to understand the impact that hyper-parameters can have on predictive accuracy. The number of layers and number of neurons per hidden layer are perhaps the most important hyper parameters in any classification or regression problem. This makes hyper-parameter optimization necessary, with various optimization techniques studied in the literature [24], [16], [4]. In our case, to improve DNNs accuracy to compare them with association rules, we experimentally studied fundamental hyper-parameters: the neuron activation function, the number of layers and the number of neurons per layer. Secondary parameters include the learning rate and neuron dropout, which we found out they had less impact on accuracy.

Overfitting is an issue in machine learning where the model is over trained on the training data to the point where accuracy gets close to 100%. Once this occurs, the model does not generalize well on the testing data (or new records) because it has become too specific, tailored to the training data set. The goal is to increase accuracy as much as possible before the model overfits and then scale back to a point where the testing accuracy does not decrease.

D. Examples

Association Rules Example

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

TABLE I: Example data set.

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

`200<=CHOL<250,SMOKE=Y => 70<=LAD<100`

In the above rule, support ψ, = 0.4, confidence α, = 0.67, and lift = 1.1. This example is a discovered rule that gives the following profile of patients: A patient that is between 60 and 100 years old, has a certain region of the heart that is considered to have defect, and has cholesterol between 200 and 250 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, $\frac{2}{5}$.

Deep Neural Network Example

An example of the DNN is made using the same mock data set as above. 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 you are 65 years old, have cholesterol of 253, and smoke, the probability of you having greater than 70% stenosis of LAD is 62%. It's 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.

TABLE II: Output for DNN on example data set.

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

IV. Experiments

This section discusses the setup used for our experimental evaluation. We explain hardware and software, and we provide detailed data set description as well as specific function call parameters for each technique. We provide a preliminary comparison evaluating predictive accuracy of both techniques.

A. Hardware and Software

Experiments were conducted on a machine with a Quad-core (4 core) Intel Pentium CPU @ 1.60GHz and 8GB of DDR3 1600 MHz memory. The operating system was Linux Ubuntu 18.04. In the future, we plan to use GPUs to accelerate the computation of DNNs as it has become the norm. Faster hardware will allow us to explore deeper nets and larger data sets.

ARs were found by a C++ program, with 5000 lines of source code with no special data mining libraries. DNNs were computed by a Python program with approximately 500 lines. The main Python libraries used were: Keras, TensorFlow, Scikit-learn, Pandas, and NumPy.

B. Medical Data Set

We should emphasize it is difficult to get access to medical data due to many security and privacy regulations. The data set explained below was shared by a US hospital, specialized in cardiology and radiology. We believe our findings hold for a other medical data sets for heart diease, but we cannot claim the challenges will be the same for other ailments. The data set used to compare the two predictive techniques contained records for cardiovascular disease. This data set was multimodal, containing both alphanumeric attributes and heart imaging data. The data set size was $n = 655$ and $p = 25$. Four additional "target" attributes were created by data scientists using data from the raw data set (details below). Alphanumeric attributes 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). Imaging attributes included a heart image divided into 9 regions (AL, AS, SA, SI, IS, IL, LI, LA, AP), and carotid artery stenosis in 4 arteries (LM, LAD, LCX, RCA). Severe disease of each artery was created based on stenosis ($\geq$ 50% for LM, $\geq$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 (stress from exercise), previous angina, previous stroke, and previous heart surgery. As previously mentioned, numeric variables are necessary for the neural network. Therefore, categorical variables need to be transformed into binary variables. 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 males and 0 for females. Remaining discrete attributes were transformed from 1 for "y" and 0 for "n".

We created binary variables for the neural network to train and test on. These variables were based on the discretized values of the arteries (LM, LAD,LCX, RCA). Each artery became a target attribute for severe disease Y/N, defined as $\geq$ 70% for LAD, LCX, and RCA and $\geq$ 50% for LM, with

1 if severe disease is present and 0 otherwise. Two additional changes were made to the patient data: missing data data values were replaced (mean for numeric, mode for categorical) and heart region image values were rescaled to prevent giving more importance to image data ([-1,1] was rescaled to [0,1]). The medical data set had a significant fraction of missing values. Missing patient data were replaced (imputed) as follows. For binary variables (smoking, previous cardiac surgery, and so on) the mode was taken for each sex. For numeric variables (cholesterol, age, or heart region images) the average was taken for each sex. Image data required a more complicated process. 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.

