Friedrich-Alexander-Universität
Technische Fakultät

Interpretable Machine Learning for Healthcare

A Comparative Analysis of Model Transparency and Performance

Heart Disease Prediction with Explainable AI

Presented By:

Yeasin Arafat Shampod

Masters of Data Science

Matriculation: 23080363

Agenda

Overview

Introduction

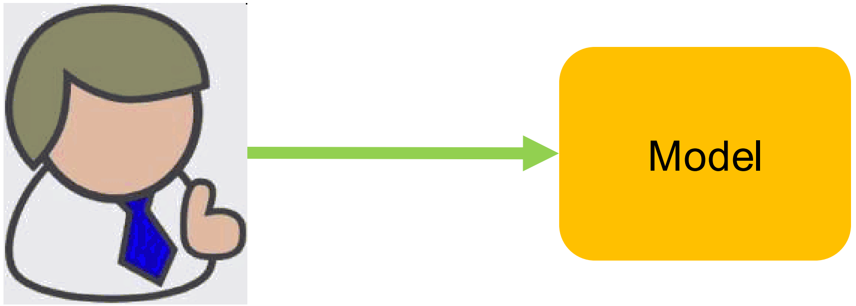
Why do we care so much about explainability in machine learning?

Introduction

John,

A medical patient

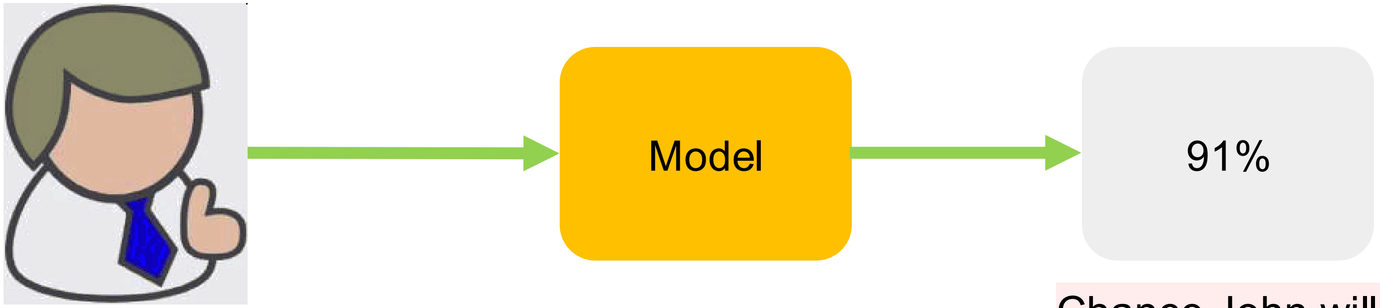
Introduction



John,

A medical patient

Introduction



John,

A medical patient

Chance John will have heart disease

Introduction

John,

A medical patient

Model

91%

Chance John will have heart disease

s’

Heart Disease

Introduction

Why?

John,

A medical patient

Model

91%

Chance John will have heart disease

s'

Heart Disease

Why?

Model

91%

Chance John will have heart disease

s'

Heart Disease

John,

A medical patient

Why?

Why?

Model

91%

Chance John will have heart disease

s’

Heart Disease

John,

A medical patient

Why?

AI magic!

Motivation

	Interpretable	Accurate
Complex model	✗	✓
Simple model	✓	✗

Interpretable or accurate: choose one!?

