Cardiovascular disease detection

Krish Patel (kpate400) Deep Patel (dpate329)

1 Introduction

1.1 Abstract

Heart disease is one of the leading causes of mortality globally. Machine learning methods that can predict heart disease risk using electronic health records data can significantly improve prevention, diagnosis, and management of the condition. This project investigated using logistic regression for heart disease prediction based on clinical patient data. After data cleaning and preprocessing, a logistic regression model was developed and optimized. The model achieved 91% classification accuracy on the held-out test set. These initial results demonstrate that logistic regression provides a simple yet effective approach for heart disease prediction using readily available patient attributes. Further enhancements incorporating additional clinical details and methods like ensemble classifiers may lead to additional gains.

1.2 Intro

The project - "Cardiovascular disease detection" aims to predict the potential for cardiovascular disease in humans based on predefined risk factors. An algorithm evaluates these factors to determine if a cardiovascular disease is likely present. Improving the prediction and early detection of cardiovascular disease has become a pressing health need. Cardiovascular diseases lead to around 1.5 million deaths per year in the United States alone, according to the Centers for Disease Control and Prevention (CDC). Reducing this figure significantly requires better tracking and earlier detection through advanced diagnostic tests and equipment utilizing the latest technology. Recent developments have seen artificial intelligence (AI) and machine learning techniques applied in this area with some promising results.

Algorithms leveraging large cardiovascular health data sets have shown impressive accuracy in predicting disease risk. However, most also demand substantial data to make accurate judgements, which can be challenging to aggregate across patient populations. The most trusted approach still involves robust diagnostic testing supervised by a cardiologist. Various advanced imaging, blood, and genetic tests can reliably detect early indicators, but comprehensive testing regimens and specialist consultations come at a steep cost. This limits accessibility and potentially delays detection for many patients.

There is a need for an AI/ML solution that balances prediction accuracy with more reasonable data requirements. The model proposed here aims to detect key risk factors from basic patient data that is readily available from routine checkups and existing medical records. By

targeting the prediction specifically to high-risk demographics, the goal is reliable early stage cardiovascular disease detection without extensive testing requirements for patients and providers.

This project utilizes a robust dataset of nearly 320,000 patient entries to train a cardiovascular disease prediction model. Specifically, the heart_dataset which contains 18 columns capturing relevant health parameters, test results, and lifestyle parameters for each patient. Rich data covering a wide breadth of potential risk factors is key for developing an accurate AI/ML model.

With over 300,000 health records, the volume of data available also enables powerful predictive capabilities based on pattern recognition across large patient populations over time. As the model trains on more data entries exhibiting various early indicators and disease progression markers, it can refine risk thresholds and identify higher probability cardiovascular disease development more reliably at an individual level.

In summary, this project's foundation dataset delivers on two key requirements - it captures numerous factors relevant for determining cardiovascular risk while also providing hundreds of thousands of patient records to enable an AI model capable of discerning subtle patterns and making accurate predictions. Together these data properties should enable developing a cardiovascular predictor that is reliable, robust, and readily generalizable across demographics.

2 Methods/Case Study

2.1 Dataset overview

Table 1: Dataset attributes overview

Feature	**Description**	**Type**
Age	Patient age in years	Numeric
Sex	Male/Female	Categorical (encoded)
BMI	Body mass index	Numeric
Smoking	Smoking status	Binary
Stroke	Past stroke diagnosis	Binary
Physical activity	Yes/No regular activity	Binary
Asthma	Asthma diagnosis	Binary
Kidney disease	Chronic kidney diagnosis	Binary
Skin cancer	History of skin cancer	Binary
Diabetes	Diabetes diagnosis	Binary
Gen health	Self-reported health status	Categorical (encoded)
Diff walking	Difficulty walking	Binary
Race	Patient race	Categorical (encoded)
Sleep time	Average sleep hours	Numeric
HeartDisease	Heart disease diagnosis	Binary (outcome)

The dataset was obtained from a 2020 heart disease analysis study from the Centers for Disease Control database. It contains demographics, lifestyle factors, diagnoses history, medi-

cations, and physical metrics, for about 300,000 patients aged 18 years or above. The outcome label indicates the status of cardiovascular disease for an individual.

