EPiC Series in Computing

Volume 83, 2022, Pages 43–52

Proceedings of 14th International Conference on Bioinformatics and Computational Biology

Simple evolutionary algorithm for quantifying how medical history factors predict disease outcomes

James Camp[1,2] and Hisham Al-Mubaid[1]

[1] University of Houston – Clear Lake, Houston, Texas

[2] Lee College, Baytown, Texas

jamespcamp@outlook.com, hisham@uhcl.edu

Abstract

The medical history information contained in electronic health records (EHR) is a valuable and largely untapped data mining source for predicting patient outcomes and thereby improving treatment. This paper presents a simple but novel evolutionary algorithm (EA) for identifying how various medical history and demographic factors predict clinical outcomes. For this initial study, our EA was tested using synthetic data concerning COVID-19 hospitalization rates and we show that the EA results are more informative than logistic regression, neural network, or decision tree results.

1 Introduction

Interpretable and explainable AI has become a major area of research in the machine learning community [1]. Clinical professional such as doctors are, in particular, thought to be highly skeptical of any machine learning models that cannot easily be explained to them [2]. In recent years, many studies have attempted to predict patient disease outcomes from patient information ranging from patient demographics to noted symptoms and clinical measurements [3], [4], [5], [6], [7]. Recently, much of the research in this field has zeroed in on electronic health record (EHR) data [2], [8]. Most of this data has been fed into machine learner algorithms that have a low human interpretability, such as SVM [9], [10]; random forest [9], [10], [11], [12], [13], [14], [15]; or neural network models [9], [14], [16]. Logistic regression, which is more interpretable, is commonly used as a baseline standard against which to judge these more advanced machine learners. Decision trees, which aim to produce a human-interpretable output, have been used to some success [6], [9], [13]. To our knowledge, however, no research has been done into using evolutionary algorithms (EA) to interpret this clinical data.

This study proposes a simple but novel EA that fits coefficients for an equation linking all medical history and demographic factors of interest, both individually and in pairwise combination, to the clinical outcome of interest. We decided to focus on five factors that have been of interest in predicting

COVID-19 outcomes: history of asthma, heart disease, hypertension, diabetes, and advanced age (defined here as 65 or older). Prior to trying this tool out on real-world medical data, much of which is still held as private [2], we decided to calibrate our model against a synthetic COVID-19 dataset where we defined (and therefore knew) the percent chance by which each model factor would lead to severe disease and hospitalization. To this end, we generated a 10,000-patient training data set and a 5,000-patient testing data set and submitted them our novel EA. As experimental controls, we used the baseline-standard logistic regression method, a multilayer perceptron neural network, and a simple but standardized decision-tree learner.

The outcome of this study, then, is not intended to give any new insights about COVID-19 but rather to serve as proof-of-concept for a novel machine learner, which can then be applied to any number of clinical datasets to determine risk factors for severe disease.

2 Methods

2.1 Evolutionary Algorithm

A genetic algorithm was constructed to find coefficients for the following equation:

$$Y = \sum_{i=1}^{n} c_i X_i + \sum_{j=1}^{n-1} \left[\sum_{k=j+1}^{n} c_{jk} X_j X_k \right]$$

where the inputs are n medical history factors, X_i, treated as binary numbers, and the output is the clinical outcome of interest, Y, again treated as a binary number. This equation captures the predictive effects of each factor in isolation as well as any pairwise interactions, or synergies, between factors.

For the present study we decided to consider five medical history and demographic factors such that

$$outcome = c_A A + c_B B + \cdots + c_E E + c_{AB} AB + c_{AC} AC + \cdots + c_{DE} DE$$

where A = asthma, B = heart disease, C = diabetes, D = hypertension, and E = advanced age (65+).

Data structures `Patient{A, B, C, D, E, outcome}` and `Gene{cA, cB, … , cDE}` were defined, along with a mutate subroutine shown in the pseudocode below:

```
Mutate(Gene g0):
    g1 = g0
    Randomly choose one coefficient
    g1.chosen_coeff = g0.chosen_coeff + normal_dist(0.0, SIGMA)
    return g1
```

The normal distribution used here was from the C++ standard library; σ was initially set to 0.5, but was reduced in stages as the model narrowed in on a solution.

