A Probabilistic Graphical Model for MACE Prediction

CSDS 491: Probabilistic Graphical Models

Sakin Kirti (BA, CWRU 2023) and Joshua Freeze (MS, CWRU 2024)
Dr. Michael Lewicki

Introduction

In the United States, heart disease and heart-related diseases are the leading cause of death for men and women of any racial or ethnic group, encompassing one in every five deaths. Additionally, heart disease costs the United States approximately $229 billion every year. Of heart diseases, Coronary Artery Disease is the most common type, taking the lives of 50% of individuals who died from heart disease. Heart disease is strongly tied to several other medical conditions including diabetes and obesity [1].

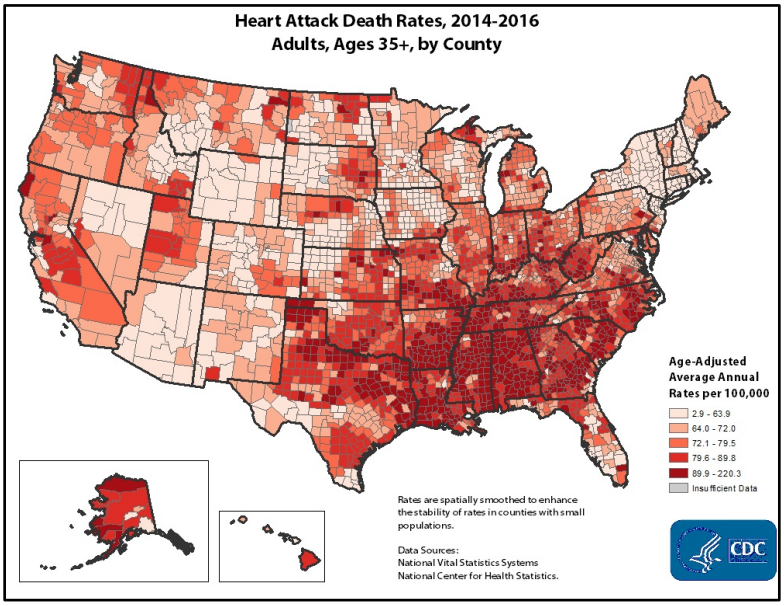


Figure 1. *A map showing the average rates of heart attack death from 2014-2016 [2].*

To encompass the many types of heart disease as well as the differing outcomes of heart disease, researchers have assembled a statistic known as Major Adverse Cardiovascular Events (MACE) which encompasses the vast majority of cardiovascular events including myocardial infarction, stroke, cardiovascular death, and more. This metric is used very regularly during randomized controlled trials to better understand the risks and efficacy of cardiovascular drugs [2].

Importantly, because of the multitude of studies that rely on MACE or use it as a method to better understand results, there are methods to predict MACE for use in a clinical environment. These methods, however, utilize a formulaic approach rather than a probabilistic approach, lacking the ability to learn from past experience or produce intelligent risk scores [3].

Motivation

Due to clinical and research reliance on MACE, there is a need for more intelligent methods to calculate risk scores. Using artificial intelligence methods to do so would produce the ability to learn from new examples and increased accuracy compared to formulas, especially when compared to the dated formula methods. The need for such systems, however, may not be only for the clinical and research departments. Instead, the release of Apple's in-depth heart sensors and rumored blood-glucose monitoring suggests that such models could also be applied to individuals, providing early detection for MACE events. In particular, our model aims to improve MACE scores with the addition of more complex calcium information from CT Calcium score (CTCS) images which should be able to supplement the existing patient information that is normally collected. Doing so is relatviely easy way to get additional diagnostic information and if it produces useful results from our model it could result in significant benefits to the patient population.

Through this course project, we aimed to produce a Bayesian network which could be used to predict the time at which a MACE event might occur as well as produce risk scores for MACE. To do so, we built a simple logistic regression model, a deep neural network, and a Bayesian network and built survival analysis curves based on the outputs of each. The outputs from each model represent risk scores while survival analysis is a form of time-to-event analysis and therefore predicts the time at which an event may occur.

Approach and Rationale

Due to the complexity of building a Bayesian network with the potential to predict the time of MACE events, we first wanted to build simpler models to understand the dataset and how this problem could best be solved.

