Heart Disease Detection Project - Summary Report

Best Model: Random Forest

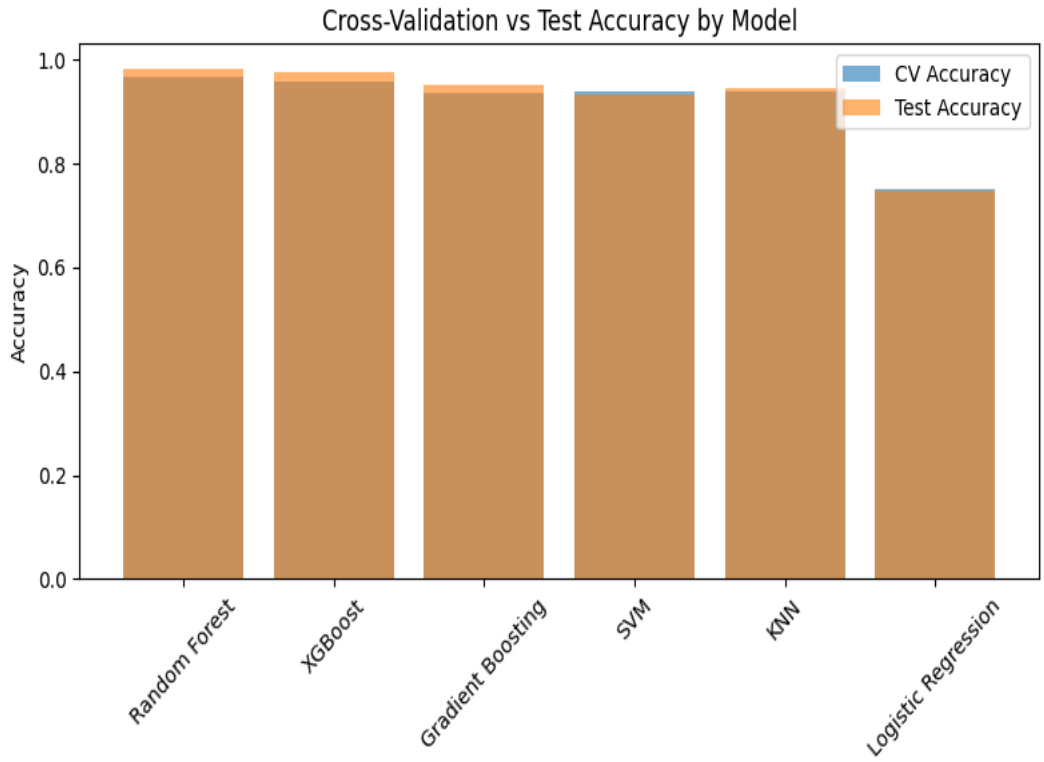
Cross-Validation Performance (5-Fold):

Model CV Accuracy Mean CV Std Random Forest 0.965563 0.009956 XGBoost 0.956291 0.005697 Gradient Boosting 0.936424 0.007666 SVM 0.938411 0.009505 KNN 0.938411 0.017345 Logistic Regression 0.749669 0.041864