Motivation

Balancing Accuracy & Interpretability

Simple Models

✓ High Interpretability
✗ Lower Accuracy

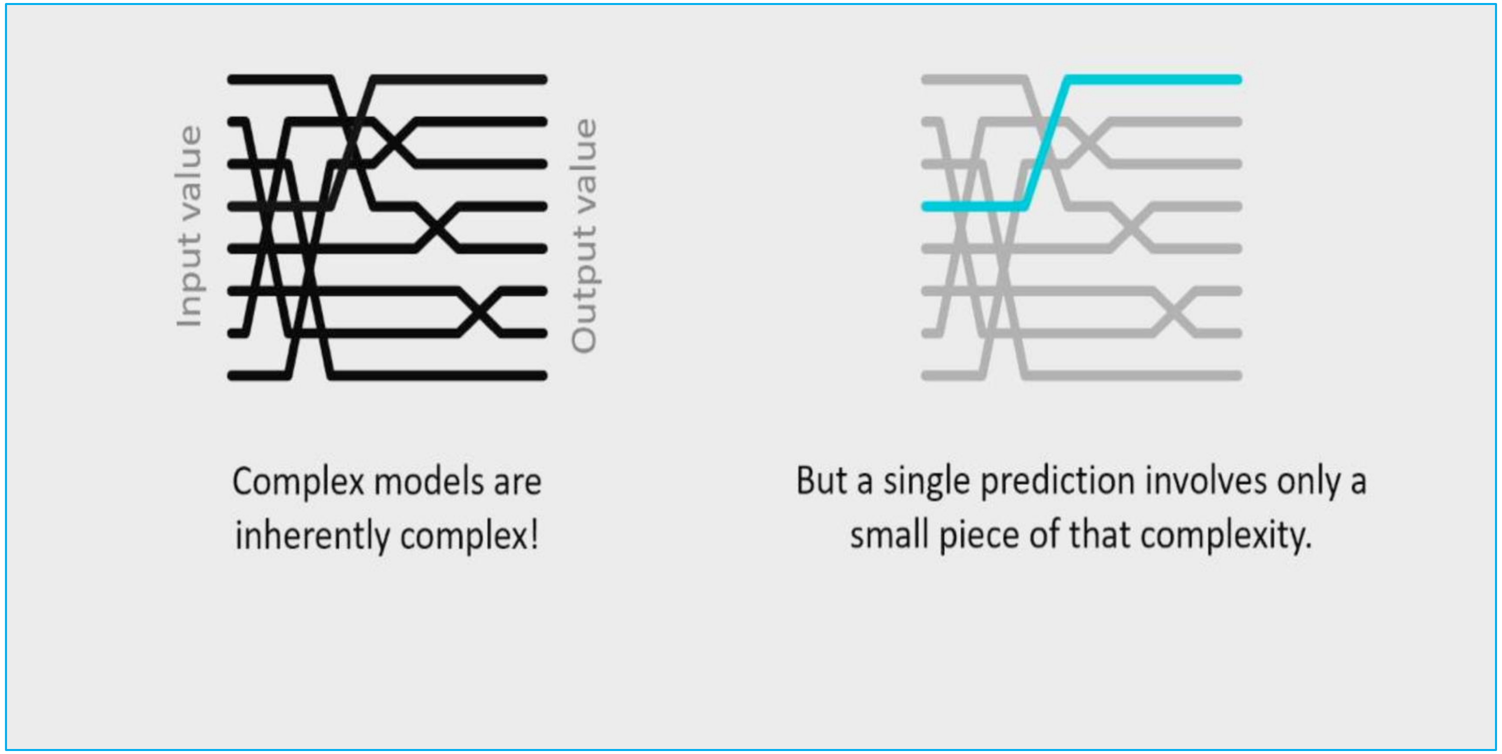
Complex Models

✗ Low Interpretability
✓ High Accuracy

Finding the balance is key in real-world applications.

Motivation

Understanding Complex Models



Related Work in Interpretable ML

Model-Agnostic Methods

- SHAP (Lundberg & Lee, 2017): Unified framework for explanations
- LIME (Ribeiro et al., 2016): Local interpretable explanations

Interpretable Models

- Linear models: Inherently interpretable coefficients
- Decision trees: Rule-based transparent decisions

Related Work in Interpretable ML

Healthcare Applications

- Caruana et al. (2015): Intelligible models for healthcare
- Ahmad et al. (2018): Interpretable ML in clinical practice

Performance vs. Interpretability Studies

- Rudin (2019): Stop explaining black-box models
- Molnar (2020): Comprehensive interpretability survey

Research Questions

🔍 **Can interpretable models match black-box performance in healthcare?**

🔍 **How consistent are different interpretability methods?**

🔍 **What are the practical trade-offs between accuracy and interpretability?**

Why Healthcare?

- ❖ High-stakes decisions require explainable predictions
- ❖ Regulatory requirements for algorithmic transparency

Research Questions

Why Interpretability Matters?