The fitness function for a Gene g and a Patient p was defined as:

$$\boldsymbol{fitness} = \mathbf{1} - [(\boldsymbol{g.c_A \cdot p.A} + \cdots + \boldsymbol{g.c_{DE} \cdot p.D \cdot p.E}) - \boldsymbol{p.outcome}]^{\mathbf{2}}$$

A few tunable parameters of the algorithm were set empirically or through a process of trial-and-error: α, the rate by which σ shrinks as the algorithm narrows in on an answer, was set to 0.9 so that the algorithm did not narrow in too quickly; λ, the number of genes in one generation, was set to the number of factors cubed (found to be the minimum necessary to get reliable improvements in each generation); and the threshold for σ below which the algorithm would exit was set to 0.005 (found to be the maximum value for which there was good agreement between subsequent runs).

Finally, then, the evolutionary algorithm itself can be expressed as:

```
Initialize SIGMA = 0.5, ALPHA = 0.9, LAMBDA = NUMFACTORS^3
Gene G0 { 0.0, … 0.0 } // initialize all coefficients to 0.0
G0.fitness = Sum of Fitness(G0, P) over all P in list Patients
Unsuccessful = 0 // count of unsuccessful mutations
While SIGMA >= 0.005 do
    If Unsuccessful >= 2 then
        SIGMA *= ALPHA
        Unsuccessful = 0
    For j = 1 to LAMBDA do
       Genes[j] = Mutate(G0)
       Genes[j].fitness = Sum of Fitness(Genes[j], P) for all P
    Sort Genes[j] in order of decreasing fitness
    If Genes[0].fitness > G0.fitness then
       G0 = Genes[0] // replace G0 with most-fit child
    Else
       Unsuccessful++ // no improvements in this generation
Output(G0)
```

2.2 Other machine learning algorithms

The widely-known WEKA machine learning workbench software [17] was used to compare the results of this EA with other common machine learning algorithms: logistic regression, multilayer perceptron neural network, and C4.5 decision tree. All algorithms were used with their default parameters to ease comparison with other work.

2.3 Synthetic data generation

Two datasets were generated: a 10,000 patient training set and a 5,000 patient testing set. Each dataset was assigned medical history and demographic factors in a random process according to the prevalence percentages given in Table 1.

Table 1. Prevalence of five medical history and demographic factors in the US population.

Factor	Population	Prevalence	Source
Asthma	US Adults (18+ yrs)	8.0%	[18]
Diabetes	US Adults (18+ yrs)	13.0%	[19]
Coronary Heart Disease	US Adults (20+ yrs)	6.7%	[20]
Hypertension	US Adults (18+ yrs)	29.0%	[21]
Advanced Age (65+)	US Adults	21.2%	[22]

For this supervised-learning study, the class variable that the learners are attempting to predict is the probability of hospitalization for patients presenting at a clinic with symptoms of COVID-19. The

baseline rate for hospitalization has been reported as 5% [23]. Based on odds ratios obtained in late 2021 [24], enhanced rates of hospitalization were calculated for each risk factor (see Table 2).

Table 2. Rates of hospitalization used to generate artificial dataset.

Risk factor	Hospitalization rate
Asthma alone	6.4%
Heart Disease (all types)	27.0%
Diabetes (both types)	19.5%
Hypertension alone	40.5%
Advanced Age (65+) alone	40.4%

No hard data was found for synergistic effects of various risk factors. Based on intuition and qualitative information found online, either 25% synergy (advanced age + either asthma, heart disease, or diabetes), 20% synergy (hypertension combined with advanced age or heart disease), or 10% synergy (all other combinations) was applied to particular risk factor combinations. The total risk percent was then calculated for each pair of factors by adding the two individual risks and multiplying by the synergy factor (i.e. 1.25 for 25% synergy).

For each risk factor or two-factor combination that a simulated patient possessed, that patient was assigned a random number between 0 and 100; their hospitalization flag was set to "true" if that number was less than or equal to the percent risk. For example, if a patient had asthma, the algorithm would assign them a Hospitalized outcome if their randomized percent score was ≤ 6.4%, but if they also had advanced age, the algorithm would assign them a Hospitalized outcome for scores ≤ 58.5%.

The following printout summarizes the training data:

```
2191 patients with adv age (21.91%) -- 1489 (67.9598%) hospitalized
768 patients with asthma (7.68%) -- 334 (43.4896%) hospitalized
651 patients with heart disease (6.51%) -- 410 (62.98%) hospitalized
1300 patients with diabetes (13%) -- 737 (56.6923%) hospitalized
2891 patients with hypertension (28.91%) -- 1779 (61.5358%) hospitalized
4223 patients with none (42.23%) -- 217 (5.13853%) hospitalized
```

As you can see from this printout, since many patients have multiple history factors, the relative rates of ICU hospitalization when patients are itemized by history factor appear to be higher than the actual risk rates. One question of this study will be whether our simple EA can sort out these effects.