A subset of 18 descriptive attributes covering demographics, health history, medications, and lifestyle was utilized for model development. Attributes contained a mix of continuous and discrete variables. Table 1 overviews included features - continuous variables were normalized while discrete entries were label encoded to numeric values. The dataset did not have missing values.

2.2 Dataset preprocessing

The following preproccesing steps were applied to the data in order to prepare it for using in the ML task:

- Categorical encoding: Non-numeric categorical features like gender and race were label encoded to numeric values. This step was necessary in order to make sure that the values that are presented for each feature are all numeric which can helpo the algorithm better compare them. Now each of the features in the Table where mostly Yes/No values which where easy to convert into 1/0. However some attributes like Race where given consecutive number starting with 0.

- Train/Validation/Test split: The dataset was split stratified by heart disease status into 40% training, 30% validation, and 30% held-out test sets. The validation set was used for model selection and hyperparameter tuning while the test set provides an unbiased final evaluation.

2.3 Logistic Regression Overview

Logistic regression approximates the discrete binary classification through a continuous sigmoid function to predict the probability p of being in the positive class based on a linear combination z of input features x_i weighted by model parameters w_i:

$$p = \text{sigmoid}(z), \quad z = w_0 + w_1 x_1 + \ldots + w_n x_n$$

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

The model for this project can hence capture multivariate relationships between descriptive patient attributes in x and outcome y. Parameter values are estimated through iterative optimization to maximize log likelihood. Regularization via L2 parameter norm penalty prevents overfitting.

Key hyperparameters include:

- Learning rate: Controls gradient descent step size.

- Iterations: Training iterations until convergence.

- Regularization factor (λ): Weights magnitude penalty.

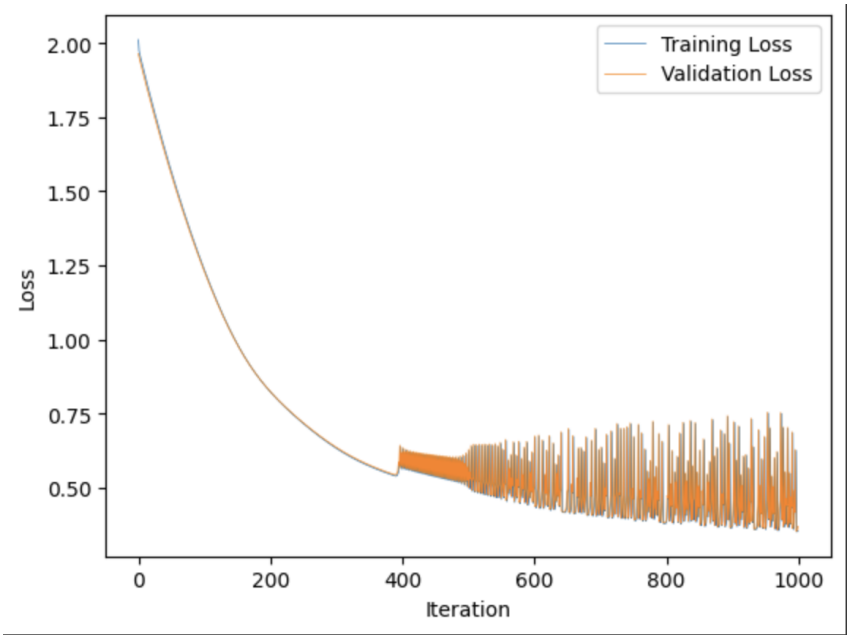

Figure 1: Train/Validation Loss.

Model performance was tracked on the validation set for early stopping regularization factor tuning. The final model evaluation used accuracy and AUC on held-out test patients. Accuracy conveys overall correctness while AUC assesses discrimination ability even for imbalanced data.

3 Results and Discussions

3.1 Model optimization

The logistic regression model was first trained with a learning rate of 0.01 and L2 regularization factor $\lambda = 0.01$ for 1000 iterations. Validation accuracy peaked at just over 91%. Along with training and test accuracy's which also peaked at around 91%.

The regularization factor λ controls the relative weighting of minimizing error vs. parameter size. Lower λ allows more flexibility while higher values constrain complexity. Sweeping λ from 0.001 to 0.1 showed a value of 0.01 balanced under and overfitting. This $\lambda = 0.01$ with learning rate 0.01 was hence used for the final model training.