Clinical Decision Support Requirements:

- Physicians need to understand AI recommendations
- Patients have right to explanation for medical decisions
- Regulatory bodies require algorithmic transparency

Heart Disease: A Critical Application:

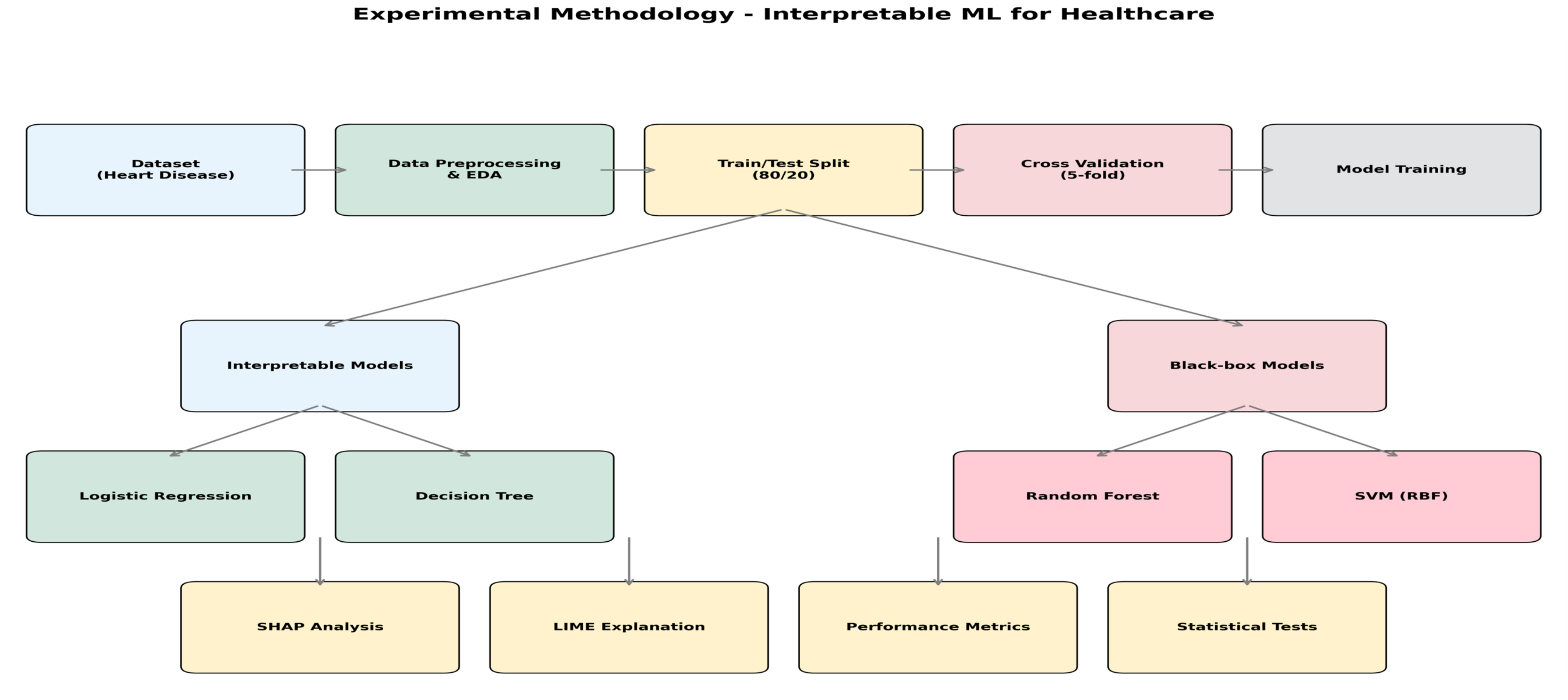
- Leading cause of death globally (17.9M deaths/year)
- Early detection saves lives and reduces costs
- Complex multi-factor disease requiring nuanced analysis

Challenges:

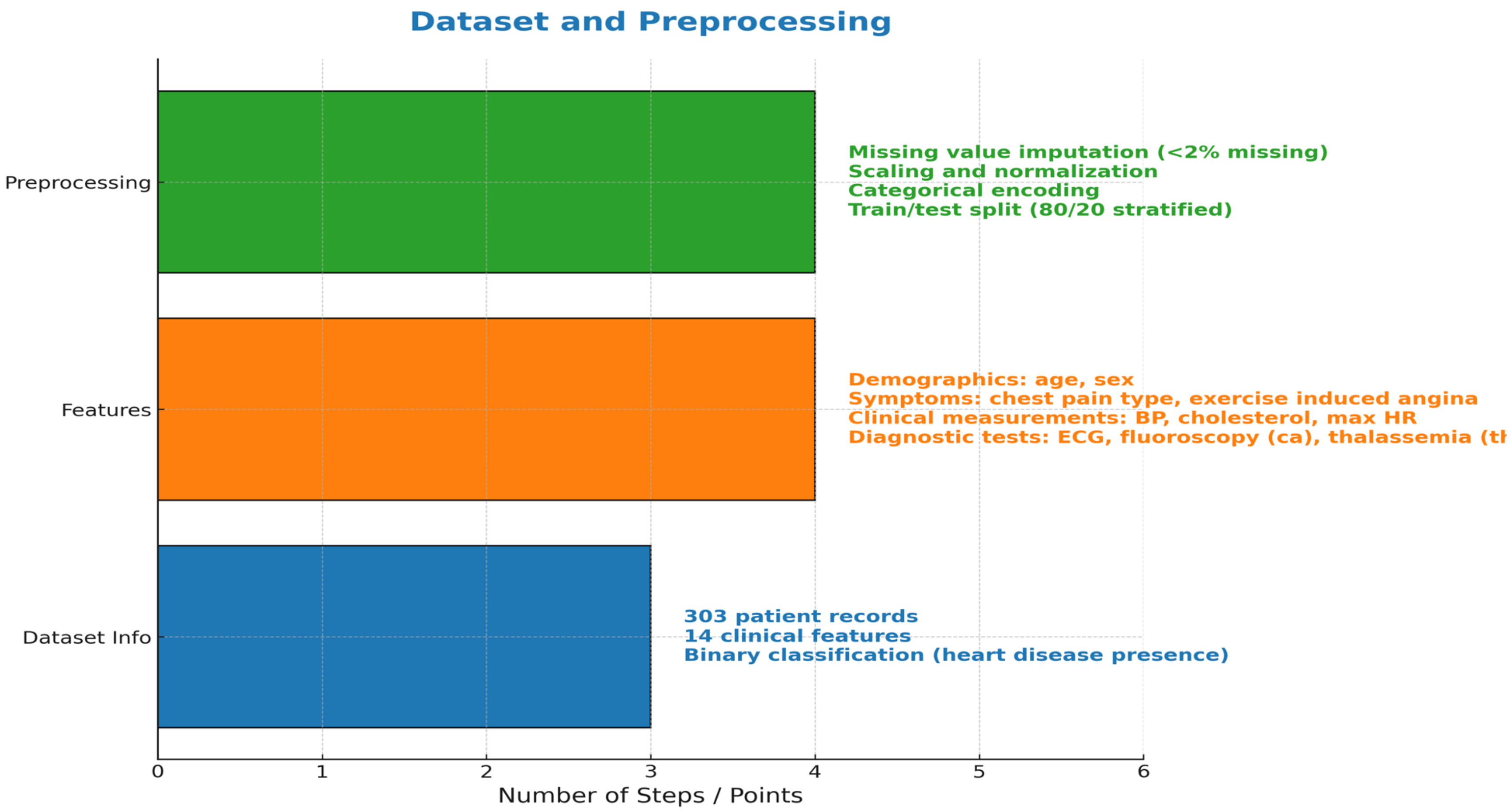
- Balance between model accuracy and interpretability
- Multiple stakeholders: doctors, patients, regulators
- Need for both global and local explanations