Before building any models, the students spent time cleaning the data. In data cleaning, missing values were checked for, performing imputation if the proportion of missing values was low enough or removing observations if not. Additionally, due to the low proportion of MACE that occurred in the dataset compared to the population size, SMOTE was performed to bootstrap the underrepresented class, resulting in a training dataset in which 50% of no MACE events occurred and 50% of MACE events occurred.

Data Availability

Data was sourced from the Biomedical Imaging Laboratory. Joshua regularly works on heart calcification data derived from CT scans along with general heart health features such as age, smoking status, diabetes status, etc. With permission from the Biomedical Imaging Laboratory, a dataset of 85 features was obtained and used for all model generation.

Logistic Regression Model

We first worked on building a logistic regression model to predict whether an individual would have a MACE event. Among the inputs to this model was the time point of the individual, which was then used to perform survival analysis.

A logistic regression model was used for this first analysis because it was incredibly simple to put together using scikit-learn, allowing us to discern any correlations in the data. Paired with a binary classification problem (MACE event or no MACE event), logistic regression was a good first model to use. A logistic regression model works by learning the coefficients of the inputted data and mapping the inputs to a 0 or 1 output. In doing so, it tries to predict the outcome of certain events. While logistic regression models are not typically associated with being the most intricate models, they provide data scientists with an observable metric for whether there is a correlation in a dataset.

After developing the logistic regression model, we were then able to use the trained model to predict MACE events and build survival analysis curves.

Deep Learning Model

After building a logistic regression model, we then built a deep learning model to compare with the Bayesian network. Deep learning models excel at classification tasks by using a combination of weights and biases, which, after being optimized are used as coefficients in individual functions to classify each observation as having a MACE event or not.

A deep neural network was good for this due to its ability to classify inputs extremely well. However, a problem that comes up often in training deep neural networks is overfitting. To reduce overfitting, we implemented early stopping, a procedure that stops training after the difference between the validation loss and training loss is larger than a user-specified threshold.

In training the network, we developed a simple fully connected network with PyTorch, which took each feature as input and produced two outputs. Each output was assigned a value of 0 or 1, and whichever output contained the larger logit probability was selected. A schematic of the fully connected network is shown below.

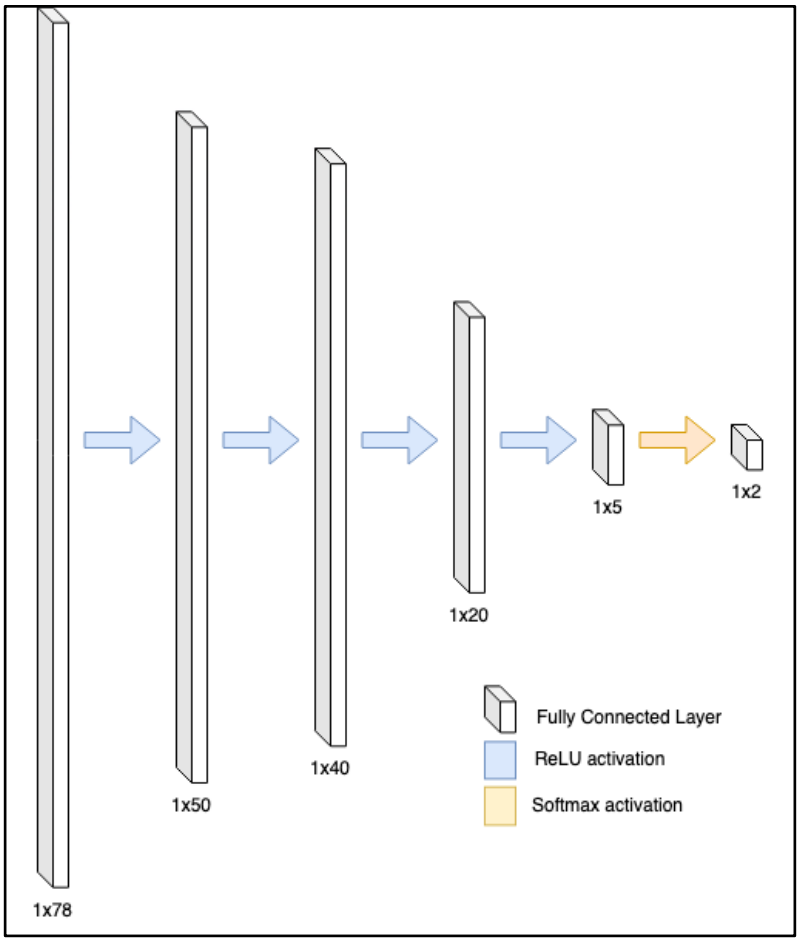


Figure 2. *A schematic of the developed fully connected network used for predicting MACE events. From left to right, vector sizes are 78, 50, 40, 20, 5, and 2.*

After training, the model weights were stored for later use in developing survival analysis plots.