3 Results

3.1 Evolutionary Algorithm

Running the evolutionary algorithm on the training dataset yielded the fit coefficients in Table 3, with a normalized fitness score of 0.876747, where A = asthma, B = heart disease, C = diabetes, D = hypertension, and E = advanced age (65+). Fit coefficients are the average of results from seven runs, with an error margin of between 0.00011 and 0.00022. On average, runs took 331 generations to converge, with a range of 295 to 380.

Table 3. Fit coefficients from Evolutionary Algorithm

Factor	Coeff	Factor	Coeff	Factor	Coeff	Factor	Coeff
A	0.0748	E	0.4415	A E	0.1321	C D	0.1096
B	0.3117	A B	0.0846	B C	0.0349	C E	0.0486
C	0.2185	A C	0.0323	B D	0.0027	D E	0.0367
D	0.3889	A D	0.1131	B E	0.0123	C0	0.0356

It appeared from a cursory observation of the coefficients that each single-variable coefficient was very close to the actual risk percent for the corresponding single factor, prompting speculation that the fit equation would predict the actual disease risk for each patient. Substituting each single factor or two-factor combination into the fit equation with the EA-determined coefficients produced a response that closely resembled the actual hospitalization risk. The results of this operation are presented in Figure 1.

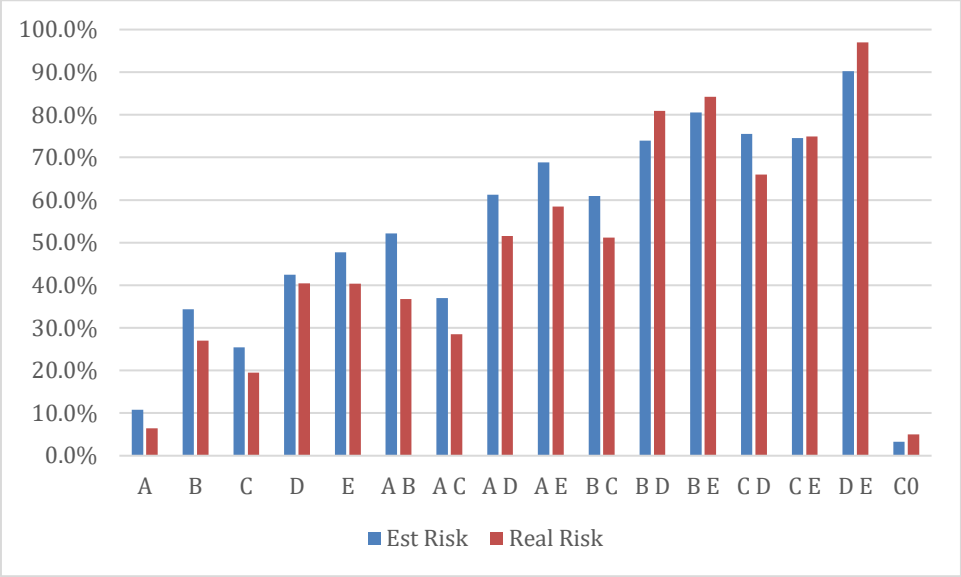

Figure 1. Percent chance of hospitalization as estimated by a novel genetic algorithm as compared to actual percent chances used to generate the data.

The test data was then processed with the regression equation to produce a risk prediction for each patient. This data was processed with the "ROCR" R package to obtain the ROC curve shown in Figure 2. Area under the ROC curve was 0.868.

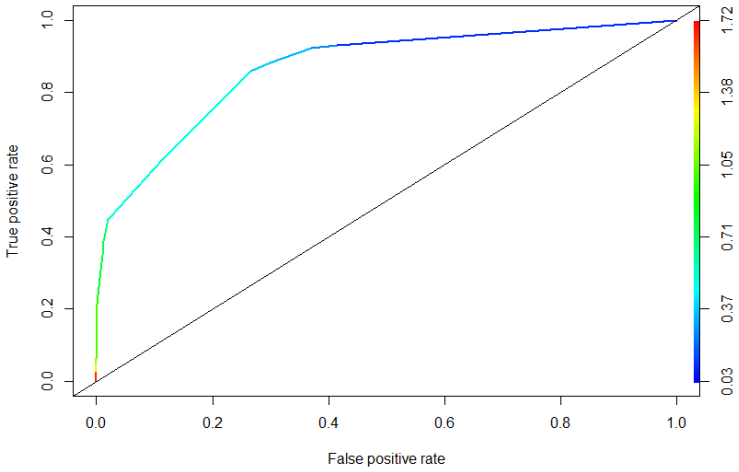

Figure 2. Receiver-Operator Characteristic (ROC) curve for EA model of synthetic COVID data.