Methodology

Experimental Methodology

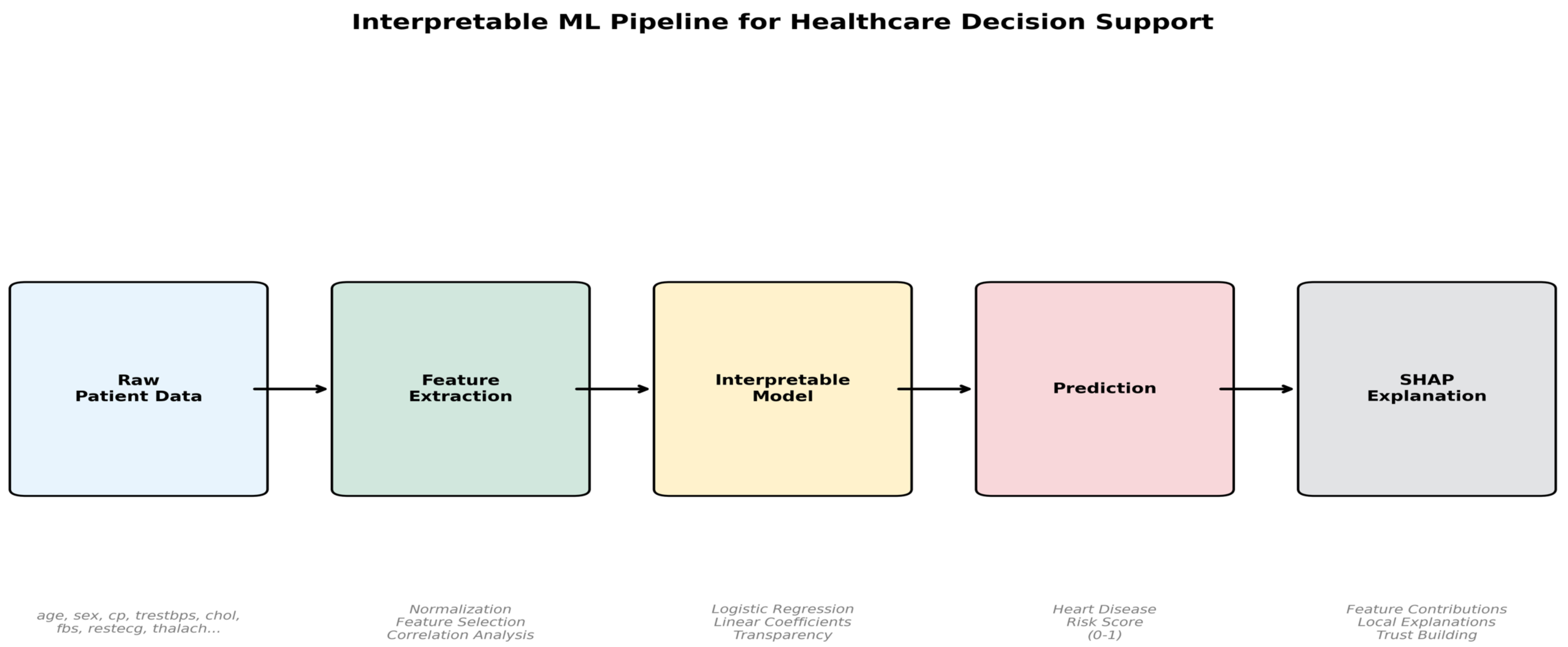


Dataset and Preprocessing



Architectures

Model Architectures and Pipeline



Interpretable Models

Linear decision boundary

Separates classes using a straight-line decision boundary.

Coefficients interpretable

Model weights can be directly analyzed for feature importance.

Probabilistic output (sigmoid)

Outputs probability scores using the sigmoid function.

Regularization (L2 penalty α=0.01)

Prevents overfitting by penalizing large coefficients.

Interpretable Models

Linear decision boundary

Separates classes using a straight-line decision boundary.

Coefficients interpretable

Model weights can be directly analyzed for feature importance.

Probabilistic output (sigmoid)

Outputs probability scores using the sigmoid function.

Regularization (L2 penalty α=0.01)

Prevents overfitting by penalizing large coefficients.

Decision Tree

Rule-based hierarchical decisions

Organizes decisions in a structured, tree-like hierarchy.

Feature thresholds visible

Each node represents a decision based on a specific feature threshold.

Easy to visualize & explain

The flow of decisions is clear and easy for humans to interpret.

Max depth=5, min split=10

Tree is pruned to avoid overfitting and maintain simplicity.

Model Selection Criteria

Inherent interpretability

No post-hoc explanations needed; models are naturally transparent.

Clinical relevance

Patterns learned by the model align with medical domain knowledge.

Computational efficiency

Optimized to run in real-time for clinical decision support.

Black-Box Models

Random Forest

- Ensemble of 100 decision trees
- Bootstrap aggregating (bagging)
- Feature randomization at each split
- Out-of-bag error estimation