C. Input Parameters

Parameter Settings: Association Rules

We consider the following parameter categories: binning cutoffs, thresholds (support, confidence) and pattern search constraints. The first step was to bin stensosis attributes into separate ranges for healthy and unhealthy patients. 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 ≥ 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 bad. We grouped patients into three different age ranges: 0-40 (young), 40-60 (adult), and 60-100 (old).

Based on the opinion of doctors and clinicians, minimum confidence was set at 70%. This number is based on balancing identifying sick and healthy patients. Minimum support was $\psi = 0.02$. Minimum confidence was $\alpha = 0.70$. Finally, Lift minimum = 1. Rules that applied to approximately fewer than 33 patients in the data set (minimum support, ψ = 2%) were filtered out. Based on medical opinion, confidence, α, was set to 70%. Medically speaking, rules whose confidence is below 70% are not useful [21]. Minimum lift was set to 1.0.

In an effort to reduce the number of association rules generated the following constraints were applied. We increased the size of the antecedent from 4 items (in previous experiments), to 6 items. Rules with more items would have higher predictive accuracy, but they would be too specific. The heart arteries were restricted to appear only in the consequent of the rule. All other attributes were set to only appear in the antecedent. Maximum run time for association rules was set to thirty minutes.

Results: Discovered Predictive Association Rules

Without search constraints the association rule discovery problem becomes intractable [21], reaching one million itemsets below 1% support ($\approx 2^{20}$) and the number of derived rules is an order of magnitude larger [22]. In other words, the medical data set is small, but the number of patterns is large, indeed being big data. 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 below.

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 chart below. 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.

Table IV shows medically significant rules, validated by a cardiologist [21]. 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.

Out of potentially millions of rules search constraints filtered them down to a few thousand. Of the 2,634 rules, 1,851 of them had 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 a 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.

TABLE III: Attribute definitions: numeric, categorical and image.

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 IV: Medically significant rules

Antecedent	Consequent	Support	Confidence	Lift
{0.2<=SA<1.1,0.2<=LI<1.1,200<=CHOL<250}	{70<=LAD<100}	0.03	1.00	3.2
{0.2<=AS<1.1,0.2<=IL<1.1,0.2<=AP<1.1,SEX=F}	{70<=LAD<100}	0.03	1.00	3.2
{0.2<=SA<1.1,0.2<=LI<1.1,SEX=M,200<=CHOL<250}	{70<=LAD<100}	0.02	1.00	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	1.00	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	0.70	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	0.70	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 V: Statistics on support values.

	Whole Set	LAD	LCX	RCA
Min	0.02	0.02	0.02	0.02
Max	0.13	0.13	0.02	0.08
Mean	0.039	0.039	0.02	0.028
Std	0.017	0.017	0.00	0.013

TABLE VI: Statistics on confidence 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	0.779	0.743	0.742
Std	0.0558	0.0557	0.035	0.044

Parameter Settings: Deep Neural Networks

Many iterations of experimentation occurred in an effort to find the model that had the highest predictive capabilities. In order to accomplish this, hyper parameters for the neural network were changed for each experimental iteration (results presented in the chart below). Accuracy is most affected by the number of hidden layers of the neural network and 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. Maximum run time for neural networks was set to thirty minutes (stopped). 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 became the target variables. We also attempted to use Bayesian Optimization [16] to select which hyperparameters perform the best, but results were not encouraging (accuracy went down) and running time increased to hours; this is an issue for future research.

Results: Deep Neural Networks

As a correctness and accuracy check, we computed classic machine learning algorithms on the data to ensure better than chance predictions, and that pursuing the models with deep neural networks was the correct choice. 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.

