Using Machine Learning Models to Predict Heart Disease and Attack with Non-Clinical and Non-Wearable Variables

Samuel Pang,
Lynbrook High
1280 Johnson Ave, San Jose, CA 95129

1. Abstract

Heart disease is the leading cause of death in the United States for over 100 years. However, electrocardiograms (EKG), the primary method of testing for heart issues, costs the average American ~100-5000 dollars per scan. Wearable technologies are much cheaper, but the vast majority of Americans do not own them. Therefore, there should be other cheaper and time efficient methods for other methods of early onset heart disease detection. We used data from the Behavioral Risk Factor Surveillance System (BRFSS), a telephoned based survey with over 200 thousand participants and partnered with theCenters for Disease Control and Prevention (CDC), to find correlations between heart disease and attacks with non-clinical factors. Multiple machine learning models were trained on BRFSS 2015 data. The best model achieved 70% accuracy, while having a 21% precision and a 79% recall. These results show that while there is potential using non-clinical data combined with machine learning to provide a cost-effective preliminary screening tool for heart disease risk, significant improvements are needed before they can be applied effectively in real-world healthcare.

2. Introduction

Heart disease has been the leading cause of death in the United States for over 100 years, accounting for approximately 17.6 million deaths worldwide in 2016 (CDC, 2025b). Heart attacks, the most common acute manifestation of heart disease, are typically caused by the buildup of fat and cholesterol in the coronary arteries, which can obstruct blood flow to the heart muscle (Mayo Clinic, n.d.). Early detection of heart disease significantly reduces the risk of serious complications, including myocardial infarction and stroke, lowers long-term healthcare costs, improves survival and quality of life, and expands treatment options (CDC, 2025a).

Electrocardiograms (EKG) are the standard clinical tool for detecting heart issues. However, EKG scans in the United States cost an average of $100-$5,000 per test (Hanzen, 2025), creating a major financial hurdle to many Americans. Surveys indicate that many Americans avoid routine medical care due to cost (24.1%), lack of health insurance (8.3%), time constraints (15.6%), or low perceived need (12.2%) (Taber et al., 2025).

This study uses non-clinical survey data from the Behavioral Risk Factor Surveillance System (BRFSS), a nationwide telephone-based survey conducted by the Centers for Disease Control and Prevention (CDC, 2022), to predict heart disease risk. The dataset contains both numerical and categorical variables, including lifestyle, and self-reported health indicators (Kaggle, 2023). Using this dataset, multiple supervised machine learning models were trained to classify individuals with a history of heart disease or heart attack. The output is binary, indicating the possibility of heart disease or a prior heart attack. This approach aims to provide a cost-effective and easy-to-use preliminary screening tool, using variables that can be easily reported via mobile devices that over 90% of Americans possess (Sidoti & Dawson, 2024).

3. Background

Several recent studies have explored the use of data-driven approaches for early detection and risk prediction of cardiovascular disease (CVD). One promising direction involves wearable sensing and big data technologies. A study by Miao et al. (2023) utilized data from wearable devices, such as Fitbit and Apple Watch, to predict heart attacks with relatively high accuracy, achieving over 80% predictive performance. The study leveraged continuous physiological monitoring, including heart rate and activity tracking, to generate individualized risk scores. While this approach demonstrates strong predictive capability, it requires participants to consistently wear devices and actively generate data, limiting its scalability for populations without widespread wearable adoption.

Another approach investigates the predictive value of non-clinical versus clinical data. Schiborn et al. (2021) demonstrated that models using easily obtainable non-clinical CVD risk factors can achieve comparable predictive performance to traditional clinical models that rely on laboratory tests and medical imaging. The study found that including additional clinical parameters provided only marginal improvements in discrimination, suggesting that accessible non-clinical data can serve as a viable alternative for population-level risk screening. The main limitation of this approach is the requirement for detailed self-reported lifestyle and diet information, which may not always be consistently collected.

These studies highlight both the potential and constraints of alternative cardiovascular risk assessment strategies. Wearable-based models offer strong real-time predictive power but depend on technology adoption, whereas non-clinical survey-based models are broadly accessible but may require extensive self-reported data. Building on these insights, our work leverages the Behavioral Risk Factor Surveillance System (BRFSS) dataset, which contains a mixture of numerical and categorical non-clinical variables, including demographic, lifestyle, and self-reported health information. By training supervised machine learning models on this dataset, our study seeks to create a cost-effective and widely accessible preliminary screening tool for heart disease, providing a complementary approach to both wearable-based and traditional clinical methods.