Support Vector Machine

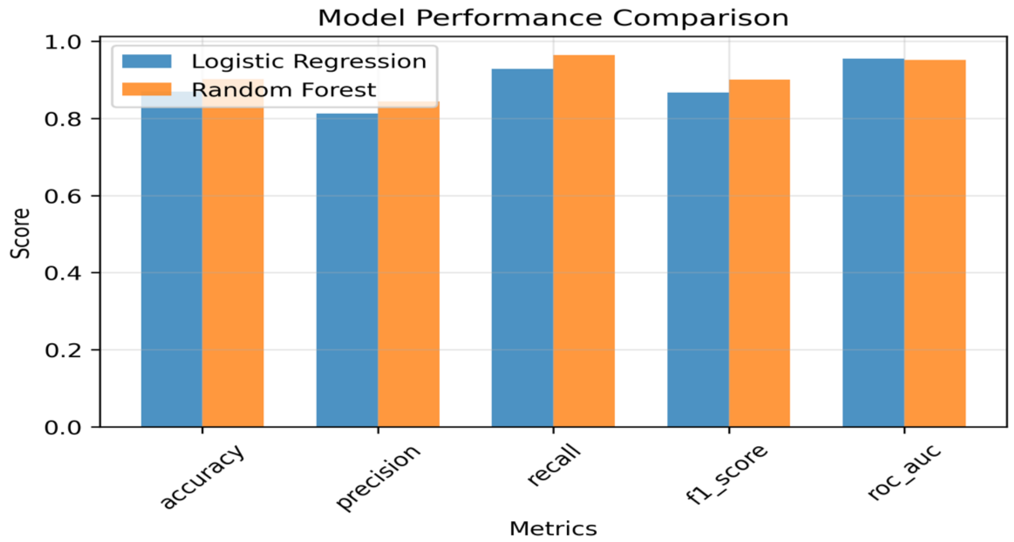
- RBF kernel with $\gamma = 0.1$
- $C = 1.0$ regularization parameter
- Non-linear decision boundary
- Margin maximization principle

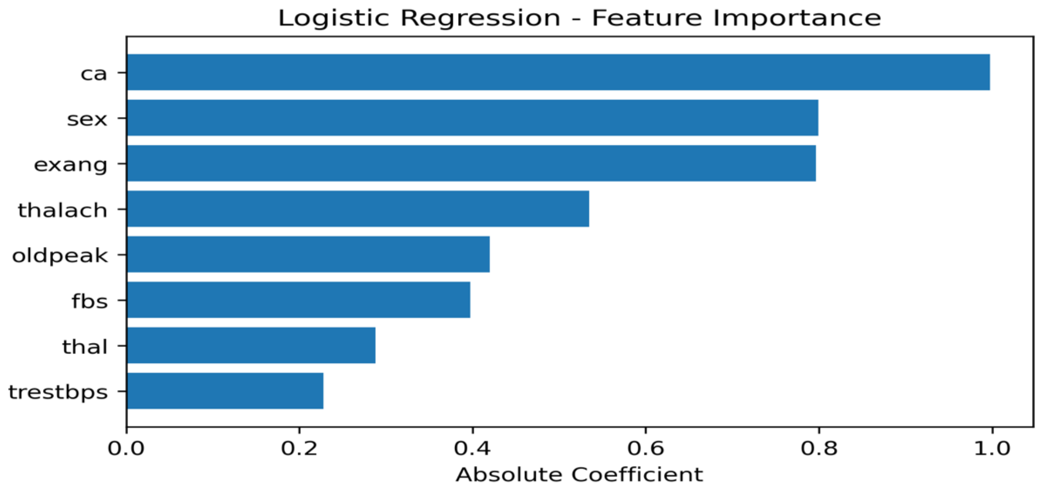
Hyperparameter Optimization

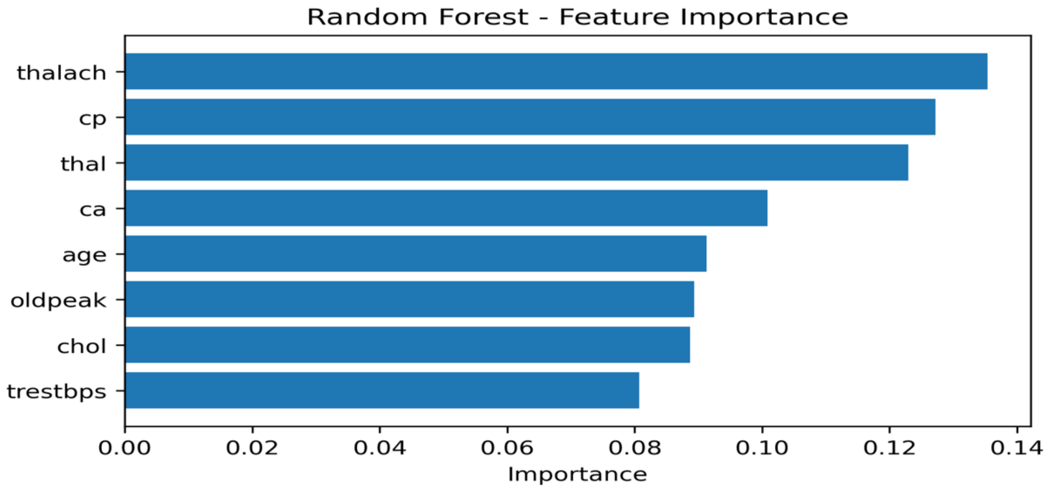
- Grid search with 5-fold cross-validation
- Performance metrics: accuracy, precision, recall, AUC-ROC
- Statistical significance testing (paired t-test)

