Heart Disease Risk Prediction

Hanna Othmal
Naga Mounica Gani
University of Texas, Arlington
nxg9704@mavs.uta.edu, hanna.othmal@mavs.uta.edu

Abstract

Cardiovascular disease is a leading cause of global mortality. We propose a predictive framework for assessing individual risk of heart disease by leveraging environmental, behavioural, and medical history factors. Our approach builds a logistic regression model in Python and further employs support vector machines and anomaly‑detection techniques for model tuning and comparative evaluation. We rigorously assess model performance using accuracy, precision, recall, and ROC‑AUC metrics on a publicly available dataset. Our experiments demonstrate that incorporating anomaly detection improves sensitivity to at‑risk individuals, while SVM‑based tuning enhances discriminative performance relative to the baseline logistic regression. The results underscore the value of combining classical and advanced machine learning methods for early detection and proactive management of cardiovascular disease risk.

1 Introduction

The heart and blood arteries are affected by a group of ailments known as cardiovascular diseases (CVDs) and cardiac disorders. These conditions include deep vein thrombosis, coronary heart disease, peripheral arterial disease, and rheumatic heart disease. One of our body's most important organs is the heart. It is a muscular organ found directly beneath and just to the left of the breastbone. The World Health Organization estimates that CVDs account for 17.9 million deaths annually, making them the leading cause of death worldwide. The majority of these fatalities, which make up 85% of them and happen in low- and middle-income nations, are mostly caused by heart attacks and strokes. Furthermore, persons under the age of 70 account for one-third of these early mortalities. Poor diet, inactivity, smoking, and binge drinking are a few risk factors for heart disease and stroke. These psychological risk factors can manifest physically as obesity, overweight, high blood lipid levels, high blood pressure, and high blood sugar. To take preventive actions and avoid the damage that these disorders can cause, early identification of cardiovascular diseases is essential. Underlying blood vessel issues may go untreated for a long time. The condition typically manifests first as heart attacks and strokes. Heart attack symptoms include pain or discomfort in the middle of the chest, in the arms, left shoulder, elbow, jaw, or back. The person may also have back or jaw discomfort, nausea or vomiting, lightheadedness, or fainting. Common stroke symptoms include trouble speaking or understanding speech, disorientation, difficulty seeing with one or both eyes, numbness on one side of the face, arm, or leg, difficulty walking, and a severe headache with no known reason. Due to a lack of basic healthcare facilities for early detection and treatment, cardiovascular diseases often have higher death rates in low and middle-income nations. People with CVDs have limited access to adequate, equitable healthcare services that can meet their needs in low and middle-income countries. Due to the disease's delayed identification, people get CVDs and pass away early. It has been demonstrated that lowering salt intake, increasing fruit and vegetable consumption, engaging in regular exercise, quitting smoking, and abstaining from excessive alcohol consumption all reduce the risk of cardiovascular disease. Additionally, recognizing people who are most at risk for CVDs and ensuring they receive the right care will help avoid early mortality. All primary healthcare facilities must have access to necessary medications and fundamental medical technology to guarantee that individuals in need receive treatment and counseling [8]

2 Related Work

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, accounting for an estimated 17.9 million deaths in 2019 alone. Recent advances in epidemiology, clinical guidelines, and computational methods have aimed to refine our understanding of CVD burden, identify key risk factors, and improve early detection and management. This review synthesizes insights from landmark epidemiological studies, global health assessments, clinical practice guidelines, and state-of-the-art machine learning (ML) and artificial intelligence (AI) applications in CVD research and care.

Vaduganathan et al. provide a comprehensive overview of the worldwide impact of CVDs, illustrating a shifting burden toward low- and middle-income countries despite substantial advances in prevention and treatment in high-income settings [1]. They highlight that ischemic heart disease and stroke together account for over three-quarters of all CVD deaths, emphasizing the need for tailored public health strategies. Complementing this, the World Health Organization reports that CVDs represent 32 % of global deaths, with risk factor prevalence, such as hypertension and diabetes, increasing in developing regions [8].

