IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Evaluating the Effectiveness of a Symbolic Metamodel in Predicting Heart Disease & Diabetes

Sneha Sundar

Lynbrook High School, 1280 Johnson Ave, San Jose, CA, 95129, USA; snehasundarr@gmail.com

ABSTRACT: Heart disease and diabetes are brought on by high blood sugar levels and can damage the blood vessels and nerves that control the heart. Early treatment can drastically improve outcomes for both heart disease and diabetes patients. This study aimed to determine whether symbolic metamodeling could be used to predict heart disease and diabetes early. The symbolic metamodel developed by Alaa and van der Schaar was evaluated on two datasets consisting of predictive factors for diabetes and heart disease, respectively, and a patient's ranking within those factors. The metamodel was fitted to both a standard logistic regression model and a more complex XGBoost model. The metamodel produced predictive scores between 0.7 and 0.8, indicating the potential for usage as a predictive device. The results from this study do demonstrate a need for increased research into using symbolic metamodeling as a methodology for disease prediction. Increased research could improve the breadth of diseases that could benefit from predictive metamodeling.

KEYWORDS: Analysis; Machine Learning; Disease Detection and Diagnosis.

Introduction

Computer Science, by nature, is an interdisciplinary field with applications across numerous domains. This evolving field is not only responsible for immense advancements in technology but also in non-computing fields such as medicine, law, and the arts. These advancements have furthered society beyond what was possible and created a positive, lasting impact. Yet, we don't always know how these advancements work, so they are labeled as "black-boxes." A black-box is a system in which we know the input and the output, but we are unable to map the process of how we obtain the said output. While black-box models are beneficial in checking code and proving calculations, getting a greater understanding of the black-box model can help domain experts better understand their specific non-computing fields.[1]

We demystify these black-box models using the process of symbolic metamodeling. A symbolic metamodel represents a black-box model using mathematical functions. Neuro-symbolic artificial intelligence focuses on bringing together the neural and symbolic traditions within AI.[2] Using neuro-symbolic artificial intelligence to gain a better understanding of our black-box models allows us to capture both domain and semantic knowledge leading to a more transparent and analytical representation of our black-box model.[1]

The Symbolic Metamodel:

Metamodeling techniques have been established across various domains, from statistics to engineering, as surrogates for expensive simulations.[3] These models can describe other models as an instance and make it possible to assign properties to classes within the model itself.[4] In a symbolic system, the output is explicit and understood by a human.[2] A symbolic metamodel framework can be used to express black-box models in terms of mathematical equations that can be interpreted easily by experts in both computing and non-computing fields. The symbolic metamodel aims to expand interpretation beyond just one mode determining whether the model is nonlinear or whether it has statistical interactions between features. A symbolic metamodel can be defined as a model that approximates f(x) - our black box function - for all x in X.1 Symbolic metamodels can be used to predict the prognosis of various ailments including breast cancer and COVID-19 using synthetic experiments.[5] These experiments use a synthetically generated model to compare predicted and proven outcomes.[6]

Heart Disease:

In 1933, heart disease was identified as one of the leading causes of death in the United States, affecting all genders, races, and ethnicities.[7,8] The death rate due to heart disease has steadily increased from 21% to 39% from 1933 to 1960.[9] Although new cardiovascular treatments have shown a decrease in mortality due to this disease, in recent decades, mortality overall has stopped decreasing, and mortality rates have continued to increase.[9,10] An accurate heart disease risk prediction could avoid potential life threats, while an incorrect prediction could prove fatal. Many factors contribute to a diagnosis of heart disease: irregular heartbeat, weight gain, and sleep issues. This makes it challenging to identify whether increases in these specific metrics lead to a diagnosis of heart disease or a condition with similar biomarkers.[11]

Diabetes:

Diabetes is cited as one of many causes of heart disease - both diseases having origins tied to high blood glucose and increased weight gain.[7] Diabetes, similar to heart disease, is a significant cause of mortality and morbidity in the United States. In recent decades, there has been an increase in the prevalence of diabetes from 4.9% in 1990 to 6.5% in 1998.[12] Outcomes of diabetes include damage to small and large blood

DOI: 10.36838/v5i2.21

vessels, kidney damage, and problems with various nerves.[13] There is a need to utilize multiple indicators of diabetes to confirm the diagnosis. Predicting diabetes early on offers the chance to start potentially life-saving treatment. Diagnosis of diabetes is considered a challenging problem in quantitative research because parameters like A1c, fructosamine, white blood cell count, fibrinogen, and hematological indices were ineffective due to some limitations.[14] In this paper, we focus our efforts on heart disease and diabetes disease processes.