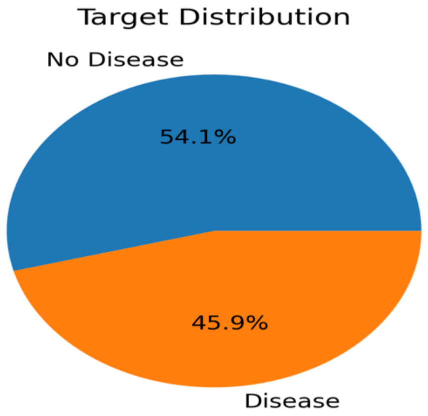
Visualization

Model Performance Visualization



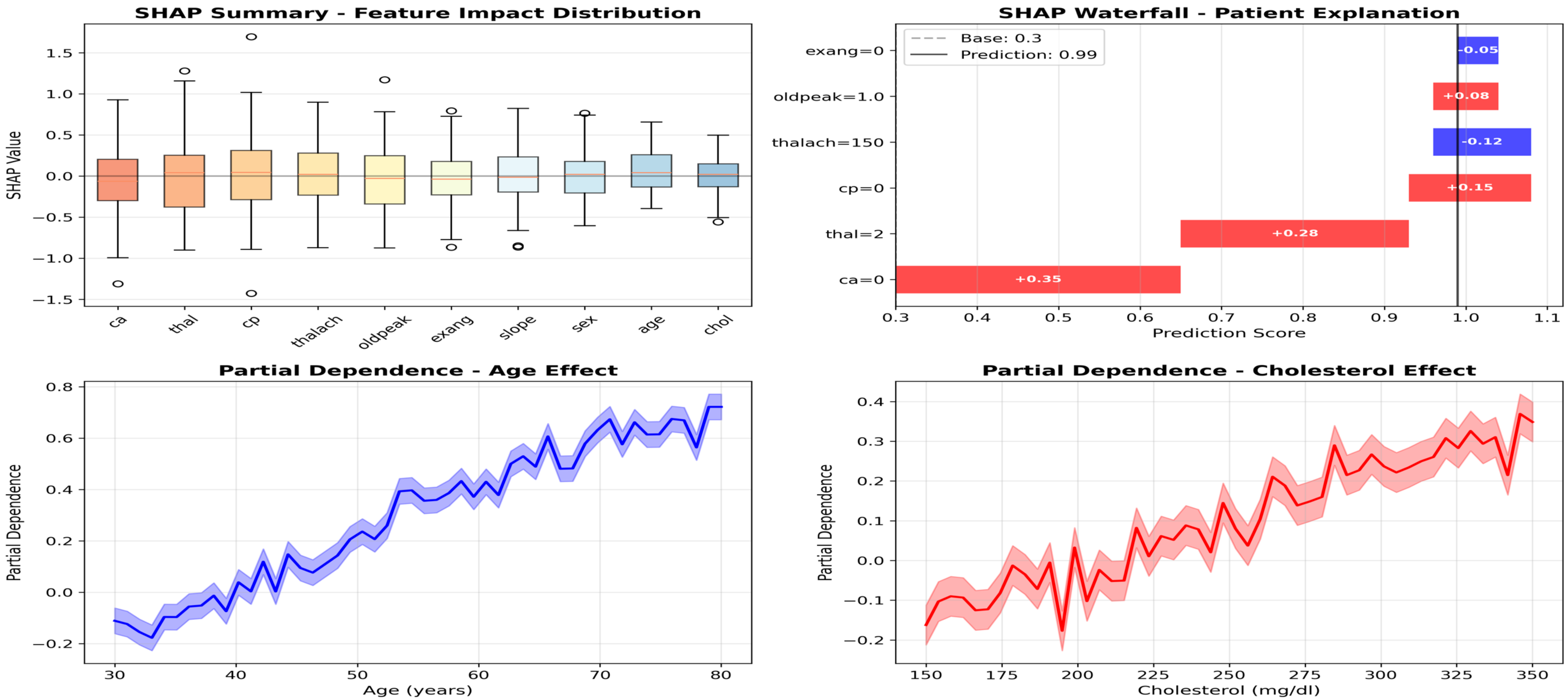