Bayesian network

Bayesian networks are a class of statistical models that utilize Bayes' theorem to make probabilistic predictions about unknown quantities. The basic idea behind Bayesian modeling is to combine prior knowledge with new data to update the beliefs about a particular phenomenon. In contrast to classical or frequentist statistics, Bayesian networks assign probabilities to parameters rather than treating them as fixed values. This means that uncertainty is explicitly quantified, and new data can be used to revise the probability of different parameter values [4].

In the case of MACE events, a Bayesian network can be used to quantify the uncertainty associated with various events, allowing clinicians and researchers to make predictions about patients while also discussing the certainty of such a prediction with patients and their families.

Due to the quantification of uncertainty with a Bayesian network, this would allow us to make predictions, while also generating confidence about the predictions made.

A Bayesian network was built using the features available in the dataset. The model, built with Pytorch and DGL, was then trained using three graph convolution layers. The final node features were then combined to create a single graph level classification. Relationships between nodes were set based on literature and the relationships between calcium features (see Appendix 1.1 for additional details). The final graphical structure is shown in Figure 4.

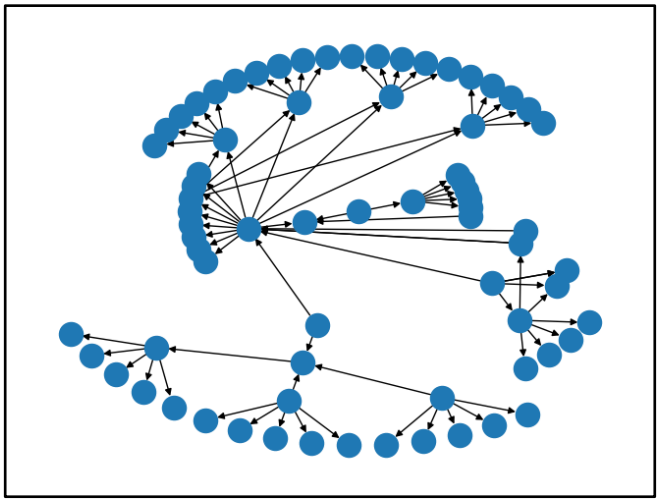
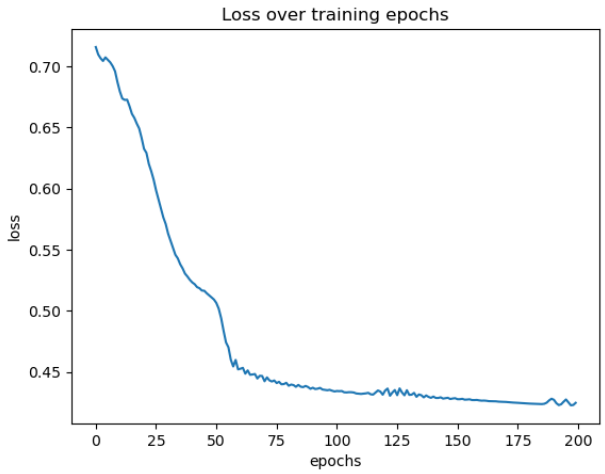


Figure 3. *Figure showing the various nodes and the relationships used to build this Bayesian network.*

After fitting the Bayesian network to the training dataset, we were then able to build a survival analysis curve from the model's predictions, allowing us to compare the survivability of the population as time continues.

Results



To evaluate the models on an even playing ground, we utilized the same training and testing set with each model. Additionally, each testing set was not bootstrapped, resulting in a more accurate proportion of events relative to the true population.

In training the neural network, an important training metric is the model's loss. While training, we want to minimize the model's loss, which reflects the difference between the true result and the model's predicted result. The chart to the left shows a chart demonstrating the model's loss during training. As can be seen in the chart, the model's first training iteration resulted in a loss near 0.7, which was then optimized during training at epoch 200 to 0.41. While training could have continued at this point, early stopping was used to reduce overfitting. Early stopping was done by comparing the loss of the validation set with the loss of the training set. If the loss was separated by a user-defined threshold of 0.05, training was stopped and the model was saved. This allowed us to create a model which adequately fit the training data, while also generalizing well to the entire population.