4. Dataset

The dataset used for this study is the “Diabetes Health Indicators Dataset” by Alex Teboul, available on Kaggle. This dataset is a cleaned subset of the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the U.S. Centers for Disease

Control and Prevention (CDC). The BRFSS is the largest continuously conducted health survey in the world, collecting responses from over 400,000 individuals each year on topics such as health behaviors, chronic conditions, and preventative health practices.

For this project, we specifically used the file diabetes_012_health_indicators_BRFSS2015.csv, which contains 253,680 survey responses from the 2015 BRFSS dataset. This version includes 21 non-clinical features, such as age, BMI, smoking status, alcohol consumption, physical activity, stroke history, high blood pressure, and high cholesterol. These features are self-reported or reflect general health conditions that respondents are typically already aware of, rather than requiring clinical measurements or wearable devices. This aligns with the project's goal of predicting heart attack risk using information accessible to the general public without medical testing.

The correlation matrix (Figure 1) showed multiple variables that could be used to predict Heart Attack or Disease. These included Age, Stroke, HighBP, HighChol, DiffWalk (days of difficulty walking in the past month), Diabetes_012, and PhysHlth (days of physical health complications in the past month).

- Diabetes_012: This feature contained 3 different values. 0 meant no diabetes. 1 is for pre-diabetes or borderline diabetes. 2 is for confirmed diabetes.
- HighBP: Adults who have been told they have high blood pressure by any health professional. 1 means they have high blood pressure, 0 means they do not have high blood pressure.
- HighChol: Adults who have ever been told they have high cholesterol pressure by any health professional. 1 means they have high cholesterol, 0 means they do not have high cholesterol.
- CholCheck: If the participant has had cholesterol checked within the past 5 years. 1 means they have received a checkup in the last 5 years, 0 means that they have not received a checkup in the last 5 years.
- BMI: The participant's Body Mass Index. It is an integer value.
- Smoker: Participants who have smoked at least 100 cigarettes in their lifetime. 1 means they have had at least 100 cigarettes, 0 means they have had less than 100 cigarettes.
- Stroke: Patients who have suffered a stroke in their lifetime. 1 means they have suffered a stroke, 0 means they have never suffered a stroke.
- HeartDiseaseOrAtack: This is the feature that we would predict. It shows if a participant has ever had a heart attack or has heart disease. 0 means they have neither, and 1 means that they have had either a heart attack and/or have heart disease.
- PhysActivity: This shows how many days in the last month the participant has engaged in physical activity in the last 30 days aside from their regular job. It is an integer value going from 0 to 30, indicating the number of days with non-job related physical activity.
- Fruits: This feature shows if the participant consumes at least one fruit per day. 0 means they consume less than 1 fruit a day, and 1 means they consume at least 1 fruit a day.
  Veggies: This feature shows if the participant consumes at least one vegetable per day. 0 means they consume less than 1 vegetable a day, and 1 means they consume at least 1 vegetable a day.
- HvyAlcoholConsump: This value shows if the participant is a heavy drinker. Heavy drinkers are classified as having over 14 drinks in the past week for male participants, and over 7 drinks per week for female participants. A value of 1 means the participant passes the threshold, and a value of 0 means they do not meet the threshold.
- AnyHealthcare: This value indicates if the participant currently has any healthcare coverage, including government plans such as Medicare. A value of 1 shows the participant has healthcare, and a value of 0 means the participant does not have healthcare.

- NoDocbcCost: This value shows if the participant has not visited a doctor due to cost reasons in the past 12 months. A value of 1 shows the participant has not seen a doctor when they needed to because of cost issues, and a value of 0 shows the participant has not avoided a doctor due to cost issues.
- GenHlth: This value is an integer scale of 1-5 describing the participant's rating of their own general health. 1 is perceived excellent health, and 5 is perceived poor health.
- MentHlth: This integer value is the number of days in the past 30 days a participant had poor mental health.
- PhysHlth: This integer value is the number of days in the past 30 days a participant had poor physical health.
- Diff\Walk: This value indicates if a participant has difficulty walking. A value of 1 means difficulty, and a value of 0 means no difficulty walking.
- Sex: This indicates the gender of the participant. 1 is male, and 0 is female .
- Age: This value is the age of the participant. Ages 18-24 are represented with 1, with increments of 5years up to a value of 13.
- Education: An ordinal categorical variable with six levels representing highest education attained. 1 = never attended school/kindergarten only, 2 = grades 1-8, 3 = grades 9-11, 4 = grade 12 or GED, 5 = some college or technical school, 6 = college graduate (4+ years).
- Income: An ordinal categorical variable with eight levels representing annual household income. 1 = <$10,000, 2 = $10,000-14,999, 3 = $15,000-19,999, 4 = $20,000-24,999, 5 = $25,000-34,999, 6 = $35,000-49,999, 7 = $50,000-74,999, 8 = $75,000 or more.

