Utilizing Machine Learning to Help Combat Cardiovascular Disease

David Strube
dstrube3@gatech.edu

Abstract—Every year, out of all known causes of death, cardiovascular disease kills the greatest number of people in the United States of America. Machine learning (a subset of Artificial Intelligence) involves using various algorithms to predict outcomes. These algorithms improve their predictive performance as the amount of data they have to work with increases. By applying certain machine learning algorithms to the problem of cardiovascular disease, vast improvements in healthcare may be made and hundreds of thousands of lives may be saved every year.

1 BACKGROUND AND SIGNIFICANCE

According to the American Heart Association, cardiovascular disease killed on average 2,353 people every day in 2017 (Virani et al., 2020). In 2019, approximately 655,000 Americans died from cardiovascular disease. According to the Centers for Disease Control and Prevention, the cost of cardiovascular disease in the United States from 2014 to 2015 was about $219 billion; and 30.3 million adults - 12.1% of all adults in the US - were diagnosed with cardiovascular disease in 2018 (CDC, 2020).

1.1 Current State of Healthcare

In 2020, COVID19 quickly became the third leading cause of death in the US. Even now in 2021, COVID19 is still making headlines every day. With limited supplies of a vaccine available to some people, COVID19 is a serious matter here and in much of the rest of the world. However, cardiovascular disease is still the #1 killer in the US and should not be forgotten as something that must be taken seriously. Indeed, given the ongoing state of quarantine and its results on Americans' already sedentary lifestyle, not to mention its effects on most citizens' mental state, COVID19 may end up exacerbating the impact cardiovascular disease has on us.

2 PROBLEM

What is cardiovascular disease? Cardiovascular disease is a group of diseases that involve the blood vessels or heart. Cardiovascular disease includes: venous thrombosis, thromboembolic

disease, peripheral artery disease, aortic aneurysms, carditis, valvular heart disease, congenital heart disease, abnormal heart rhythms, cardiomyopathy, rheumatic heart disease, hypertensive heart disease, heart failure, and stroke. Coronary artery diseases are a subset of cardiovascular disease which include myocardial infarction (also known as a heart attack) and angina. Suffice it to say cardiovascular disease is not one thing.

The main risk factors for cardiovascular disease are: high blood pressure, high blood cholesterol, and smoking. Other medical conditions and lifestyle choices that can also contribute to higher chances for cardiovascular disease include: diabetes, obesity, unhealthy diet, lack of physical activity, and drinking too much alcohol. (Those last three items may sound familiar to anyone who has had to self-quarantine for the past several months, and will continue to have to do so for the foreseeable future.)

2.1 Specific demographics

Cardiovascular disease does not affect all demographics equally, but it is a leading cause of death among most demographics. The risk of death from cardiovascular disease among black people is almost double that of white people (CDC, 2013). Cancer is the leading cause of death among women from the Pacific Islands and Asian American, American Indian, Alaska Native, and Hispanic women; cardiovascular disease is the second leading cause of death for them (CDC, 2020).

Race of Ethnic Group	**% of Deaths**	**Men, %**	**Women, %**
American Indian or Alaska Native	18.3	19.4	17.0
Asian American or Pacific Islander	21.4	22.9	19.9
Black (Non-Hispanic)	23.5	23.9	23.1
White (Non-Hispanic)	23.7	24.9	22.5
Hispanic	20.3	20.6	19.9
All	23.4	24.4	22.3

Table 1—Percentages of deaths caused by cardiovascular disease, broken up by ethnicity, race, and sex (CDC, 2020)

3 PROPOSED IDEA

The proposed idea is to gather patient data, feed the data into multiple machine learning algorithms to determine which one(s) (and which configurations) are most likely to yield accurate predictions, and use those algorithms to make predictions on how likely a patient is to suffer from a serious cardiovascular disease incident. The patient data would include blood pressure, what medications they're taking, what their diet is, how much exercise they're getting, their income, location, and whether they've recently suffered a serious cardiovascular disease incident. The patient data would be gathered voluntarily either at varying time periods (constant blood pressure monitoring from a smart device, daily diet reporting, and weekly cardiovascular general health surveying) or at the same time period (daily or weekly).

