Predicting Heart Disease with Machine Learning: A Comparative Analysis of Neural Networks, SVM, and Naive Bayes

By Katarina Barbosa

Cardiovascular disease remains the leading cause of death globally, accounting for millions of deaths each year. Early detection and accurate risk assessment are critical for prevention and intervention. This project applies machine learning to predict heart disease presence using clinical and demographic data from the UCI Heart Disease Dataset. The goal was to compare three distinct machine learning approaches: a Neural Network, Support Vector Machine (SVM), and Naive Bayes classifier, to determine which model provides the most reliable and balanced predictions for clinical decision support.

The Data: Heart Disease Dataset

The dataset originates from a comprehensive collection of heart disease studies compiled from multiple medical institutions and made available through the UCI Machine Learning Repository and Kaggle. The dataset contains 918 patient observations with 11 clinical and demographic features capturing cardiovascular health indicators, lifestyle factors, and diagnostic measurements.

The features include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST depression (old peak), and ST slope. The target variable is a binary indicator of heart disease presence, with 508 patients diagnosed with heart disease and 410 without.

Preparing the Data

Before building any models, I implemented a comprehensive data preprocessing pipeline to transform raw clinical data into a format suitable for machine learning. This pipeline consisted of several critical steps: categorical encoding, missing value imputation, feature scaling, and feature selection.

Categorical Encoding

Five categorical variables required encoding for numerical analysis. Sex was encoded as binary (Male=1, Female=0). Chest pain type was encoded ordinally (Typical Angina=1, Atypical Angina=2, Non-Anginal Pain=3, Asymptomatic=4). Resting ECG was encoded as (Normal=0, ST-T Wave Abnormality=1, Left Ventricular Hypertrophy=2). Exercise-induced angina was binary (Yes=1, No=0). ST slope was encoded as (Upsloping=1, Flat=2, Downsloping=3).

Cholesterol Imputation

A critical data quality issue was identified: 172 observations (18.74% of the dataset) contained biologically impossible cholesterol values of zero. These missing values required careful imputation to preserve the integrity of the analysis while retaining the full sample size. This

preprocessing challenge highlighted the importance of domain knowledge in clinical machine learning applications.

The 172 zero cholesterol values presented a significant challenge. I evaluated three imputation strategies: (1) overall median imputation, (2) sex-stratified median imputation, and (3) sex and age group stratified median imputation. Method 3 was selected because it produced distributions closest to the original non-zero cholesterol values while accounting for the biological reality that cholesterol levels vary by sex and age. The imputed values had a mean of 243.24 mg/dL and median of 236.00 mg/dL with standard deviation of 53.56, closely matching the distribution of valid measurements.

Feature Transformation and Selection

All numerical features were standardized using StandardScaler to have zero mean and unit variance. This transformation is critical for distance-based algorithms like SVM and improves convergence for neural networks. Feature importance was assessed using Mutual Information, which measures the dependency between features and the target variable. The analysis revealed that ST slope (MI=0.2393) was the single most predictive feature, followed by chest pain type (MI=0.1531), exercise-induced angina (MI=0.1234), old peak (MI=0.0888), and maximum heart rate (MI=0.0766).

The dataset was split into 80% training (734 samples) and 20% testing (184 samples) subsets, maintaining equal class proportions through stratified sampling. This split provides sufficient data for model training while reserving an adequate holdout set for unbiased performance evaluation.

Building the Models

To capture different modeling paradigms, I implemented three distinct machine learning approaches: a feedforward Neural Network using PyTorch, a Support Vector Machine with multiple kernel options, and a Gaussian Naive Bayes classifier. Each model underwent systematic hyperparameter tuning and cross-validation to ensure optimal performance.

Neural Network Architecture

The Neural Network was implemented in PyTorch with two architecture variants tested: a small network with layers [32, 16] and a medium network with layers [64, 32, 16]. Both architectures used ReLU activation functions and dropout regularization (rate=0.2) to prevent overfitting. The model was trained using the Adam optimizer with learning rate 0.001 and binary cross-entropy loss. Critically, I tested both the full 11-feature set and a reduced 8-feature set containing only the most informative variables. The small network trained on the top 8 features emerged as the best configuration, achieving validation F1-score of 0.8503.