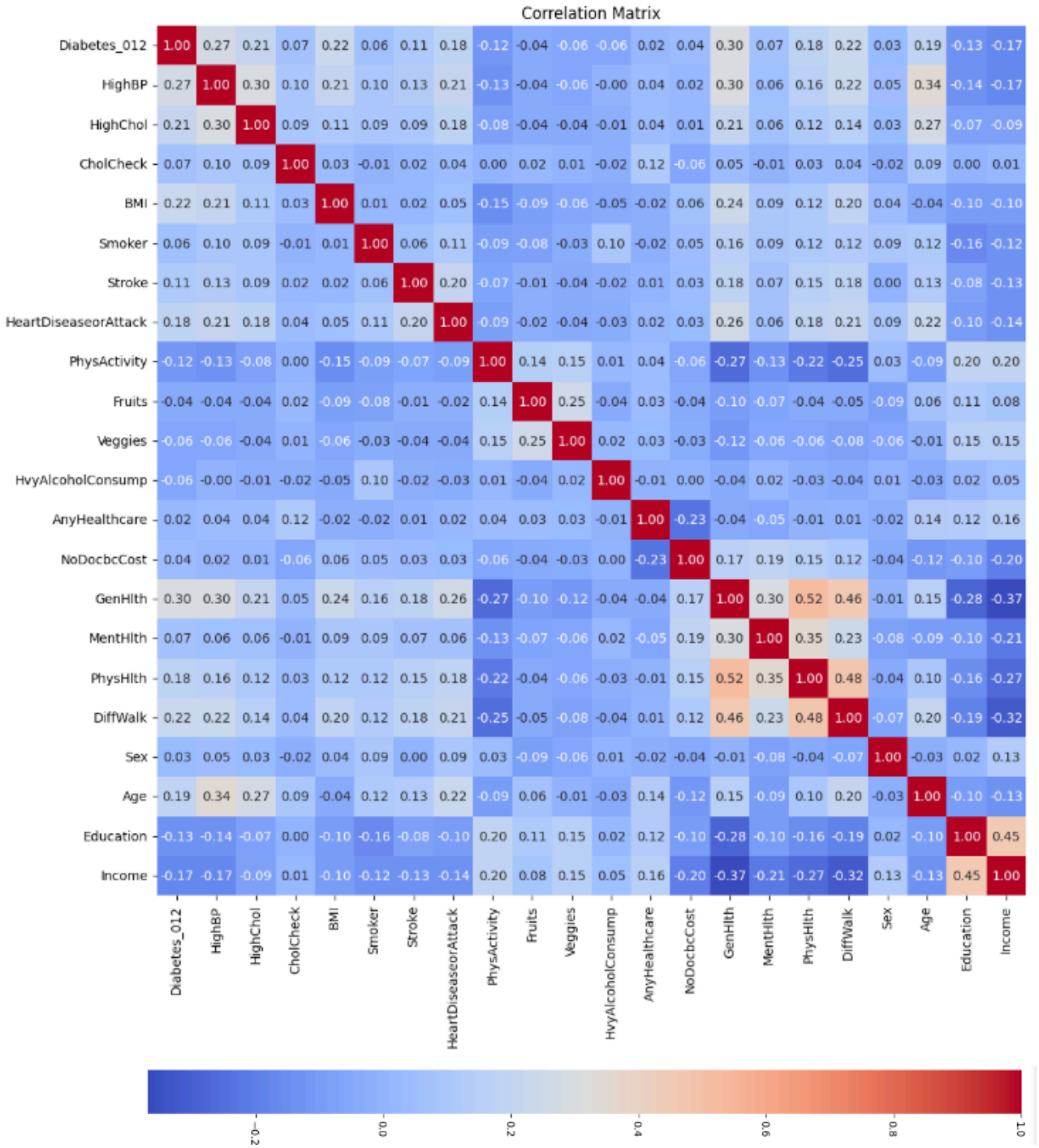

Figure 1. Correlation Matrix

The dataset was split using an 80-20 train-test split. Of the 253,680 records, approximately 202,944 were used for training, and 50,736 were reserved for testing.