Adhikary et al. systematically review principal modifiable and non-modifiable risk factors, including hypertension, dyslipidemia, tobacco use, poor diet, physical inactivity, obesity, and genetic predisposition [2]. They note that while traditional factors like elevated blood pressure and cholesterol remain dominant, emerging concerns such as air pollution and psychosocial stress are gaining recognition. The Framingham Heart Study (FHS), initiated in 1948, first quantified the role of these factors and developed risk scoring algorithms still in use today [6]. Findings from FHS revealed that age, sex, blood pressure, cholesterol levels, smoking status, and diabetes collectively predict 10-year CVD risk, laying the groundwork for contemporary risk calculators.

The Framingham Heart Study, sponsored by the National Heart, Lung, and Blood Institute, represents a seminal cohort study that followed over 5,000 individuals for decades to identify determinants of CVD [6]. Its contributions include establishing the concept of “risk factor” epidemiology, developing multivariate risk scores, and informing preventive interventions. Continued follow-up and expansion into multi-ethnic cohorts have refined these risk models, ensuring relevance across diverse populations. The FHS exemplifies how longitudinal data can transform population health understanding and guide clinical practice. Based on accumulating evidence, Virani et al. published the 2023 AHA/ACC guideline for managing chronic coronary disease, integrating updated risk stratification, lifestyle interventions, pharmacotherapy, and revascularization strategies [7]. Recent years have seen a rapid integration of ML and AI techniques into CVD research. Baghdadi et al. review advanced ML algorithms—such as support vector machines, random forests, and deep learning—for early detection and diagnostic support, highlighting improved predictive performance over traditional statistical models [3]. Sun et al. further discuss AI applications in diagnostic imaging (e.g., automated echocardiographic interpretation), risk stratification, and therapeutic decision support, while acknowledging challenges related to data quality, interpretability, and regulatory approval [4]. Srinivasan et al. demonstrate an active learning approach using the UCI heart disease dataset, where model accuracy improves iteratively by intelligently selecting the most informative cases for labeling [5]. Collectively, these studies showcase the potential of AI to enhance early detection, personalize treatment, and optimize resource allocation, albeit with the need for rigorous validation in clinical settings.

3 Dataset

This study leverages a Kaggle dataset containing 70,000 records and 19 variables, capturing patient profiles, symptoms, lifestyle factors, and medical history alongside a binary indicator of heart-disease risk. Symptom features include chest pain, shortness of breath, unexplained fatigue, palpitations, dizziness, swelling, radiating pain, and cold sweats. Risk factors encompass age (recorded as a continuous variable), hypertension, high cholesterol, diabetes, smoking history, obesity, and family history of heart disease. Prior studies have shown that many of these variables can serve as predictors of cardiac risk. With all features encoded as binary values except for age, we apply binary classification techniques to predict each patient's likelihood of developing heart disease.

4 Methodology

The methodology follows a structured approach, divided into four phases: data pre-processing, model creation and training, model testing and evaluation, comparison and final model classification. Initially, preprocessing the obtained data was done by handling any missing values, normalizing the features, and checking for imbalanced features, as these would affect model creation. Then, an exploratory analysis was conducted to analyze the features and determine which factors are more highly related in the dataset. Feature selection is then applied to obtain the most relevant features that determine the risk of heart disease. The following is a figure of the heatmap created to check the features for the most correlated features from the dataset.