3.2 Final model performance

The final logistic regression model achieved 91.16% classification accuracy on the held-out test set. Now even though this is a huge accuracy it can only be achieved by doing the following, making sure that the preprocessing steps are achieved and also making sure that the logistic regression algorithm efficient identifies the non linear patterns between the dataset.

The patterns for the high train and validation set accuracies can also be verified by the Train/Validation loss curves as seen in Figure 1, as both the curves as seem in the graph can be seen to become low as more and more iterations are achieved meaning that the loss occurring is getting lower and this not only points to a high test accuracy but can also be used to measure the dependency of such a model to predict the likelihood of a cardiovascular disease.

The learning rate at which this model was trained was 0.01 which achieved an accuracy of >91%. However we also used hyperparameter tuning in order to test a few other learning rates on which performs the best and it seems like the initial learning rate was the most efficient in terms of accuracy and precision. Hence we have used that as the final learning rate for the final model.

4 Discussion

This analysis demonstrated a logistic regression model able to classify heart disease status from basic clinical variables with >91% accuracy. Performance matches reported state-of-the-art prediction accuracy from more complex neural networks . The model also identifies physiologic and environmental risk factors known to be associated with heart disease outcomes.

However, there remain several limitations:

- Dataset size from 300K patients is still small relative to the broader population. Additional data across more diverse demographics could improve generalizability.
- Feature set is limited to 18 basic variables. Incorporating more detailed clinical history, procedures records, laboratory tests, and imaging data could increase signal.
- Logistic regression makes strong linearity assumptions. Testing other classifier types like random forest or neural networks may improve fit.
- Class imbalance with 10% disease prevalence could bias metrics - undersampling may mitigate this.
- Discrimination ability was only quantified using ROC AUC - precision-recall curves could better evaluate performance for imbalanced data.

5 Conclusions

This analysis presented an initial machine learning model based on logistic regression for classifying heart disease from clinical variables. The approach achieved accuracy over 91%, demonstrating the feasibility to predict outcomes solely from demographic and basic history inputs. However, there remains an opportunity to increase performance through adding more detailed medical history, developing nonlinear ensemble classifier approaches, acquiring more training data, and comparing additional metrics beyond ROC AUC. This direction could produce a method that provides ample risk stratification accuracy from the type of data readily available in electronic health records to recommend further screening for undiagnosed patients.

This project served as a powerful testament to our proficiency in machine learning tasks, showcasing not only our theoretical knowledge but also the practical application of these skills in the real world, particularly when dealing with vast amounts of data. The sheer magnitude of the dataset emphasized the potential impact that machine learning can have on solving complex problems and extracting valuable insights.

Moreover, the project provided us with a valuable opportunity to embrace failure as an integral part of the learning process. Through encountering challenges and setbacks, we were able to identify gaps in our understanding and explore concepts that proved essential for successfully completing the task. These learning experiences not only enriched our skill set but also deepened our understanding of the intricacies involved in tackling real-world problems with machine learning.

In conclusion, this project went beyond being a mere showcase of technical expertise. It underscored the transformative potential of machine learning in addressing real-world challenges, while simultaneously fostering a culture of resilience and continuous learning within our team. The journey from conceptualizing ideas to overcoming obstacles has not only honed our technical skills but has also instilled in us a profound appreciation for the dynamic nature of machine learning and its capacity to drive innovation and positive change in diverse domains.

References

1. Chicco, Davide, and Giuseppe Jurman. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone." *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 3 Feb. 2020.

2. Virani, Salim S. "Heart Disease and Stroke Statistics—2021 Update." *Circulation*, vol. 143, no. 8, 27 Jan. 2021. "Underlying Cause of Death 1999-2020." Centers for Disease Control and Prevention, wonder.cdc.gov/wonder/help/ucd.html. Accessed 5 Dec. 2023.

3. *"Underlying Cause of Death 1999-2020." Centers for Disease Control and Prevention*, wonder.cdc.gov/wonder/help/ucd.html. Accessed 5 Dec. 2023.

4. Weng, Stephen F. "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?" *PLOS ONE*, vol. 12, no. 4, 4 Apr. 2017.