However, due to poor performance, the dataset was then manually balanced for the column "HeartDiseaseOrAttack" to balance the weights.

5. Methodology/Models

We used multiple machine learning models in order to get the highest performance. The following models were used. All models are part of the Scikit-learn python package, a very popular free python machine learning package, and would use a random state of 42 if applicable.

- Logistic Regression: A linear model used for classification that estimates the probability of class membership using a logistic (sigmoid) function.
- Decision Tree Classifier: A non-linear model that splits the data into branches based on feature values, producing a tree-like structure for classification.
- Random Forest Classifier: An ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
- Neural Network: A model inspired by biological neurons that learns complex patterns through layers of interconnected nodes (neurons).
- XGBooster: An optimized gradient boosting algorithm that builds trees sequentially, where each new tree corrects errors made by the previous ones, providing high performance for structured data.

The first model uses only Age and BMI on the Scikit-learn Logistic Regression model.

The second batch of models used the same set of features: *Age, Stroke, HighBP, HighChol, DiffWalk, Diabetes_012,* and *PhysHlth*. An 80/20 train-test split was applied for model evaluation. A Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Neural Network were used.

The second model involved using every single variable on a Scikit-learn Random Forest Classifier model with 100 estimators.
Afterwards, a decision tree was also used.
Finally, a neural network with the size of 100, 50 nodes, with 1000 iterations, was trained using the same data and features.

After looking at the results, both neural networks and random forest classifier models were run again multiple times. This time we decided to use SMOTE to balance the training data with a 50-50 split on rows with HeartDiseaseOrAttack and without. The Neural network was run multiple times with varying shapes and iterations.

- A layer size of 100, 50 for 1000 iterations
- A layer size of 50, 25 for 1000 iterations
- A layer size of 100, 100, and 100 for 10 iterations
- A layer size of 7, 7 for 1 iteration
- A layer size of 7, 14 for 1 iteration
- A layer size of 100, 25, 100 with 100 iterations

Random Forest Classifier models with balanced inputs were run 886 times, going from 1 estimator to 886 estimators.

Finally, an XGBoost model with the same new balanced training data was trained with the same features.

6. Results and Discussion

The metrics used to evaluate the models were Accuracy, ROC AUC, Precision, and Recall (Google, 2024).

- Accuracy: The proportion of all classifications that were correct, whether positive or negative.
- ROC AUC: A metric that summarizes how well a binary classification model can distinguish between positive and negative classes across various classification thresholds.
- Precision: The proportion of all the model's positive classifications that are actually positive.
- Recall: The proportion of all actual positives that were classified as positives.

All models had high Accuracy (~90%) and ROC AUC (~90%). However, this performance was misleading due to class imbalance: ~90% of the dataset were of negative cases, leading to models that could achieve high Accuracy by almost only predicting the negative class. Precision and Recall provided a more realistic evaluation, with most models performing poorly (≈50% Precision and ≈7% Recall). The best precision was achieved by the Neural Network (55.7%), while the highest Recall came from Logistic Regression (8.7%).

To address the imbalance, SMOTE was applied to balance the training data with a 50-50 split between positive and negative cases. Neural Networks were retrained with several different shapes, including:

- Layer sizes [100, 50], 1000 iterations
- Layer sizes [50, 25], 1000 iterations
- Layer sizes [100, 100, 100], 10 iterations
- Layer sizes [7, 7], 1 iteration
- Layer sizes [7, 14], 1 iteration
- Layer sizes [100, 25, 100], 100 iterations

Across these trials, results were nearly identical, averaging ~70.6% Accuracy, 81.8% ROC AUC, 21.3% Precision, and 79.1% Recall.

Random Forest Classifier models were trained 886 times with balanced inputs using estimators from 1 to 886. Despite the wide range, performance on all metrics barely changed, averaging ~73.6% Accuracy, 77.9% ROC AUC, 21.8% Precision, and 69.8% Recall.