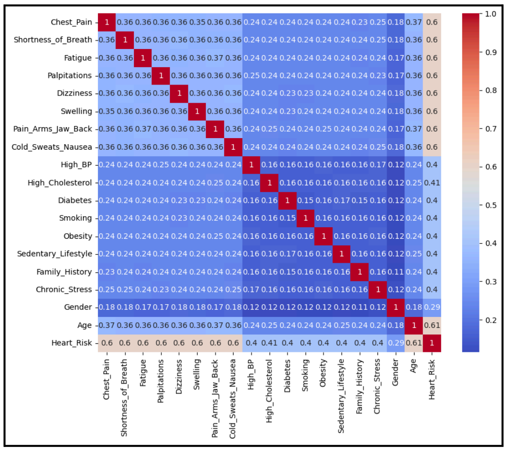

Figure 1: Correlation heat map

For modelling, the two machine learning algorithm built from scratch are used to compare are logistic regression and support vector machines (SVM) to create classification models to categorize the patient to be either high risk of heart disease or low risk The models are then assessed using confusion matrix, classification report, and rox curve to determine how accurately each of the models classified the patient. Anomaly detection can also be done to analyze any specific outliers or data that deviates from the norm in the dataset. Finally, the results are visualized and analyzed to determine which algorithm is best at predicting the patient's risk of heart disease. The following figures are some examples of the features that will be used to assist in predicting the risk of heart disease.

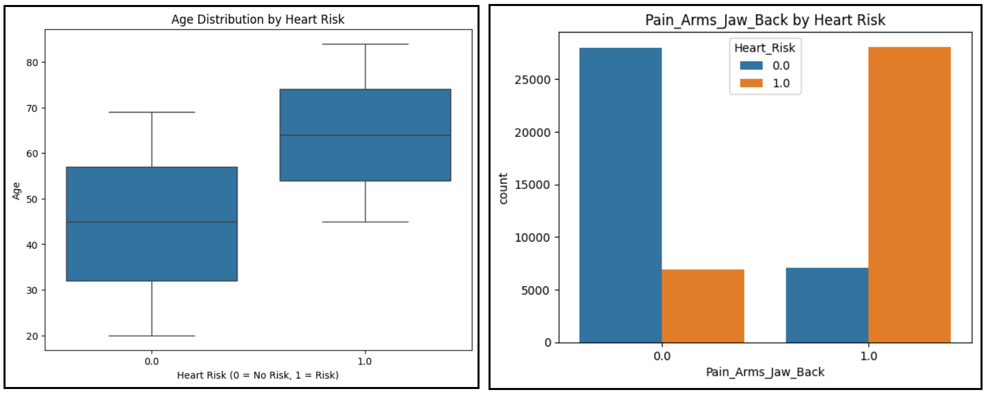

Figure 2: Age distribution and Pain in Arm/Jaw/Back

5 Algorithms

Logistic regression is a probabilistic classification algorithm that models the log-odds of a binary outcome as a linear combination of input features. When creating the algorithm from scratch, we employed the logistic function to map raw model outputs to predicted probabilities between 0 and 1. Model parameters are estimated by maximising the likelihood of observed labels, typically via gradient-based optimisation. The logistic regression class implemented a fit, predicted probability, predict, accuracy, save, and load method. Despite its simplicity, logistic regression is interpretable, provides calibrated probabilities, and is a strong baseline for many applications.

Support Vector Machines (SVM) are margin-based classifiers that seek the hyperplane maximizing the distance between classes in feature space. SVMs handle non-linear separations through kernel functions without explicitly transforming data into higher dimensions. For the SVM class from scratch, the code implemented a compute loss, fit, and predict method. SVMs are robust to overfitting in high-dimensional spaces but can be computationally intensive for large datasets.

Anomaly detection encompasses methods for identifying outliers or rare events that deviate from normal patterns, often without labelled examples of anomalies. Effective anomaly detection supports early identification of critical issues, such as fraud, equipment failures, or unusual patient health indicators.

6 Results

Based on the feature selection, three different models were created using both logistic regression and support vector machines to compare which method was able to better predict the risk of heart disease: a model with the highest relevant features, symptom and risk features, and separate symptom and risk feature sets. Before training and testing each model, the dataset was split into an 80% training set and a 20% test set. After creating the models, the accuracy results were compared to see which machine learning algorithm is better in detecting the risk of heart disease.