Related Work in Heart Disease & Diabetes Prediction:

Current machine learning models in diabetes prediction use data mining techniques to showcase various significant attributes and the relationship between these attributes. Alam and Iqbal found that an artificial neural network (ANN)-based technique provided the best accuracy of 75.7% through RF (random forest) and K-means clustering techniques.[14] These experiments show the use of an ANN in disease prediction, which is a key stepping stone in symbolic metamodel prediction for the same disease processes. Delen *et al*.'s work in the use of logistic regression in predicting breast cancer can be further extended to diabetes and heart disease prediction as well.[15] Pattekari and Parveen developed an intelligent system using data mining techniques to predict myocardial infarction.[16] Palaniappan *et al*. combined decision trees, an artificial neural network, and Naïve Bayes to predict heart disease.[17]

Transparency in predicting many disease processes, including heart disease and diabetes, is essential; symbolic metamodeling can provide this transparency in new ways that other machine-learning models cannot. The symbolic metamodel can produce predictive scores and symbolic equations that clinicians can use and understand. This transparency allows for increased trust in the metamodel and a deeper understanding/insight into the contributing factors of a patient's risk. Alaa and van der Schaar created and used this symbolic metamodel to predict breast cancer prognosis.[1] Extending their work to heart disease and diabetes, the same fundamental principles of transparency and trustworthiness prevail. The predictive scores the model produces allow us to compare the model based on both XGBoost and logistic regression.

Methods

Metamodeling:

A metamodel can also be defined as a model made up of models where the outputs from the submodels are combined in a particular sequence to form a model of the entire system.[18] Metamodeling is widely applicable when a model needs to play more than one role. For example, a particular concept could be portrayed as either a class or an individual idea. Metamodels can be applied to any field that contains semantic data that can be exchanged or stored. Semantics can formally be defined as a tool to translate concepts into their domain language. Semantics translate knowledge from formal techniques such as mathematics to a natural language such as English.[4]

The model we are evaluating in our research is a symbolic metamodel used to perform computations using Meijer-G functions. Three datasets evaluate the model's ability to explain predictions regarding instance-wise feature importance with an altered 9-dimensional feature space. The three datasets are fitted with a 2-layer neural network to predict the labels based on the dimensional feature spaces and then equipped with a symbolic metamodel for the trained network using algorithms that optimize symbolic metamodels via gradient descent. The model predicts both an AUC-ROC score that can be used to compare model accuracy with other predictive models not using symbolic regression. Alaa and Van der Schaar tested the model's accuracy versus a predicted model on the prognosis of breast cancer using this predictive score.[1]

Data:

This experiment employed two data sets containing predictive indicators for heart disease and diabetes.

The first dataset contains 253,680 survey responses derived from a cleaned Behavioral Risk Factor Surveillance System (BRFSS) in 2015 to be used as a binary classification of heart disease. The BRFSS is a health survey that looks at behavioral factors, chronic health conditions, and the use of preventative services. The dataset contains one binary target variable determining an instance of heart disease or heart attack and nine feature variables that are either binary or ordinal. The feature variables - high blood pressure, high cholesterol, cholesterol check, body mass index, the incidence of smoking, prior stroke, diabetes, and physical activity - have been previously identified as indicators of heart disease.[19]

The second dataset contains data from 769 females over the age of 21 of Pima Indian heritage. The dataset includes one binary target variable determining the presence of a form of diabetes and eight feature variables - number of times pregnant, plasma glucose concentration calculated by 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, age, and insulin level - all identified as indicators of diabetes.[20] For experimentation, the datasets were divided into training and testing groups with a ratio of 1:2.[1]