Next, we conducted experiments with various neural network toplogies. The next paragraph describes why certain network choices were made. 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.

The best results were obtained from a deep neural network with four layers of 50 neurons per layer, an activation LeakyReLU function and dropout of 0.5. It is worth noting that the artery LM had higher prediction values than the

TABLE VII: Lift 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 VIII: Sanity check: comparing DNNs with other ML models.

Artery	Logistic Regression	SVM	DNN
LAD	71.8	74.8	77.9
LCX	77.1	77.9	79.4
RCA	67.2	67.2	71.0
LM	93.1	94.7	93.1

TABLE IX: Accuracy of DNNs tuning hyper-parameters.

Description of Network	LAD	LCX	RCA	LM
[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	77.0	77.4	71.4	92.2
[50,50,50,50] linear	76.1	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	71.0	93.1

other arteries, as expected. From a medical standpoint, LM is the base trunk from which the other arteries branch out of. Because of this anatomy, severe disease in LM is much more rare than disease in the other three arteries. Therefore, our data set lacks positive (sick) LM instances. In other words, the high predictive values from the neural network is due to predicting no disease in LM, ignoring those few patients who have LM stenosis.

We tried Bayesian Optimization to search for the best hyper-parameters. Bayesian Optimization was applied on two different networks as a trail with fifty iterations. Results showed lower accuracy than deep neural networks computed with default values. In addition, the run time was approximately between 40 and 75 minutes. We must mention 50 iterations is considered low for Bayesian Optimization, which may require thousands of iterations. In short, Bayesian optmization is a research issue for future work.

Comparing ARs and DNNs

We gave each technique its best opportunity. In comparing them, 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 is one of the main issues with association rules, the relationship between support and confidence: as one goes up the other goes down. 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 accuracy in the neural network, 71%. The mean confidence for rules containing RCA was 74.17%. The maximum confidence was 89%, but again it reached the minimum support of 2%. In this case, it is less clear which technique is better overall. However, neural networks do have the advantage of being a widely applicable and generalizable predictive model. From a time peformance perspective, as can be seen in Table X, ARs are 4X faster than DNNs because all rules involving the four arteries can be obtained on a single run, whereas a separate DNN is needed for each artery. Nevertheless, both techniques take several minutes.

TABLE X: Running times in minutes and seconds.

DNNs				ARs
LAD	LCX	RCA	LM	
6m 59s	7m 6s	7m 10s	7m 6s	4m 12s

V. Related Work

This section reviews closely related work on association rules and deep neural networks in the medical field.

Data mining on medical data presents unique challenges [26]. 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 [10], [19]. Association rules have been used to help infection detection and monitoring [5], [6], to understand what drugs are co-prescribed with antacids [7], to discover frequent patterns in gene data [3], to understand interaction between proteins [23] and to detect common risk factors in pediatric diseases [9]. [13] optimizes an algorithm

that incorporates search constraints into the association rule mining process. In [27], algorithms are proposed to include constraints that exclude or include certain items in the association rule, like we do.

Neural networks have been successfully applied in many medical and healthcare problems [2], [14]. Association rules and neural networks are used in tandem on a medical data set in [17], but this work does not compare their strengths and weaknesses for disease prediction.