Machine learning is used to build models around some input data known as *training data* in order to make predictions on another set of data known as *test data* which is used to test the accuracy of predictions. Machine learning uses a variety of algorithms that can be organized into a few categories including *supervised learning* and *unsupervised learning*. (There are other categories, but this paper will stick with covering just these two.) In supervised learning, the input data is classified into categories called *labels* or *features*; in unsupervised learning, there are no labels, just data points that leave it up to the algorithm to find patterns. For this proposal, supervised learning algorithms will be the focus because the categories of data will be known.

Using a supervised machine learning algorithm, it may be possible to predict with some degree of accuracy an individual's likelihood of having cardiovascular disease based on a large amount of data coming from a number of variables that are risk factors for cardiovascular disease.

3.1 Algorithms

There are many supervised learning algorithms from which to choose. Some of the most widely used supervised learning algorithms are:

- `Decision trees`
- `K-nearest neighbor`
- `Neural networks`

There are other widely used supervised learning algorithms, but these are the ones explored in this paper because they seem to be the most applicable to this problem. The decision tree machine learning algorithm uses a model representing the data in the form of a decision tree where each decision of a feature value is a branch in the tree and the final value arrived at is the

leaf. In K-nearest neighbor, for each data point in the data set, it is determined what category that data point is in based on what are the next nearest data points. In neural networks, each input is fed into a function known as a perceptron which gives a calculated weight to each input to determine how likely that input is to lead to an expected result.

Picking out the right algorithm for the task has to take into account many factors. One consideration is the issue of *bias versus variance*. Bias can be described as the probability that a machine learning algorithm will give an incorrect prediction based on some input. Variance is the reliability with which an algorithm will give a certain prediction given various inputs. There is a tradeoff between these two factors - high bias generally means low variance and high variance generally means low bias.

One should also take into account the *dimensions* of the data. Data dimensionality refers to the number of labels an algorithm must use to make a prediction. Greater dimensionality can lead to more accurate results, but it can cause exponentially longer processing times, so much so that the increased accuracy one gets from more dimensions isn't worth it.

Depending on which algorithm one decides on, there may be some variations on that algorithm one may have to choose among. For example, if one decides to use a decision tree algorithm, one may have to choose between using pre-pruning, post-pruning, or no pruning. Pre-pruning is the process of letting the algorithm know that it can stop searching at a certain point so as to not waste time exploring certain nodes of a decision tree. Post-pruning is the process of deciding which nodes of a decision tree can be eliminated to yield a greater accuracy of the results.

Another thing to consider is the amount of data being processed by the algorithm. The greater the amount of data, the more accurate the prediction will be, but the longer it will take for the algorithm to make predictions. Note the accuracy of the predictions increasing with the amount of data assumes the algorithm is well suited for the task and the parameters chosen are well tuned.

3.2 Parameters

Once an algorithm has been decided upon, one must also consider the various combinations of parameters that an algorithm has to explore. For example, I have found neural networks to be a very good algorithm for finding a higher degree of accuracy for some input data, but if one decides to use a neural network algorithm, there are many parameters one must consider. In the Python library for machine learning known as scikit-learn, the type of neural network class

most suited for this undertaking would be the Multi-layer Perceptron classifier (Pedregosa et al., 2011). In that class, the parameters include: `hidden_layer_sizes`, `activation`, `solver`, and `learning_rate`. The `hidden_layer_sizes` parameter indicates the number of neurons at each layer. The possible values of `activation` are: *identity*, *logistic*, *tanh*, and *relu*. The possible values of `solver` are: *lbfgs*, *sgd*, and *adam*. The possible values of `learning_rate` are: *constant*, *invscaling*, and *adaptive*. In searching for an optimal setting for an algorithm, one that yields the most accurate predictions, all possible combinations of these last three parameters and more may have to be searched through, and various combinations of `hidden_layer_sizes` should be tried as well.

4 COMPLEXITY

4.1 Policy, Privacy, Security

A policy of trying to save the greatest number of lives with the technology that we now have while trying to balance concerns of privacy and security would be tricky at best. Given the complexity and political difficulty of passing and navigating current healthcare policies such as HIPAA and ACA, it may take years to accomplish authoring, lobbying for, and passing a policy with the stated goal of utilizing machine learning to help combat cardiovascular disease. Researching all the information necessary to prepare such policy would be a costly and time consuming undertaking. Finding funding for this policy would be another huge hurdle to overcome.