Heart disease and Diabetes Prediction:

In the process of computing the predictive scores, we had to assign each predictive indicator a feature name that the model could understand. The original symbolic metamodel had a 10-dimensional feature space, while our metamodel had to be altered to accommodate a 9-dimensional feature space. Once the feature names had been imputed, the model classifier was determined. In our experimental setup, we wanted to test the model's ability to imitate the standard logistic regression model as well as the XGBoost model. The next parameter to alter was the maximum number of iterations used to train the model. The original metamodel had settled on ten iterations, so we decided to test from 0 to 15 iterations in our experiment.[1] Through these iterations, we decided to leave the learning rate as a constant to evaluate better how the iterations affected the model's performance.

Results and Discussion

Three types of data - in the form of roc auc scores - were collected from the symbolic metamodel for both heart disease and diabetes. The first set of data collected was the performance of the fitted XGBoost model. The second set of data collected was the standard logistic regression. The third data

DOI: 10.36838/v5i2.21

set was the metamodel's evaluation for both the XGBoost model and the standard regression model.

The roc auc (roc → receiver operating characteristic, auc → area under the curve) scores provide a uniform metric to evaluate each type of prediction method within each set of data which helps us gauge the accuracy of each type of prediction (Tables 1,2,3). These scores can be evaluated against each other to see how effectively a particular model performs compared to another model. We consider roc auc scores between 0.9-1.0 to be excellent, 0.8-0.9 to be good, 0.7-0.8 to be fair, and 0.6-0.7 to be poor. Any score below 0.5 would be considered a failed score.[21]

Figure 1 shows these predictive scores when ten iterations were used for all tested models. The symbolic metamodel was fitted to the initial fitted XGBoost model and standard logistic regression model with a learning rate of 0.01 to see how close the roc auc scores were. The highest roc auc score was for the fitted XGBoost model, while the standard logistic regression was about 0.02 lower. We hypothesize that the lower predictive score could be due to the lower complexity of the model. With a more complex model, such as XGBoost, we could see higher roc auc values than the standard logistic regression. Yet, there is no concrete result when we compare the symbolic metamodel fitted to the XGBoost and the standard logistic regression. For the heart disease data, the metamodel performs better when fitted to the logistic regression, and for diabetes, the metamodel performs better when fitted to the XGBoost model. To analyze these results further, we tested various iterations of each fitted model.

Table 1: Summary of the predictive roc auc values obtained from running our different metamodels against the heart disease and diabetes datasets.

	heart disease	diabetes
fitted XGBoost model	0.781520779233839	0.8103543743078627
standard logistic regression	0.7790508318619848	0.7965808416389812
symbolic metamodel (based on xgboost) for 10 iterations	0.7334147622645342	0.7787236987818383
symbolic metamodel (based on logistic regression) for 10 iterations	0.735112941221171	0.7661960132890365

Table 2: The predictive roc auc values obtained from running our XGBoost trained metamodels against the heart disease and diabetes datasets.

	heart disease	diabetes
symbolic metamodel for 0 iterations	0.7349389204996443	0.7802464008859357
symbolic metamodel for 1 iteration	0.7349550555174734	0.7794158361018827
symbolic metamodel for 2 iterations	0.7349817466104752	0.7771317829457364
symbolic metamodel for 3 iterations	0.7349920099131467	0.7758859357696566
symbolic metamodel for 4 iterations	0.7349671893111233	0.7778239202657806
symbolic metamodel for 5 iterations	0.7349988900669188	0.7793466223698782
symbolic metamodel for 7 iterations	0.7350388860435739	0.7787236987818383
symbolic metamodel for 10 iterations	0.735112941221171	0.7771317829457365
symbolic metamodel for 12 iterations	0.7351826861369488	0.7772702104097452
symbolic metamodel for 13 iterations	0.7351986259731107	0.7783776301218162
symbolic metamodel for 15 iterations	0.7351927217276752	0.7779623477297896
symbolic metamodel for 50 iterations	0.7298715389140197	0.7776854928017719

The results from testing these various iterations, as shown in Figure 2, show that for the iterations of 0, 1, 5, the model performs better than the metamodel fitted to the standard logistic regression model shown in Figure 1. We hypothesize that the closer fit could be due to the number of iterations it takes the metamodel to fit XGBoost for this particular dataset. Looking at Figure 3, we can see that the metamodel for iterations 0, 1, 5 has a higher roc auc score when fitted to XGBoost than when it is fitted to the standard logistic regression. This pattern follows for both the heart disease and diabetes datasets. We hypothesize that this could be due to the complexity of the XGBoost model, which in turn makes the metamodel more complex and accurate, but we would need to conduct more tests on the nature of these results to confirm.