VI. Conclusions

We compared predictive accuracy and speed between association rules and deep neural networks in a difficult multi-target predictive problem: predicting heart disease in the four arteries of the heart. For our experiments on a medical data set, results indicated a distinct advantage of DNNs over ARs in some cases and less clear results in others. Overall, basic DNNs beat ARs, but by a small margin, due to consistently higher predictive scores and a model generalizable on the entire data set. On the other hand, association rules found accurate predictive rules for LAD and RCA arteries with very high predictive accuracy (e.g. confidence=0.89, 1.0 impossible to reach with a DNN), but only on small subsets within the data set. In contrast, LM and LCX arteries had higher predictive accuracy in the neural network than in association rules. Furthermore, only three association rules were found with LCX. The deep neural network produced a model able to predict LCX with 79% accuracy, clearly preferable. RCA had a peak prediction accuracy of 71%. While not as high as ARs confidence, it still has the benefit of being generalizable to the entire data set. Unfortunately, higher accuracy in DNNs could not be obtained for several reasons. Perhaps the most important factor is the small data set size, which is common with medical data (especially if it comes from a single hospital, patient records are confidential, difficult to share between hospitals). A second reason is that we did not use deeper DNNs or CNNs. In contrast, with big data, large data sets have the luxury of being able to drop incomplete samples and still have a large training and testing set. Another machine learning issue with data sets is that DNNs tend to perform better on a balanced data set, which is hardly the case with patients having diverse heart ailments. We double checked our DNNs were better than other machine learning models (i.e. a sanity check): DNNs were indeed more accurate than SVMs and logistic regression, but by a small margin. In short, ARs and DNNs complement each other.

Our work opens several research issues. For the medical data set the predictive accuracy of deep neural networks was not as outstanding as it happens with benchmark image classification problems (e.g. CIFAR). We want to understand in more depth why that happens. Medical images are noisier and have less well defined patterns than images in other domains. We plan to use Convolutional Neural Networks (CNNs), which can identify hidden patterns on images. To get more accurate models based on conditional probabilities, it is necessary to find subsets from the data set, where the deep neural network fit is better. We have shown ARs can complement DNNs. Thus we believe association rules can guide and explain the deep neural network, given their simplicity and exhaustive search capabilities. On the other hand, deep neural networks can enhance association rules, given their ability to capture non-linear behavior. Finally, we plan to apply Bayesian Optimization with a large number of iterations to automatically determine DNN best parameter settings.

Acknowledgments

The authors thank Dr Cesar Santana who provided initial guidance to identify which rules were medically significant. The authors thank the Emory Hospital (Atlanta, USA), for sharing the data set used in this paper.

References

[1] Alex Aizman, Gavin Maltby, and Thomas Breuel. High performance I/O for large scale deep learning. In *IEEE International Conference on Big Data (BigData)*, pages 5965–5967. IEEE, 2019.

[2] Émilien Arnaud, Mahmoud Elbattah, Maxime Gignon, and Gilles Dequen. Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. In *IEEE International Conference on Big Data (IEEE BigData)*, pages 4836–4841. IEEE, 2020.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] Jeffrey Dean, Greg Corrado, Rajat Monga, and et al. Large scale distributed deep networks. In *Proc. Advances in Neural Information Processing Systems*, pages 1232–1240, 2012.

[9] 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.

[10] 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.

[11] Shweta Garg, Raghavan Krishnan, Sarangapani Jagannathan, and V. A. Samaranayake. Distributed learning of deep sparse neural networks for high-dimensional classification. In *IEEE International Conference on Big Data, Big Data 2018*, pages 1587–1592. IEEE, 2018.

[12] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2006.

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

[14] Song Ju, Yeo Jin Kim, Markel Sanz Ausin, Maria E. Mayorga, and Min Chi. To reduce healthcare workload: Identify critical sepsis progression moments through deep reinforcement learning. In *IEEE International Conference on Big Data (BigData)*, pages 1640–1646. IEEE, 2021.

[15] Daniel Justus, John Brennan, Stephen Bonner, and Andrew Stephen McGough. Predicting the computational cost of deep learning models. In *IEEE International Conference on Big Data (IEEE BigData)*, pages 3873–3882. IEEE, 2018.

[16] Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Neural architecture search with bayesian optimisation and optimal transport. In *Advances in Neural Information Processing Systems*, pages 2016–2025, 2018.

[17] Murat Karabatak and M Cevdet 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.

[18] Bekir Karlik and A Vehbi 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Fabrício José Pontes, GF Amorim, Pedro Paulo Balestrassi, AP Paiva, and João Roberto Ferreira. Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing*, 186:22–34, 2016.

[25] Sebastian Raschka and Vahid Mirjalili. Python machine learning: Machine learning and deep learning with python. *Scikit-Learn, and TensorFlow. Second edition ed*, 2017.

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

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

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