Cross-validation strategy comparison revealed interesting trade-offs. I tested 3-fold, 5-fold, and 8-fold strategies. While 8-fold achieved the highest mean F1-score (0.8674), 3-fold showed the

lowest variance (std=0.0230), and 5-fold offered a middle ground. The 8-fold strategy was selected for its superior discriminative performance despite slightly higher variance.

Support Vector Machine

The SVM implementation tested three kernel functions: linear, radial basis function (RBF), and polynomial. Each kernel underwent hyperparameter tuning via GridSearchCV with 3-fold, 5-fold, and 8-fold validation strategies. For the linear kernel, I tested C values of [0.1, 1, 10]. For RBF, I tested C values [0.1, 1, 10] combined with gamma settings ['scale', 'auto']. For polynomial kernels, I tested C values [0.1, 1, 10] with degrees [2, 3] and gamma ['scale'].

Interestingly, SVM showed opposite feature preferences compared to the Neural Network. The RBF kernel with all 11 features achieved the highest validation F1-score (0.8571) with parameters C=1 and gamma='auto', outperforming the reduced feature set. This demonstrates that different algorithms have different feature requirements. The 3-fold cross-validation strategy was selected for SVM based on stability analysis, achieving cross-validation F1-score of 0.8734 with standard deviation of only 0.0150.

Naive Bayes Classifier

The Gaussian Naive Bayes model was tested with four var_smoothing parameters (1e-9, 1e-8, 1e-7, 1e-6) to control numerical stability. Surprisingly, all smoothing parameters produced identical results, suggesting the default value (1e-9) was sufficient. Like SVM, Naive Bayes performed best with all 11 features, achieving validation F1-score of 0.8199. I tested 3-fold, 5-fold, and 8-fold strategies. The 3-fold cross-validation strategy was selected, yielding cross-validation F1-score of 0.8512 with remarkably low standard deviation of 0.0126, indicating highly stable performance.

Classification Results: Comparing Three Approaches

After comprehensive training and validation, each model was evaluated on the held-out test set of 184 patients. The results revealed subtle but meaningful differences in how each algorithm approaches the heart disease prediction problem.

Test Set Performance

Table 1 Model Performance Comparison on Test Set

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Neural Network	0.8315	0.8381	0.8627	0.8502	0.8998
SVM (RBF)	0.8533	0.8378	**0.9118**	0.8732	0.9024
Naive Bayes	**0.8587**	**0.8519**	0.9020	**0.8762**	**0.9205**

Note: Bold values indicate best performance for each metric.

The results demonstrate remarkably competitive performance across all three models, with differences of less than 3% for most metrics. Naive Bayes emerged as the overall winner, achieving the highest accuracy (85.87%), precision (85.19%), F1-score (87.62%), and ROC-AUC (92.05%). However, SVM achieved the highest recall (91.18%), making it the best choice for minimizing false negatives. This is a critical consideration in medical diagnosis where missing a disease case has severe consequences.

Confusion Matrix Analysis

Table 2 Confusion Matrix Comparison

Model	True Negatives	False Positives	False Negatives
Neural Network	65	17	14
SVM (RBF)	64	18	**9**
Naive Bayes	66	16	10

Note: Bold value indicates fewest false negatives (missed disease cases).

The confusion matrices reveal important trade-offs in error types. SVM produced only 9 false negatives, meaning it missed the fewest disease cases, though at the cost of 18 false positives. Naive Bayes achieved the best balance with 10 false negatives and only 16 false positives, contributing to its superior F1-score. The Neural Network, while competitive, produced 14 false negatives, the highest among the three models.

Comparing Three Machine Learning Paradigms

This project demonstrates that model selection involves more than simply comparing accuracy scores. Each algorithm represents a different machine learning paradigm with distinct strengths, computational requirements, and practical trade-offs.

Feature Requirements and Model Complexity

An unexpected finding was the divergent feature preferences between models. The Neural Network performed best with only the top 8 most informative features (ST slope, chest pain type, exercise angina, old peak, maximum heart rate, age, sex, and cholesterol). This suggests that neural networks can be sensitive to noise from less predictive features and benefit from aggressive feature selection.