Table 3: The predictive roc auc values obtained from running our metamodel trained on standard logistic regression against the heart disease and diabetes datasets.

	heart disease	diabetes
logistic regression model for 0 iterations	0.7345276076661114	0.7682724252491694
logistic regression model for 1 iterations	0.7345149859182933	0.7685492801771873
logistic regression model for 2 iterations	0.7345077316663258	0.7685492801771873
logistic regression model for 3 iterations	0.7345168726744105	0.7681339977851607
logistic regression model for 4 iterations	0.7345152461605164	0.7684800664451827
logistic regression model for five iterations	0.734551013201047	0.7678571428571428
logistic regression model for 7 iterations	0.7345646271223404	0.7676495016611294
logistic regression model for 9 iterations	0.7346400648367475	0.7673726467331118
logistic regression model for 10 iterations	0.7347149170061527	0.7661960132890365
logistic regression model for 12 iterations	0.734829228402628	0.7661267995570321
logistic regression model for 13 iterations	0.7347825799841461	0.7666112956810631
logistic regression model for 50 iterations	0.7349129938681728	0.7656423034330011

To find how the number of iterations affects the auc roc score that the metamodel produces, we tested values from 0 to 50 and plotted them on the graph in Figures 1 and 2. The results from these experiments show that the heart disease and diabetes datasets' roc auc scores are the inverse of one another for both standard logistic regression and the fitted XGBoost model. Heart disease, for instance, is directly proportional to the number of iterations increasing as the number of iterations increases for both XGBoost and standard logistic regression. However, diabetes shows the exact opposite: decreasing as the number of iterations increases, possibly due to the specific metamodel used. These results show intricate details of the metamodel in its attempt to predict prognosis for both heart disease and diabetes but do not provide conclusive evidence. While similar to the original metamodel's ability to predict breast cancer prognosis, the metamodel does reveal more than the original XGBoost model; we cannot definitively say to what extent the model accurately predicts heart disease and diabetes. Through the analysis of the roc auc scores, there is some evidence that indicates the model has the potential to predict heart disease and diabetes well. So, our model's results demonstrate fair roc auc scores.[21]

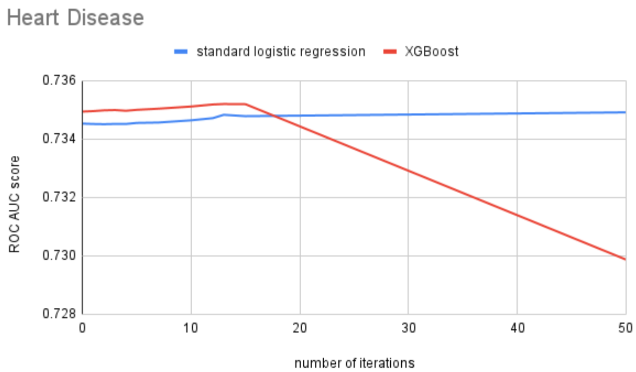

Figure 1: Heart Disease.

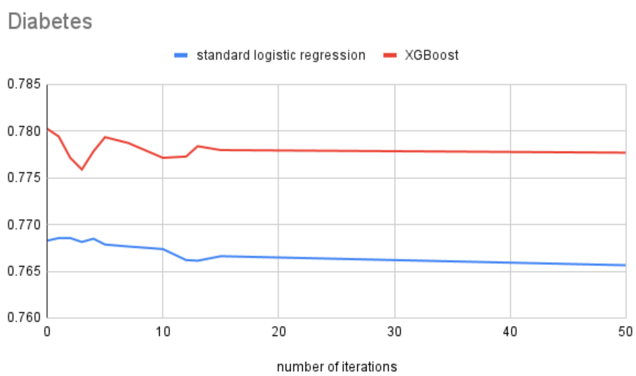

Figure 2: Diabetes.