6.1 Logistic regression

When training, the model that produced the highest accuracy both in training and in testing was the model of the most relevant features of the dataset. These relevant features included age, pain in arm/jaw/back, cold sweats/nausea, chest pain, dizziness, fatigue, swelling, shortness of breath, and palpitations. The training accuracy of the model was around 98%, and around 97% accuracy during testing. The figure below displays the confusion matrix that was obtained from the model, and the table displays the classification report from testing the model. This figure displays that the confusion matrix for the relevant features gives the best in terms of predicting false positives and false negatives

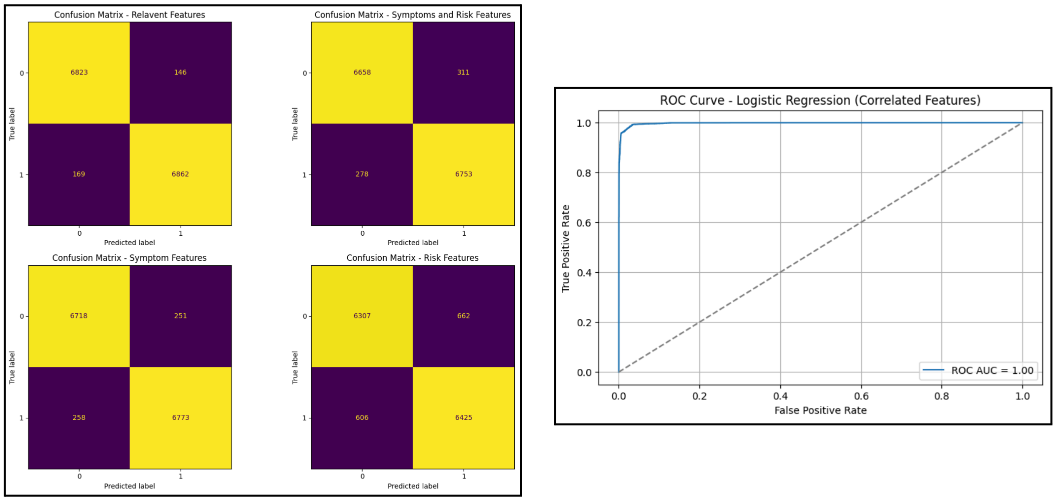

Figure 3: Confusion matrix for Logistic Regression and ROC curve

6.2 Support vector machine

When training with the SVM model created from scratch, the model accuracy displayed the best results with both testing and training was also the model with the features that were most relevant. The training accuracy for this model was 96% and the testing accuracy was also around 96%, which is less accurate than the model created from using logistic regression. The confusion matrix for the models created is shown in the figure below, along with the training loss for the first SVM model. The confusion matrix displays that the first model predicted more false positives than the others, but it also predicted fewer false negatives than the other models. In this case, out of all of the SVM models, it displayed the best results as predicting a patient has the risk of heart disease when they don't is a better scenario than predicting a patient has a low risk when they do.