Models using All Specified Features

Model	Confusion Matrix	Accuracy	ROC AUC	Precision	Recall
Logistic Regression	[[45611, 357], [4355, 413]]	90.71%	82.04%	53.64%	8.66%
Decision Tree	[[45570, 398], [4443, 325]]	90.46%	78.47%	44.95%	6.81%

Random Forest	[[45547, 421], [4411, 357]]	90.47%	80.07%	45.87%	7.49%
Neural Network	[[45611, 357], [4355, 413]]	90.71%	82.04%	53.64%	8.66%
Neural Network (Balanced)	[[32033, 13935], [998, 3770]]	70.57%	81.78%	21.29%	79.07%
Random Forest (Balanced)	[[33999, 11969], [1439, 3329]]	73.57%	77.98%	21.76%	69.82%
XGBoost (Balanced)	[[37572, 8396], [1973, 2795]]	79.56%	79.87%	24.95%	58.62%

Table 1. Model metrics and performances. Confusion Matrices are represented in the following shape [[True Negative, False Positive],[False Negative, True Positive]]

Finally, an XGBoost model trained on the balanced dataset achieved results comparable to Random Forest, with Accuracy and ROC AUC in the same range and similar Precision and Recall tradeoffs.

7. Conclusion

This study demonstrated that machine learning models trained on non-clinical survey data from the BRFSS can identify individuals at risk of heart disease with high recall but low precision. While models such as Neural Networks, Random Forest, and XGBoost had recall rates near 80%, their precision remained low (~21%), highlighting the challenge of false positives in predictive screening. These findings suggest that non-clinical factors –such as age, BMI, and lifestyle behaviors– have predictive value and could be used for cost-effective, large-scale preliminary risk assessments.

The primary use case for such models is an initial screening tool to flag individuals who otherwise seem unlikely to have a heart attack or disease, enabling targeted follow-up with clinical testing without the need to screen the entire population. This approach could reduce healthcare costs, prioritize preventive care, and increase early detection in populations that otherwise would be unlikely to have a checkup. However, current models are not precise enough to serve as the standalone diagnostics due to the high rate of false positives.

Future work should focus on refining feature selection, exploring more sophisticated methods for handling class imbalance, and integrating additional datasets to improve predictive accuracy. With these enhancements, non-clinical machine learning models have the potential to become a practical supplement to traditional diagnostic methods, guiding preventive interventions and helping to reduce the burden of heart disease in the general population.

8. Acknowledgements

I would like to give special thanks to Dr Rami Abi-Akl for his guidance throughout this study. His insightful feedback and expertise were instrumental in shaping this research.

9. References

CDC. (2022). *Behavioral Risk Factor Surveillance System*. CDC. https://www.cdc.gov/brfss/index.html

CDC. (2025a, February 14). *About heart attack symptoms, risk, and recovery*. Heart Disease. https://www.cdc.gov/heart-disease/about/heart-attack.html

CDC. (2025b, July 10). *Heart disease facts*. Heart Disease. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

Diabetes health indicators dataset. (n.d.). Retrieved September 14, 2025, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Google. (2024). *Classification: Accuracy, recall, precision, and related metrics. Google for Developers*. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Hazen, T. (2025, April 11). *How much does an EKG cost?*. BetterCare. https://bettercare.com/costs/ekg-cost

Heart attack. (n.d.). Mayo Clinic. Retrieved September 14, 2025, from https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106

Miao, F., Wu, D., Liu, Z., Zhang, R., Tang, M., & Li, Y. (2023). *Wearable sensing, big data technology for cardiovascular healthcare: current status and future prospective*. Chinese Medical Journal, 136(09), 1015-1025.

Schiborn, C., Kühn, T., Mühlenbruch, K., Kuxhaus, O., Weikert, C., Fritsche, A., Kaaks, R., & Schulze, M. B. (2021). A newly developed and externally validated non-clinical score accurately

predicts 10-year cardiovascular disease risk in the general adult population. *Scientific Reports*, *11*, 19609. https://doi.org/10.1038/s41598-021-99103-4

Sidoti, O., & Dawson, W. (2024, November 13). *Mobile fact sheet*. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/mobile/

Taber, J. M., Leyva, B., & Persoskie, A. (2014). Why do People Avoid Medical Care? A Qualitative Study Using National Data. *Journal of General Internal Medicine, 30*(3), 290–297. https://doi.org/10.1007/s11606-014-3089-1