Further research on this topic could utilize the same XGBoost model as well and improve upon the training of the metamodel on this XGBoost model. Additional ideas for research would be to see why the heart disease and diabetes datasets behave so differently when the number of iterations is changed. Other models outside of XGBoost can be evaluated to see if the metamodel will produce higher roc auc scores which can then be translated to increased model accuracy.

Conclusion

The ability to harness the power of metamodeling and symbolic artificial intelligence to predict the prognosis of cardiovascular conditions, including heart disease and diabetes, can improve patient outcomes by making early treatment possible. Using predictive scores, such as ROC AUC scores, has made it possible to evaluate the metamodel's performance quantitatively.

We assigned the predictive factors from each dataset as feature names to our model. We then were able to set up models for testing: a metamodel trained on standard logistic regression and another trained on the XGBoost model.

The results from these experiments indicate the metamodel's potential to be used as a predictive method for both heart disease and diabetes. The range of predictive roc auc scores fell between 0.7 and 0.8 for both heart disease and diabetes in both the standard logistic regression model and the XGBoost model, suggesting that the symbolic metamodel has room for improvement in predicting both heart disease and diabetes.

Our experiments suggest a slight increase in roc auc scores with the XGBoost model than with the standard logistic regression model. Future research can be conducted to improve the metamodel's training on the XGBoost model with an emphasis on training the model specifically for heart disease and diabetes. The future of disease prediction with artificial intelligence can benefit greatly from using symbolic metamodeling for various diseases.

Acknowledgments

I would foremost like to thank my mentor, Dr. Catherine Wah, for supporting me through this project and offering the most helpful and insightful feedback. I would also like to thank Tyler Moulton for his exceptional input on my writing style and help throughout this publication journey. Next, I would like to thank my parents for fostering my love for research and discovery throughout my life.

References

1. Alaa, A. M.; van der Schaar, M. Demystifying black-box models with symbolic metamodels. https://papers.nips.cc/paper/2019/hash/567b8f5f423af15818a068235807edc0-Abstract.html (accessed Jul 18, 2022).
2. Susskind, Z.; Arden, B.; John, L. K.; Stockton, P.; John, E. B. Neuro-symbolic ai: An emerging class of AI workloads and their Characterization. https://arxiv.org/abs/2109.06133 (accessed Jul 18, 2022).
3. Wang, G. G.; Shan, S. Review of metamodeling techniques in support of Engineering Design Optimization. https://asmedigitalcollection.asme.org/mechanicaldesign/article-abstract/129/4/370/466824/Review-of-Metamodeling-Techniques-in-Support-of?redirectedFrom=fulltext (accessed Jul 18, 2022).
4. LLC, D. A. W. O.; Allemang, D.; LLC, W. O.; Working Ontologist LLCSearch about this author; Institute, J. H. R. P.; Hendler, J.; Institute, R. P.; Rensselaer Polytechnic InstituteSearch about this author; INRIA, F. G.; Gandon, F.; Inria; author, I. N. R. I. A. S. about this; Metrics, O. M. V. A. Semantic web for the working ontologist: effective modeling for Linked Data, RDFS, and owl. https://dl.acm.org/doi/10.1145/3382097.3382114 (accessed Jul 18, 2022).
5. Jana, A.; Minacapelli, C. D.; Rustgi, V.; Metaxas, D. Global and local interpretation of black-box machine learning models to determine prognostic factors from early COVID-19 data. https://www.researchwithrutgers.com/en/publications/global-and-local-interpretation-of-black-box-machine-learning-mod (accessed Jul 18, 2022).
6. Computing, Y. X. S. of; Xie, Y.; Computing, S. of; Computing, Z. X. S. of; Xu, Z.; Mohan S Kankanhalli School of Computing; Kankanhalli, M. S.; Kuldeep S. Meel School of Computing; Meel, K. S.; Computing, H. S. S. of; Soh, H.; Metrics, O. M. V. A. Embedding symbolic knowledge into deep networks: Proceedings of the 33rd International Conference on Neural Information Processing Systems. https://dl.acm.org/doi/ abs/10.5555/3454287.3454668 (accessed Jul 18, 2022).
7. Multiple cause of death, 1999-2020 request. https://wonder.cdc.gov/mcd-icd10.html (accessed Jul 20, 2022).
8. Heart disease and stroke statistics—2021 update - circulation. https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000950 (accessed Jul 20, 2022).
9. Leading causes of death, 1900-1998 - centers for disease control and ... https://www.cdc.gov/nchs/data/dvs/lead1900_98.pdf (accessed Jul 22, 2022).
10. Ritchey, M. D.; Wall, H. K.; George, M. G.; Wright, J. S. US trends in premature heart disease mortality over the past 50 years: Where do we go from here? https://www.sciencedirect.com/science/article/pii/S1050173819301343 (accessed Jul 18, 2022).