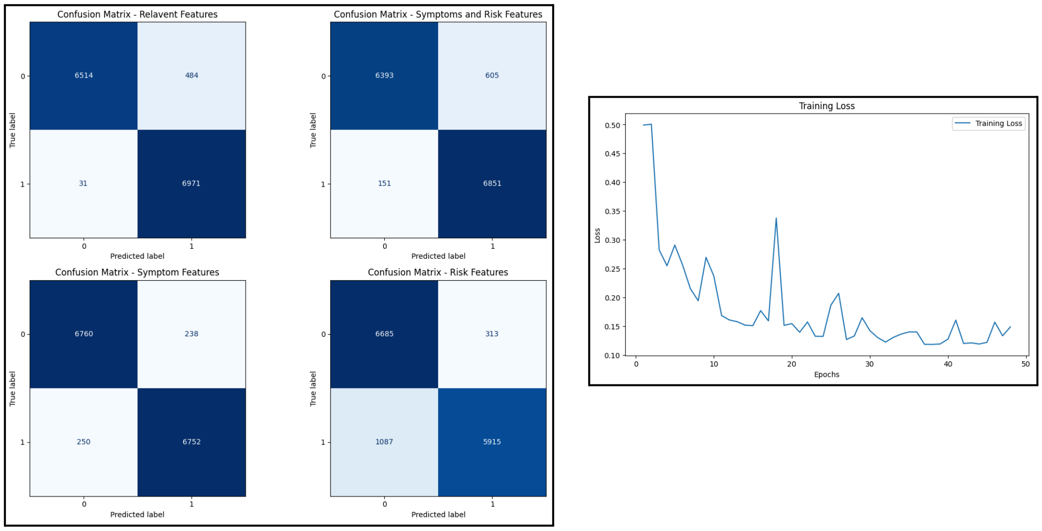

Figure 4: Confusion matrix for SVM and Training loss

6.3 Anomaly detection

The objective of anomaly detection is to identify patients exhibiting atypical combinations of symptoms and risk factors, potentially signifying unique health profiles or data inconsistencies. Presenting a concise summary of detected anomalies allows healthcare professionals and data analysts to rapidly pinpoint cases that merit further investigation. This approach proves particularly valuable in clinical environments, facilitating the prompt detection of unusual patient conditions and the identification of potential data-entry errors within extensive healthcare datasets.

The histogram displays the frequency of anomaly scores ranging roughly from –0.11 to +0.15. Most scores fall between 0 and 0.12, indicating that the bulk of observations exhibit mild deviation from normal. There is a long left tail of lower (negative) scores, but very few data points below –0.05. The distribution is right-skewed, with counts peaking around scores of 0.10 – 0.12 (over 4,000 occurrences). This shape suggests that extreme negative anomalies are rare, while moderate positive anomaly scores are common.

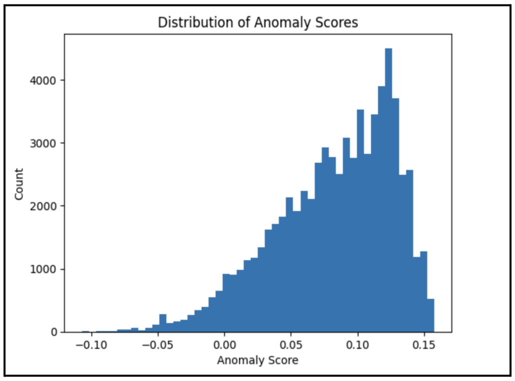

Figure 5: Distribution of Anomaly Scores

Based on the anomalies that were detected, the values that were detected as anomalies were removed from the dataset and a completely new model was created from the feature engineering done during anomaly detection. The selected features included age, gender, symptom count, risk count, high blood pressure and cholesterol, and symptom risk ratio. With these as new features, they were trained and tested using logistic regression, as it was the model that displayed the best accuracy in predicting the risk of heart disease. The result of both the training and testing from these features gave around a 99% accuracy and lower amounts of false positives and false negatives predicted, as shown in the figure of the confusion matrix below.

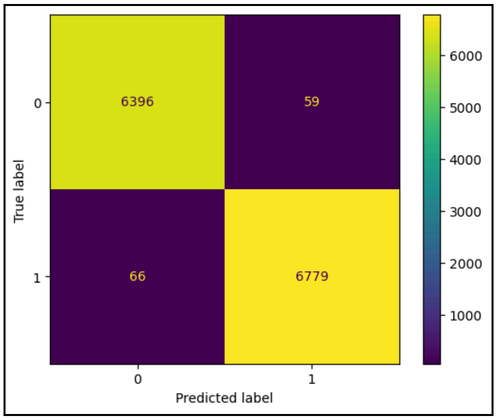

Figure 6: Confusion Matrix for Improved Logistic Regression Model

6.4 Performance Evaluations of Machine Learning Algorithms

The models were evaluated according to these four (4) measures: Accuracy, Precision, Recall, and F1-score. The following two tables show the performance metrics for the models from the logistic regression model and from the SVM models. From these two tables, logistic regression performed the best when using feature selection, but overall performed well in the other models. SVM had also performed well, but was slightly less accurate than logistic regression, especially when feature selection is optimized. This creates the assumption that using risk factors alone is not sufficient for predictive factors.