3.2 Logistic Regression

The logistic regression model produced the coefficients seen in Table 4 and the confusion matrix seen in Table 5. The ROC area for this model was 0.868.

Table 4. Coefficients for logistic regression.

Variable	Class No
ASTHMA=Yes	-1.0909
HRT_DIS=Yes	-2.3872
DIABETES=Yes	-1.9578
HYPERTEN=Yes	-2.9075
ADV_AGE=Yes	-3.1404
Intercept	3.2621

Table 5. Confusion matrix for logistic regression.

Classified as No	Classified as Yes	
3376	79	True No
845	700	True Yes

3.3 Decision-Tree Learner

The C4.5 decision tree algorithm in WEKA produced the tree shown in Figure 3, and again the confusion matrix was identical to that shown in Figure 5. The ROC area for this model was 0.856.

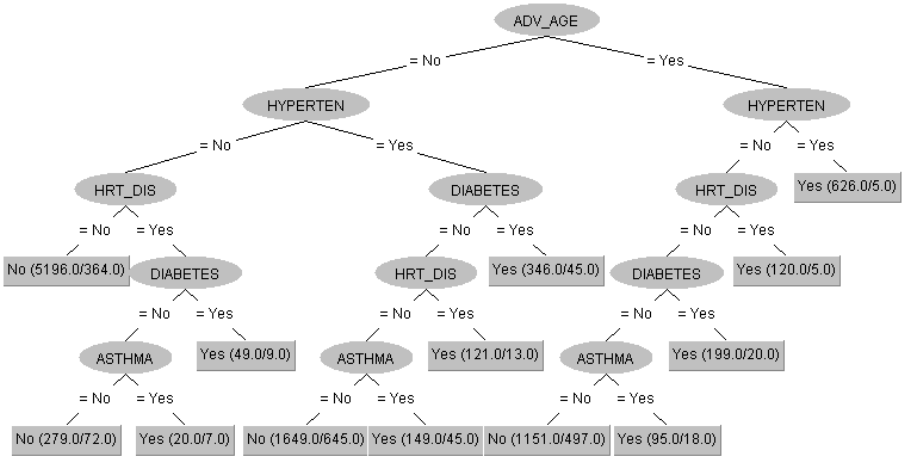

Figure 3. C4.5 decision tree based on unequalized dataset.

3.4 Multilayer Perceptron

The multilayer perceptron neural-network model produced five nodes with weights given in Supplemental Figure 4 at the end of this paper. The confusion matrix was identical to the one shown in Table 5 for the logistic regression model. The ROC area for this model was 0.868.

4 Discussion

The single-factor coefficients produced by the EA clearly assign correct weights to each factor – giving hypertension a much larger coefficient than asthma, for example – and even assigns coefficients to each factor that roughly correspond to the percent chance of each attribute predicting hospitalization. The intercept coefficient, C_0, also comes close to identifying the correct percent chance of hospitalization for a patient with none of these risk factors. The remaining coefficients of the equation look at interactions between two attributes. These interaction coefficients range from miniscule – less than 0.005 – to substantial – up to 0.14. None of the coefficients in this model is negative, which might indicate that that factor or interaction between factors was somewhat protective against strong disease, though this researcher has seen negative coefficients in preliminary studies of real-world data.

Figure 1 indicates that while the EA model predictions do not agree exactly with the percent risks – they particularly seem to overestimate the risks for each single factor, though this is perhaps an artifact of the process by which the data were generated – the predictions are sufficiently close to the actual risks that it is reasonable to expect that models produced with this method could be used to gauge the quantitative risks that each medical history and demographic factor – and each two-factor interaction – poses for severe disease, potentially a large step forward in interpretable machine learning. The somewhat poor ROC curve was no worse than any of the other models tested, indicating that this is simply a difficult dataset to predict.