DOI: 10.36838/v5i2.21

11. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of heart disease using a combination of machine learning and Deep Learning. https://www.hindawi.com/journals/cin/2021/8387680/ (accessed Jul 18, 2022).
12. Mokdad, A. H.; Ford, E. S.; Bowman, B. A.; Nelson, D. E.; Engelgau, M. M.; Vinicor, F.; Marks, J. S. Diabetes trends in the U.S.: 1990-1998. https://diabetesjournals.org/care/ article/23/9/1278/21466/Diabetes-trends-in-the-U-S-1990-1998 (accessed Jul 20, 2022).
13. Trikkalinou, A.; Papazafiropoulou, A. K.; Melidonis, A. Type 2 diabetes and quality of life. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5394731/ (accessed Jul 18, 2022).
14. Alam, T. M.; Iqbal, M. A.; Ali, Y.; Wahab, A.; Ijaz, S.; Baig, T. I.; Hussain, A.; Malik, M. A.; Raza, M. M.; Ibrar, S.; Abbas, Z. A model for early prediction of diabetes. https://www.sciencedirect.com/science/article/pii/S2352914819300176 (accessed Jul 18,2022).
15. Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer surviva bility: A comparison of three data mining methods: Artificial Intelligence in Medicine: Vol 34, no 2. https://dl.acm.org/doi/10.1016/j.artmed.2004.07.002 (accessed Jul 18, 2022).
16. Dulhare, U. N. Prediction system for heart disease using naive Bayes and particle ... https://www.researchgate.net/publication/326717088_Prediction_system_for_heart_disease_using_Naive_Bayes_and_particle_swarm_optimization (accessed Jul 22, 2022).
17. Palaniappan, S.; Awang, R. Intelligent heart disease prediction system using data mining techniques. https://ieeexplore.ieee.org/document/4493524 (accessed Jul 18, 2022).
18. Nisbet, R.; Elder, J. F.; Miner, G. Model evaluation and enhancement.https://www.researchgate.net/publication/284075239_Model_Evaluation_and_Enhancement (accessed Jul 22, 2022).
19. Teboul, A. Heart disease health indicators dataset. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset (accessed Jul 22, 2022).
20. Akturk, M. Diabetes dataset. https://www.kaggle.com/datasets/mathchi/diabetes-data-set (accessed Jul 22, 2022).
21. TEl Khouli Barker PB; Habba MR; Jacobs MA; Bluemke DA; R. H.; Katarzyna, M. J.; Barker, P. B.; Habba, M. R.; Jacobs, M. A.; Bluemke, D. A. Relationship of temporal resolution to diagnostic performance for dynamic contrast-enhanced MRI of the breast. https://pubmed.ncbi.nlm.nih.gov/19856413/ (accessed Jul 18, 2022).

Authors

Sneha Sundar is a rising senior at Lynbrook High School in San Jose, CA. Her interests include computer science, medicine, and mathematics. Sneha's goal is to bring an interdisciplinary perspective to computer science and discover more about the evolving field in college. She hopes to major in computer science with a minor in statistics or public health.