Similar to the neural network, training the Bayesian network also requires optimization. Therefore, a similar loss curve was generated for the construction of this model. In training the Bayesian network, early stopping was not used, but the model was trained several times with differing numbers of epochs to generate a model that fit the testing set well.

Classification

Logistic regression and deep learning are optimized for classification problems. It, therefore, makes sense that the predictions produced by these two formats are close to 0 and 1 with very little data in between. The goal of these two methods is in fact to minimize the number of non-0s and 1s. Conversely, with the Bayesian network, the point is to produce points in between to input confidence into the predictions. Below, figure 4 shows the output predictions of each model. As can be seen, the logistic regression model and deep learning model output purely 0s and 1s while the Bayesian network produces outputs that reflect the probability of an event occurring. These values, as one might expect, are between 0 and 1 (or could be the absolute 0 or 1, though this is unlikely.

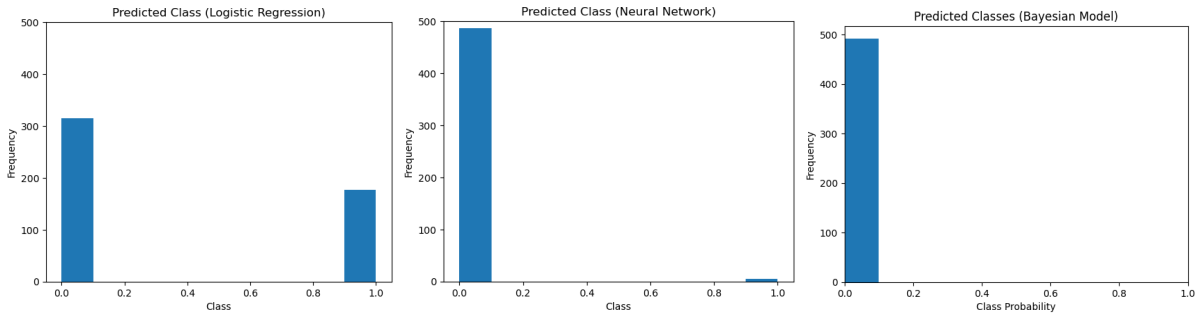


Figure 4. *Plots demonstrating the difference in classification ability between the logistic regression, deep learning, and Bayesian networks.*

MACE Prediction

Each model predicted whether or not a MACE would occur. The Bayesian network produced the highest accuracy, followed by the deep neural network, and the logistic regression model. Accuracy for the Bayesian network was calculated as follows: if the probability of an event occurring was greater than 50%, then it was classified as an event, and otherwise no event. The models produced accuracy scores of 93.7%, 87.4%, and 66.1% respectively (Table 1).

	Logistic Regression	**Deep Neural Network**	**Bayesian network**
Accuracy	0.661	0.874	0.905
AUC	0.685	0.907	0.720

Table 1. *Table showing accuracy and AUC of each model. The metrics give a general idea of how well each model was able to fit the data.*

This suggests that the Bayesian network was better able to fit the data than the logistic regression model but couldn't quite outperform the traditional deep neural network. Additionally, the Receiver Operating Characteristic was calculated and plotted (Figure 4), demonstrating how well each model could fit the data. When interpreting an ROC plot, a model which produces a high true positive rate paired with a low false positive rate is a model which fits the data well. These data demonstrate that deep neural network was able to fit the data best.

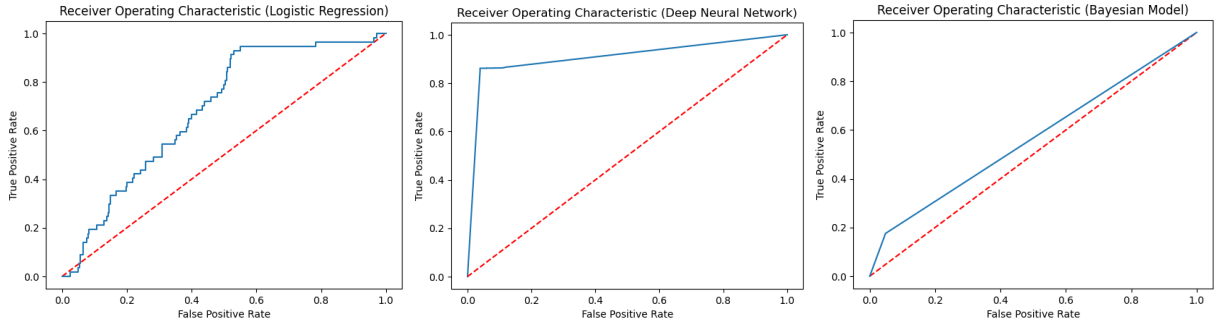


Figure 5. *Receiver Operating Characteristic curves for each model. left: logistic regression, middle: deep learning, right: Bayesian network*