Two of the three "control" models – the logistic regression and the multilayer perceptron – were equal to the EA model in terms of predictiveness as measured by ROC area. When it comes to assessing the potential of each for interpretable AI, however, the neural network model fails to meet this standard. Supplemental Figure 4 shows that, other than the fact that asthma is consistently weighted less than the other factors, there are few clear patterns that can be discerned from the network node descriptions. The logistic regression model can be seen to produce factor coefficients that scale somewhat proportionally to the percent risk that each factor represents, but unlike the EA model the LR model predictions cannot simply be read as equal to the risk percent for each factor or combination of factors.

The decision tree model was somewhat less predictive, from a ROC area standpoint, than the other two control models. Its confusion matrix, however, was the same as the other two: it was good at avoiding false positives, but predicted more false negatives than true positives. While one might expect the decision tree to be a clear winner in the interpretability category, the tree produced by this model was in fact difficult to interpret. It correctly predicts advanced age and hypertension as the two top-level concerns, and asthma as a bottom-level concern, but its treatment of heart disease and diabetes is somewhat confusing, and the exact interactions it predicts as most important in the middle of the tree is not very informative.

5 Conclusions

The simple but novel genetic algorithm presented in this paper may represent a leap forward in interpretable machine learning for clinical applications. Not only did it assign model coefficients in proportion to the risk that each factor represented, but it also detected interactions between risk factors, something that is lacking from many other models. Perhaps more significant, however, the model predictions for each factor or two-factor combination seem to be almost a numerical match for the percent risk that those factors or combinations represent. The goal of an interpretable or explainable AI is not only to make accurate predictions but to point out patterns in the data such that human domain

experts, such as clinicians, could use the model to gauge actual risks for each patient and to see how each factor contributes to that risk. Such a model could have implications in other fields as well, such as estimating the percent risk that particular defects might represent for device failure. In future studies, this machine learner, or a more advanced form of it, will be trained on publicly-available data in an effort to gain insight into real world disorders.

References

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2018, pp. 80–89. doi: 10.1109/DSAA.2018.00018.

[2] F. Shamout, T. Zhu, and D. A. Clifton, "Machine Learning for Clinical Outcome Prediction," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 116–126, 2021, doi: 10.1109/RBME.2020.3007816.

[3] J. R. Curtis *et al.*, "Data-Driven Patient Clustering and Differential Clinical Outcomes in the Brigham and Women's Rheumatoid Arthritis Sequential Study Registry," *Arthritis Care Res.*, vol. 73, no. 4, pp. 471–480, 2021, doi: 10.1002/acr.24471.

[4] D. Gregori, R. Bigi, L. Cortigiani, F. Bovenzi, C. Fiorentini, and E. Picano, "Non-invasive risk stratification of coronary artery disease: an evaluation of some commonly used statistical classifiers in terms of predictive accuracy and clinical usefulness," *J. Eval. Clin. Pract.*, vol. 15, no. 5, pp. 777–781, 2009, doi: 10.1111/j.1365-2753.2008.01034.x.

[5] B. C. Munsell *et al.*, "Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data," *NeuroImage*, vol. 118, pp. 219–230, Sep. 2015, doi: 10.1016/j.neuroimage.2015.06.008.

[6] N. Nakayama *et al.*, "Algorithm to determine the outcome of patients with acute liver failure: a data-mining analysis using decision trees," *J. Gastroenterol.*, vol. 47, no. 6, pp. 664–677, Jun. 2012, doi: 10.1007/s00535-012-0529-8.

[7] J. Kwon, K.-H. Kim, K.-H. Jeon, and J. Park, "Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography," *Echocardiography*, vol. 36, no. 2, pp. 213–218, 2019, doi: 10.1111/echo.14220.

[8] Md. E. Hossain, A. Khan, M. A. Moni, and S. Uddin, "Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 18, no. 2, pp. 745–758, Mar. 2021, doi: 10.1109/TCBB.2019.2937862.

[9] R. Akula, N. Nguyen, and I. Garibay, "Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes," in *2019 SoutheastCon*, Apr. 2019, pp. 1–8. doi: 10.1109/SoutheastCon42311.2019.9020358.

[10] N. S. Rajliwall, G. Chetty, and R. Davey, "Chronic disease risk monitoring based on an innovative predictive modelling framework," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Nov. 2017, pp. 1–8. doi: 10.1109/SSCI.2017.8285257.