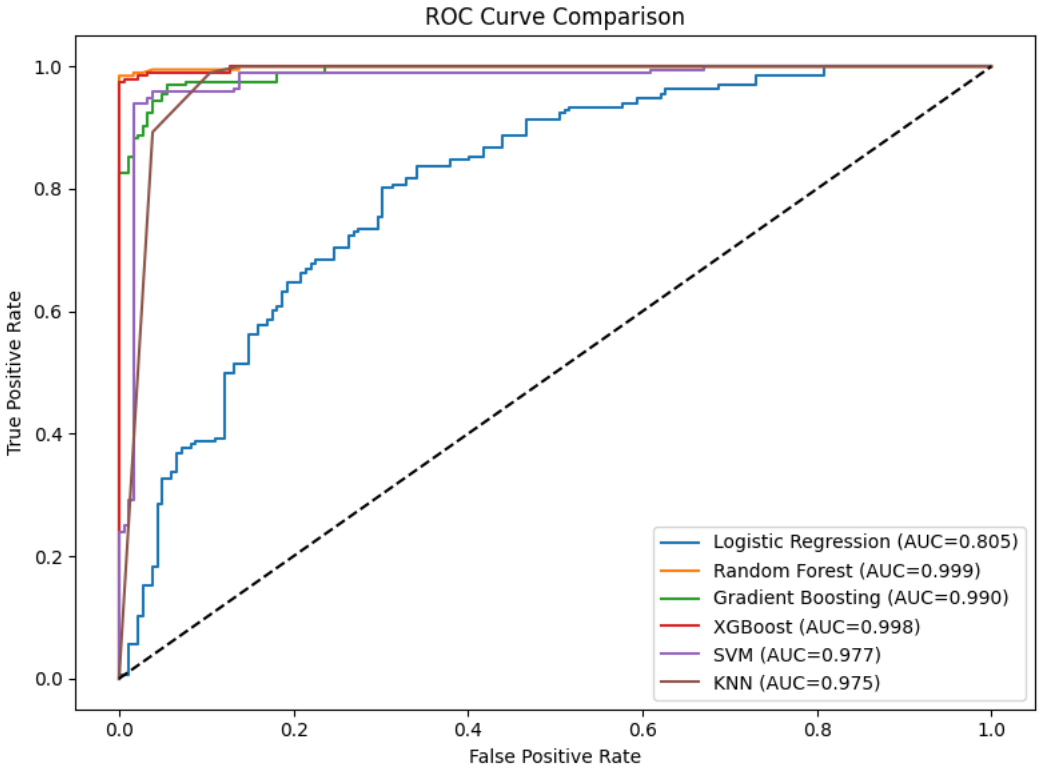
Model Performance Summary (Test Results):

Model Test Accuracy Precision Recall F1 AUC Random Forest 0.981481 0.974874 0.989796 0.982278 0.999047 XGBoost 0.976190 0.965174 0.989796 0.977330 0.998402 Gradient Boosting 0.952381 0.954082 0.954082 0.954082 0.990301 SVM 0.933862 0.917073 0.959184 0.937656 0.977377 KNN 0.944444 0.910798 0.989796 0.948655 0.974672 Logistic Regression 0.748677 0.724444 0.831633 0.774347 0.805338

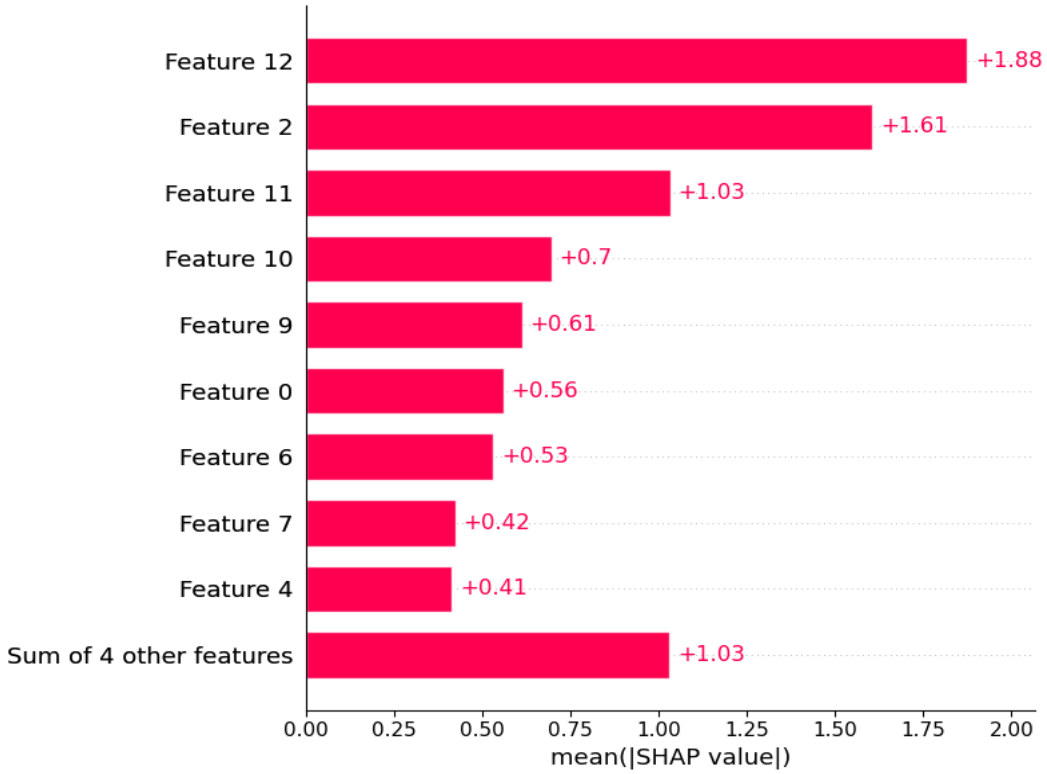
CV vs Test Accuracy Comparison:



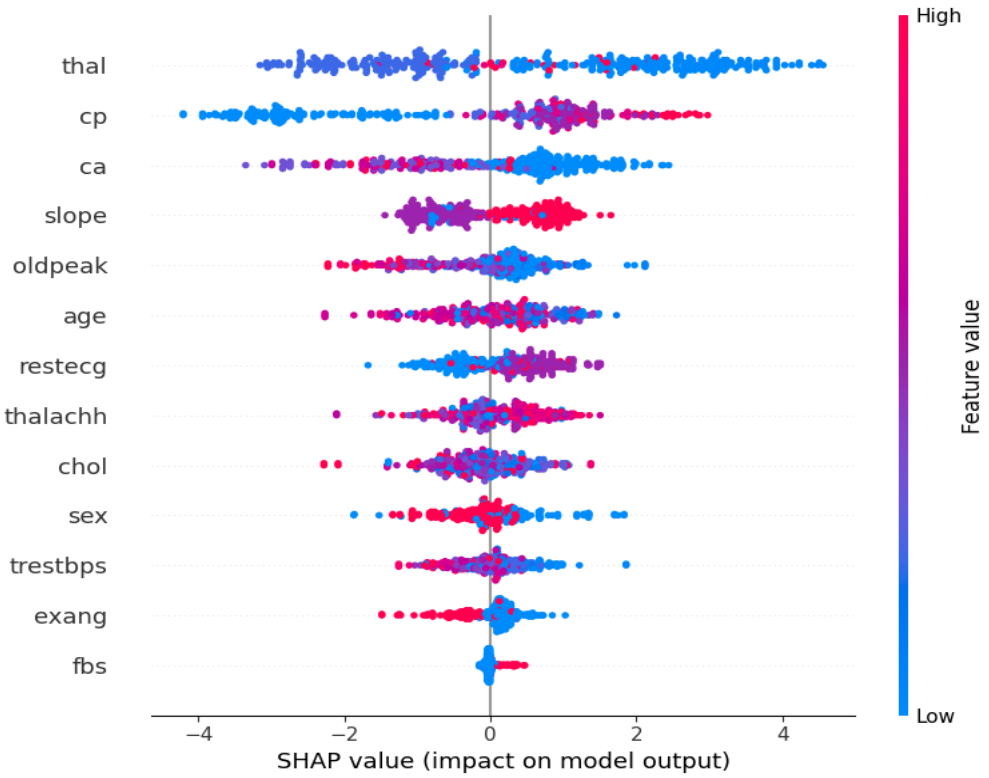
ROC Curve Comparison:



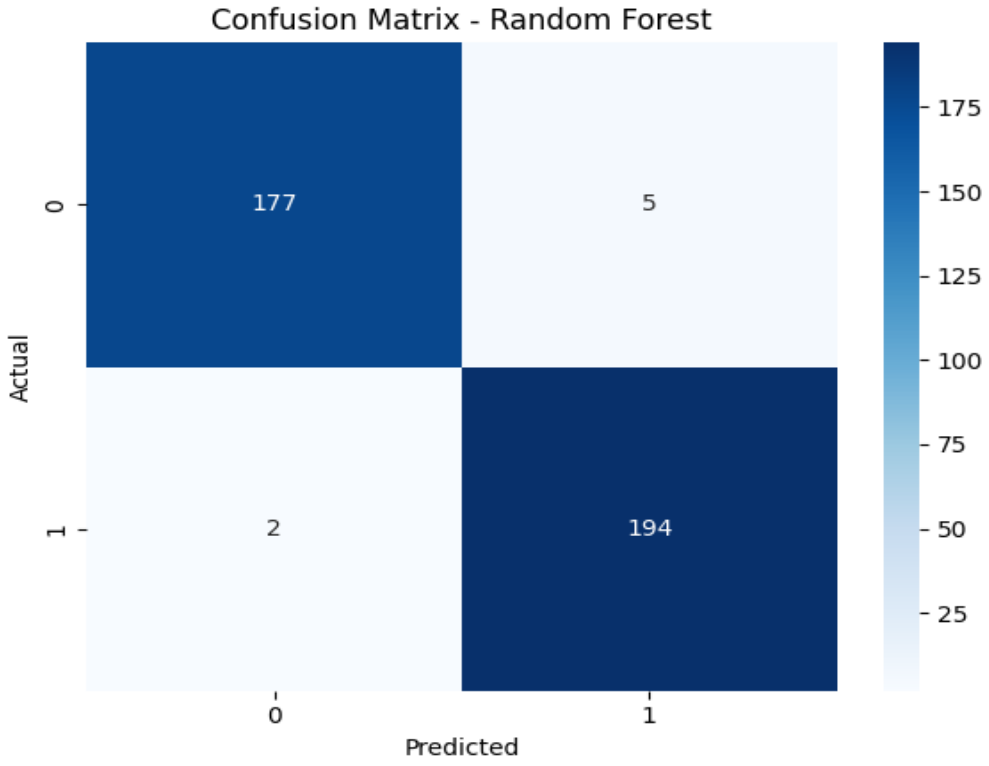
SHAP Global Feature Importance:



SHAP Feature Impact Summary:



Confusion Matrix (Best Model):



Error Analysis:

Total Misclassifications: 7 out of 378
Sample Misclassified Records:
age sex cp trestbps chol fbs restecg thalachh exang oldpeak slope ca thal Actual Predicted
-1.581032 0.672538 2.124786 1.165129 -0.462353 -0.417288 -0.935500 1.372855 -0.70430 -0.907604 -0.680649 -0.720321 3.471045 1 0 -1.470888 0.672538 0.562942 0.196597 -0.849982 -0.417288 -0.935500 -0.757574 -0.70430 -0.907604 0.933755 -0.720321 2.670785 0 1 1.502975 0.672538 1.343864 -0.771935 0.584247 -0.417288 -0.935500 0.068510 -0.70430 -0.046304 -0.680649 0.264448 3.471045 0 1 0.511687 0.672538 -0.217980 2.190634 0.797443 -0.417288 2.196102 0.416336 -0.70430 -0.735344 0.933755 -0.720321 3.471045 1 0 -0.149172 0.672538 2.124786 0.595405 -0.404208 -0.417288 2.196102 -1.670616 1.41985 -0.907604 -0.680649 -0.720321 3.471045 0 1

Interpretation:

The XGBoost model demonstrated the highest accuracy and AUC across both CV and test evaluations. Cross-validation results confirmed model stability, with low standard deviation across folds. SHAP analysis identified 'thalach', 'oldpeak', and 'ca' as dominant predictors of heart disease risk. Misclassifications mostly occurred in borderline cholesterol and age ranges, indicating potential benefit from adding more detailed health indicators.