In addition to creating accurate predictions, one of the goals that was set for this project was to produce risk scores rather than just predicting MACE. The probabilities outputted by the Bayesian network inherently represent risk scores as they directly correlate with the likelihood of an event. In this sense, the goal of the project was reached. Similarly, the outputs of the logistic regression and neural network also represent risk scores, however, they are not good ones. Both of these models optimze their parameters to classify outputs rather than represent the likelihood of an event. Therefore, these outputs should not be used as risk scores.

Survival Analysis

Survival analysis was performed to understand the time until a MACE event was likely for a certain individual. Survival analysis is a type of time-to-event analysis where the probability of an individual being alive is shown on the y-axis while time is shown on the y-axis. Interpreting such plots allows clinicians to confidently tell their patients their probability of being alive in *x* number of months.

Survival analysis plots were produced for each model as shown in Figure 6.

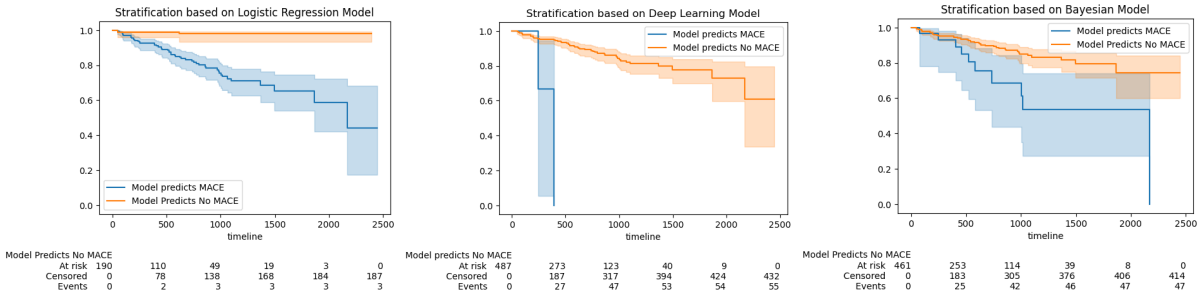


Figure 6. *Survival Analysis plots for each of the three models built. Left: logistic regression, middle: deep learning, right: Bayesian network*

Even with the simple logistic regression model, it is clear that each model is able to perform time-to-event analysis with the MACE and no MACE groups separating clearly. This separation clearly demonstrates a correlation between the features that were examined in the model and the outputs. Based on Figure 6, it is clear that although the logistic regression model is not the most sophisticated model, for performing time-to-event analysis, it does a fairly reasonable job of separating the MACE and no MACE event groups, with the Bayesian and deep neural network improving on this performance. On the other hand, notice thate the deep learning model's classification power is its greatest strength when it comes to risk prediction. The curve for MACE predicted drops sharply, demonstrating that the model easily picks out the highest risk patients who are likely to have a MACE event soon. The Bayesian network produces a middling output here, where the groups are very well separated during the first few years but later the groups begin to converge again due to a misclassified case or two. This leads to the

models confidence intervals overlapping at around 2000 days which is a clear indicator of poor stratification at this time point.

Overall, it is clear that the deep neural network had the clear upper hand because of its ability to not only fit the data in a classification sense but also to produce confidence in each of its predictions.