[11] B. J. Mortazavi *et al.*, "Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1719–1729, Nov. 2017, doi: 10.1109/JBHI.2017.2675340.

[12] M. Fu, J. Yuan, and C. Bei, "Early Sepsis Prediction in ICU Trauma Patients with Using An Improved Cascade Deep Forest Model," in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, Oct. 2019, pp. 634–637. doi: 10.1109/ICSESS47205.2019.9040774.

[13] C. Mugisha and I. Paik, "Pneumonia Outcome Prediction Using Structured And Unstructured Data From EHR," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 2640–2646. doi: 10.1109/BIBM49941.2020.9312987.
[14] B. Wang *et al.*, "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients With Electronic Health Records," *IEEE Access*, vol. 7, pp. 178392–178400, 2019, doi: 10.1109/ACCESS.2019.2956859.
[15] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "A multi-modal machine learning approach towards predicting patient readmission," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 2027–2035. doi: 10.1109/BIBM49941.2020.9313588.
[16] T. Wanyan *et al.*, "Relational Learning Improves Prediction of Mortality in COVID-19 in the Intensive Care Unit," *IEEE Trans. Big Data*, vol. 7, no. 1, pp. 38–44, Mar. 2021, doi: 10.1109/TBDATA.2020.3048644.
[17] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques', Morgan Kaufmann, Fourth Edition, 2016." 2016.
[18] "Most Recent National Asthma Data | CDC," Apr. 07, 2021. https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm (accessed Sep. 16, 2021).
[19] "National Diabetes Statistics Report 2020. Estimates of diabetes and its burden in the United States.," p. 32, 2020.
[20] "Heart Disease Facts | CDC," Apr. 07, 2021. https://www.cdc.gov/heartdisease/facts.htm (accessed Sep. 16, 2021).
[21] "Hypertension Prevalence and Control Among Adults: United States, 2015–2016: NCHS Data Brief No. 289, October 2017," Sep. 16, 2021.
[22] "U.S. Census Bureau QuickFacts: United States." https://www.census.gov/quickfacts/fact/table/US/AGE295219 (accessed Sep. 16, 2021).
[23] "COVID-19 and the heart: What have we learned?," *Harvard Health*, Jan. 08, 2021. https://www.health.harvard.edu/blog/covid-19-and-the-heart-what-have-we-learned-2021010621603 (accessed Jan. 09, 2022).
[24] G. M. Vahey *et al.*, "Risk factors for hospitalization among persons with COVID-19—Colorado," *PLoS ONE*, vol. 16, no. 9, p. e0256917, Sep. 2021, doi: 10.1371/journal.pone.0256917.

6 Supplemental Figures

```
Sigmoid Node 0
    Inputs    Weights
    Threshold    -2.8642559324600314
    Node 2    1.5946146633516534
    Node 3    3.527503074183863
    Node 4    0.9909031723493787
Sigmoid Node 1
    Inputs    Weights
    Threshold    2.8642559324600323
    Node 2    -1.5946146633516531
    Node 3    -3.5275030741838638
    Node 4    -0.9909031723493786
Sigmoid Node 2
    Inputs    Weights
    Threshold    -13.094164748671725
    Attrib ASTHMA=Yes    -0.6212144376072047
    Attrib HRT_DIS=Yes    -4.271070318570962
    Attrib DIABETES=Yes    -3.288334229321039
    Attrib HYPERTEN=Yes    -3.624855217539419
    Attrib ADV_AGE=Yes    -5.255370584350011
Sigmoid Node 3
    Inputs    Weights
    Threshold    -4.444264876641236
    Attrib ASTHMA=Yes    -2.5544984737877097
    Attrib HRT_DIS=Yes    -3.6606207697222493
    Attrib DIABETES=Yes    -3.1031640917531895
    Attrib HYPERTEN=Yes    -4.1253385066905315
    Attrib ADV_AGE=Yes    -4.663454633981677
Sigmoid Node 4
    Inputs    Weights
    Threshold    -12.781133705247957
    Attrib ASTHMA=Yes    -2.6124478050913127
    Attrib HRT_DIS=Yes    -3.7596690848630274
    Attrib DIABETES=Yes    -1.8687255514640533
    Attrib HYPERTEN=Yes    -5.362747926014692
    Attrib ADV_AGE=Yes    -7.042599921992943
Class No
    Input
    Node 0
Class Yes
    Input
    Node 1
```

Figure 4. Node structure of multilayer perceptron neural network.