Table 1: Performance metrics from Logistic Regression

Model	Class	precision	recall	f1	support	accuracy
Model 1 - top correlated/relevant	0	0.9758	0.9791	0.9774	6969	
	1	0.9792	0.976	0.9776	7031	0.9775
Model 2 - symptoms and risk	0	0.9599	0.9554	0.9576	6969	
	1	0.956	0.9605	0.9582	7031	0.9579
Model 3 - just symptoms	0	0.963	0. 9640	0.9635	6969	
	1	0.9643	0.9633	0.9638	7031	0.9636
Model 4 - just risk	0	0.9123	0.905	0.9087	6969	0.9094
	1	0.9066	0.9138	0.9102	7031	

Table 2: Performance metrics from SVM

Model	Class	precision	recall	f1	support	accuracy
Model 1 - top correlated/relevant	0	0.9953	0.9308	0.962	6998	
	1	0.9351	0.9956	0.9644	7002	0.9632
Model 2 - symptoms and risk	0	0.9769	0.9135	0.9442	6998	
	1	0.9189	0.9784	0.9477	7002	0.9460
Model 3 - just symptoms	0	0.9643	0. 9660	0.9652	6998	
	1	0.966	0.9643	0.9651	7002	0.9651
Model 4 - just risk	0	0.8061	0.9553	0.9052	6998	0.9000
	1	0.9497	0.8448	0.8997	7002	

7 Conclusion

Accurately forecasting cardiovascular disease risk holds significant promise for enhancing early diagnosis in clinical settings. Rather than supplanting physicians, machine learning approaches serve as decision-support tools that can streamline patient evaluation and reduce the financial and logistical burdens of extensive clinical and laboratory testing. Our feature selection and model-evaluation processes identified patient age, specific symptomatic presentations, and the presence of hypertension and high cholesterol as the strongest predictors of heart-disease risk, underscoring that traditional risk-factor profiles alone are insufficient. Incorporating feature engineering further boosted predictive performance, and logistic regression consistently outperformed support vector machines in both accuracy and precision. Collectively, these findings demonstrate that machine-learning-driven risk models can facilitate early intervention and long-term prevention strategies. Future work should explore additional algorithms and leverage more diverse datasets encompassing a wider array of variables and value scales to further refine and generalize cardiovascular-risk prediction.

8 Acknowledgements

We thank Soumaya Gharsallaoui and Prajwal Venkat Venkatesh for insightful comments and feedback.

References

[1] Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., and Roth, G. A. The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. J. Am. Coll. Cardiol., 80(25):2361–2371, 2022. doi:10.1016/j.jacc.2022.11.005

[2] Adhikary, D., Barman, S., Ranjan, R., and Stone, H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus, 14(10):e30119, 2022. doi:10.7759/cureus.30119

[3] Baghdadi, N. A., Farghaly Abdelaliem, S. M., Malki, A., Gad, I., Ewis, A., and Atlam, E. Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis. J. Big Data, 10:144, 2023. doi:10.1186/s40537-023-00817-1

[4] Sun, X., Yin, Y., Yang, Q., et al. Artificial Intelligence in Cardiovascular Diseases: Diagnostic and Therapeutic Perspectives. Eur. J. Med. Res., 28:242, 2023. doi:10.1186/s40001-023-01065-y

[5] Srinivasan, S., Gunasekaran, S., Mathivanan, S. K., et al. An Active Learning Machine Technique Based Prediction of Cardiovascular Heart Disease from UCI-Repository Database. Sci. Rep., 13:13588, 2023. doi:10.1038/s41598-023-40717-1

[6] Framingham Heart Study (FHS): A Landmark Study Identifying Key Risk Factors for Cardiovascular Disease. National Heart, Lung, and Blood Institute, National Institutes of Health, 1948–present. Available at https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs. Accessed April 22, 2025.

[7] Virani, S. S., Newby, L. K., Arnold, S. V., et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the Management of Patients With Chronic Coronary Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. Circulation, 148(9):E9–E119, 2023. doi:10.1161/CIR.0000000000001168

[8] World Health Organisation (WHO), *Cardiovascular Diseases*, Retrieved from Who, June 11, 2021. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)