Here, the second goal of the project was met as the Bayesian network was able to produce accurate predictions for the time at which a MACE was likely to occur but it was not able to outcompete the traditional deep learning approaches

Conclusion

Altogether, this project resulted in three models - a logistic regression, deep learning, and a Bayesian network. The logistic regression model was a first pass which allowed us to perform more in-depth analysis. Following this, a deep learning model was built as a step above the logistic regression model to compare the Bayesian network. The deep learning model was built with a simple architecture starting with the number of inputs (features) - 78, followed by 50, 40, 20, 5, and finally 2 outputs. These two outputs were then compared and whichever was higher, was the predicted output. A Bayesian network was also built, performing a graphical convolution across the nodes to propogate the information from one node to the next which was then condensed to output a probability of an event happening. This is inherently better than the deep learning or logistic regression model because it allows clinicians to understand the and control the risk factors involved to better understand the nature of the disease state. This is extremely useful in both clinical and research settings because oftentimes, families and patients are not only interested in what disorder/disease/condition they have but why the event is likely to occur as in the case of the model that is focused on in this project. The model we constructed for the purposes of this project did not output results better than more traditional models but this could simply be because of the improper node connections when constructing the graph. Making sure that the relevant variables as causaly linked is critical in Bayesian analysis and can be challenging to identify properly with the medical image scan data we have now.

Broadly speaking, this is nowhere near the first example of a probabilistic graphical model used to predict heart health outcomes (See Appendix 1.1). It does, however, add to the growing body of literature surrounding this topic by not only using the normal patient features, but also, including heart calcification data collected from CTCS images. This added set of features provides additional data and therefore should make calculating risk scores and predicting outcomes easier and more accurate. However, this model clearly demonstrated that node and edge selection are critical for determining the success of the network and improper node connections can lead losses in model performance.

Overall, while deep learning and neural networks have taken over the AI world, there is obviously much room where probabilistic models and methods clearly have the upper hand.

Overall, especially in health settings, it is not enough to have a prediction, but the certainty of a particular prediction is often just as important as the prediction itself.

Appendix

1.1 Node Justification

For general heart health connections, relationships between nodes were produced based on knowledge gained from current literature. One paper that the authors investigated, generated a graphical structure for general heart health features as shown below. All of the general heart feature connections were made to model the graphical structure shown on the left. This graph was originally shown by Gupta et. al. while the authors were aiming to create a network to estimate the risk of coronary artery disease.

The dataset that we possess, however, also contains data on heart calcification, which is a process by which atherosclerotic plaques develop. These plaques can then progressively block arteries, eventually leading to myocardial infarction or stroke. The relationships for these nodes were made based on industry best practices, with the help of Dr. Wilson, the PI for the lab where Josh works. The full list of connections are listed in the GrapicalModel.ipynb document.

1.2 Code Availability

To find code relating to all of the work done, please visit this link to a google drive folder which contains all of the python notebooks, python files, and data files related to the project.
> https://drive.google.com/drive/folders/1kIk4Xl4trIL5RSqFhsz7_-CXrqYtQFX-?usp=sharing

Bibliography

[1] Artificial Intelligence (AI): What it is and why it matters. Accessed April 14, 2023. https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

[2] van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. *Nat Rev Methods Primers*. 2021;1(1):1-26. doi:10.1038/s43586-020-00001-2

[3] Bandopadhyay S. *Cardiovascular Risk Prediction from Electronic Health Records Using Probabilistic Graphical Models*. Doctor of Philosophy. The University of Minnesota; 2016.

[4] Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation*. 2022;145(8):e153-e639. doi:10.1161/CIR.0000000000001052

[5] CDC. Heart Disease Resources. Centers for Disease Control and Prevention. Published March 21, 2023. Accessed April 14, 2023. https://www.cdc.gov/heartdisease/about.htm

[6] Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification. *Mobile Information Systems*. 2020;2020:e8843115. doi:10.1155/2020/8843115

[7] Bosco E, Hsueh L, McConeghy KW, Gravenstein S, Saade E. Major adverse cardiovascular event definitions used in observational analysis of administrative databases: a systematic review. *BMC Medical Research Methodology*. 2021;21(1):241. doi:10.1186/s12874-021-01440-5

[8] Mpanya D, Celik T, Klug E, Ntsinjana H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. *IJC H&V*. 2021;34:2352-9067. doi:https://doi.org/10.1016/j.ijcha.2021.100773

[9] Chowdhury S, Chen Y, Wen A, et al. Predicting Physiological Response in Heart Failure Management: A Graph Representation Learning Approach using Electronic Health Records. Published online February 1, 2023. doi:10.1101/2023.01.27.23285129

[10] Gupta A, Slater JJ, Boyne D, et al. Probabilistic Graphical Modeling for Estimating Risk of Coronary Artery Disease: Applications of a Flexible Machine-Learning Method. *MDM*. 2019;38(8):1032-1044. doi:https://doi.org/10.1177/0272989X19879095

[11] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. Published online February 22, 2017. doi:10.48550/arXiv.1609.02907