In contrast, both SVM and Naive Bayes achieved optimal performance with all 11 features. SVM's ability to operate effectively in high-dimensional spaces through the kernel trick allows it to handle additional features without overfitting. Naive Bayes, despite its strong independence assumption, similarly benefited from the complete feature set, likely because the additional features provide complementary information even if not perfectly independent.

Cross-Validation Strategy Selection

Each model independently selected its optimal cross-validation strategy through empirical testing. The Neural Network preferred 8-fold cross-validation, achieving the highest mean F1-score despite slightly higher variance. This suggests that neural networks benefit from larger training sets available with higher fold counts. Both SVM and Naive Bayes selected 3-fold cross-validation, prioritizing stability (lower variance) over marginal gains in mean performance. This reflects the different bias-variance trade-offs inherent to each algorithm family.

Computational Efficiency

Training time varied dramatically across models. Naive Bayes was by far the fastest, completing training and evaluation in approximately 30 seconds. SVM required 1-2 minutes including hyperparameter tuning via GridSearchCV. The Neural Network was the most computationally expensive, requiring 3-4 minutes for architecture search, cross-validation comparison, and training. For real-time clinical deployment or resource-constrained settings, Naive Bayes offers compelling efficiency advantages with minimal performance sacrifice.

Feature Importance and Clinical Implications

Understanding which features drive predictions is critical for clinical interpretability and trust. The Mutual Information analysis consistently identified ST slope as the single most predictive feature (MI=0.2393), which aligns with clinical knowledge. The ST segment represents the interval between ventricular depolarization and repolarization on an electrocardiogram, and abnormalities in ST slope are well-established indicators of myocardial ischemia and coronary artery disease.

The second most important feature was chest pain type (MI=0.1531), which makes biological sense as different pain patterns reflect distinct pathophysiological mechanisms. Asymptomatic presentations, while counterintuitive, are associated with higher risk as they often indicate advanced disease or atypical presentations in vulnerable populations. Exercise-induced angina (MI=0.1234) and old peak (ST depression, MI=0.0888) ranked third and fourth, both reflecting cardiac stress response during exertion.

These findings validate the models' clinical relevance. The algorithms learned patterns consistent with established cardiovascular pathophysiology rather than identifying spurious correlations. This alignment between machine learning feature importance and medical knowledge increases confidence in model predictions and facilitates clinical adoption.

Conclusion

Through this comprehensive comparison, I developed and evaluated three distinct machine learning approaches for heart disease prediction. After rigorous preprocessing, hyperparameter optimization, and cross-validation, the results demonstrate that Naive Bayes provides the best overall performance for this clinical prediction task, achieving 85.87% accuracy, 87.62% F1-score, and 92.05% ROC-AUC. Its combination of strong predictive performance, computational efficiency, and probabilistic interpretability makes it the recommended choice for deployment in clinical decision support systems.

However, the choice of model should depend on the clinical context. For scenarios where minimizing missed diagnoses is paramount, such as screening high-risk populations, SVM's superior recall (91.18%) makes it the preferred alternative despite slightly lower overall accuracy. The Neural Network, while competitive, offers limited advantages to justify its computational overhead for this dataset size. Its performance would have been improved had the dataset been much larger with at least a minimum of 10,000 values.

This study reinforces several key principles in applied machine learning for healthcare. First, thoughtful preprocessing, particularly handling missing data with domain knowledge, is as important as model selection. Second, different algorithms have different feature requirements and optimal configurations that must be determined empirically. Third, evaluation must consider multiple metrics aligned with clinical priorities rather than focusing solely on accuracy. Finally, the most complex model is not always the best model, simplicity, interpretability, and efficiency matter greatly in real-world deployment.

With all three models achieving over 85% accuracy and 87% F1-scores, this project demonstrates that machine learning can provide reliable cardiovascular risk assessment to support clinical decision-making. By combining routine clinical measurements with data-driven modeling, such systems can augment physician expertise, enable earlier intervention, and ultimately improve patient outcomes for one of the world's leading causes of mortality.

References

Ali, M. M., Paul, B. K., Ahmed, K., et al. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, 136, 104672.

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*, 319(13), 1317-1318.

Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/45/heart+disease

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. *Journal of the American College of Cardiology*, 69(21), 2657-2664.

Roth, G. A., Mensah, G. A., Johnson, C. O., et al. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*, 76(25), 2982-3021.