One other major concern is privacy. Many people may feel uneasy about a government undertaking that involves citizens' personal health data being collected on a massive scale to be processed in a set of complex mathematical algorithms. These concerns may be somewhat alleviated by emphasizing that all data collection is done voluntarily and that participants are free to discontinue participation at any time. Further steps may be taken to allow participants to have their data removed from the training and testing data sets, but this may prove to be an impossible task if patient data is to be securely anonymized.

Another major concern is the security of patient data. Any breaches of the security of patient data would undermine the entire operation, not only bringing us back to square one, but also setting back the whole enterprise of using machine learning to better the lives of everyone. Extremely careful precautions must be taken from the outset and every step along the way to ensure that patient data is obtained, retrieved, processed, and stored securely.

4.2 Feasibility

It is difficult to say how feasible this proposal would be. On the one hand, it sounds like a good idea to put our current state of machine learning algorithms to good use in preventing deaths from the greatest source of death in our nation, an idea so good that it would override the personal concerns of privacy and security pushing back against the idea. On the other hand, it is easy to underestimate how much political resistance one can encounter from the opponents of an idea. Moreover, the technical challenges preventing this proposal from coming to fruition may be too great to overcome.

5 DISCUSSION

In the set of data collected from participants, instead gathering **whether** they've recently suffered a serious cardiovascular disease incident, we could alternatively narrow down the data to specify **when** was their last cardiovascular disease incident and what was its **severity**; or even further: when was their last heart attack, when was their last stroke, etc. These data points could be used to make predictions that are more specific and targeted. However this would require more data to make accurate predictions.

There is also the problem of how to deal with situations when a patient has died from cardiovascular disease, in which case this data would have to be entered by a caretaker, family member, or friend who the participant has designated ahead of time as authorized to do so. This may bring with it another set of security and ethical concerns.

Since the risk of death from cardiovascular disease among some demographics is so much higher than that of some others, it might make sense to advertise this idea to those demographics first. This advertisement may be particularly interesting once a sufficient amount of data has been collected to make accurate predictions.

5.1 Related Technical Challenges

The technical challenges of this proposal are many. Some of the challenges include: signal-to-noise, overfitting, computing resources, and algorithm / parameters selection. Signal-to-noise is the problem of determining which data are useful and which are not. Overfitting is the problem of figuring out whether the machine learning algorithm has been tailored too closely to the training data so that the predictive power on the training data is maximized and any anomalous data are accounted for leading to a model that will be inaccurate on test data. Computing resources will have to be allocated for the storage and processing of all

the patient data. A greater number of participants will require larger disk space and more computer processing power. The selection of which algorithms to use and which parameters of those algorithms can quickly compound in such a way that drastically slows the progress of the project, especially if one wants to make an exhaustive search of all options.

6 CONCLUSION

There are some things everyone can do to reduce their risk of cardiovascular disease. These include talking to one's health care provider about their chances of having cardiovascular disease, not smoking, going out for a walk multiple times a day throughout the week, not drinking too much, and eating a healthier diet.

If an individual could see predictive numbers indicating they have a much higher than normal likelihood for cardiovascular disease based on reproducible mathematical methods, then they may be more inclined to pursue some of the preventative measures available to avoid cardiovascular disease. Conversely, if an individual could see that they have a much lower than normal likelihood for cardiovascular disease, then they might learn that their efforts to avoid cardiovascular disease would be better spent on avoiding a more likely cause of death, like cancer or COVID19.

7 REFERENCES

1. Virani, S. S., Alonso, A., Benjamin, E. J., et al. (2020). Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. https://www.ahajournals.org/doi/10.1161/CIR.0000000000000757. Published January 29, 2020. Accessed January 30, 2021.
2. CDC. (2020). Heart Disease Facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm. Published September 8, 2020. Accessed January 30, 2021.
3. CDC. (2013). Preventable Deaths from Heart Disease & Stroke. Centers for Disease Control and Prevention. https://www.cdc.gov/vitalsigns/heartdisease-stroke/index.html. Published September 3, 2013. Accessed January 31, 2021.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. pp. 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Published January 1, 1970. Accessed January 31, 2021.