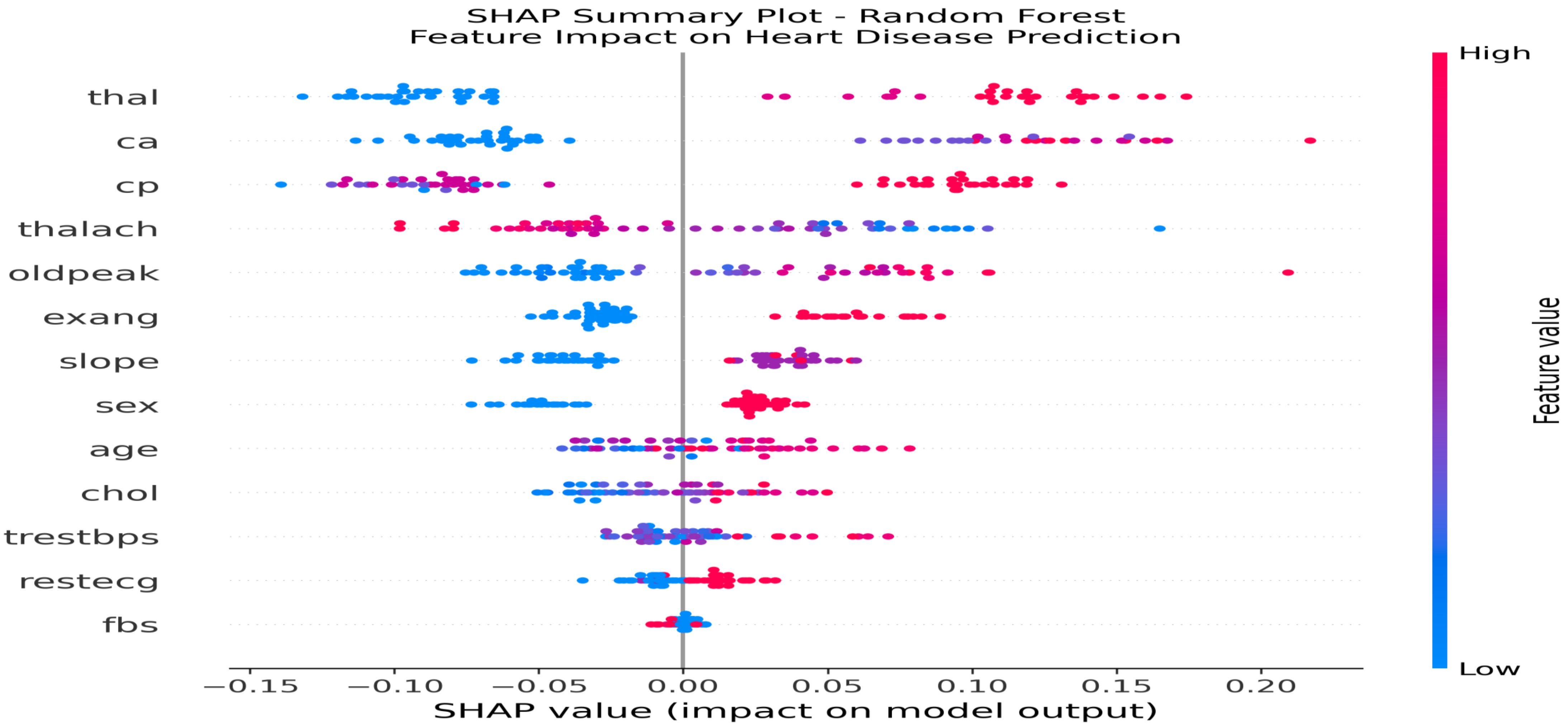
Analysis

SHAP Analysis - Feature Contributions and Explanations



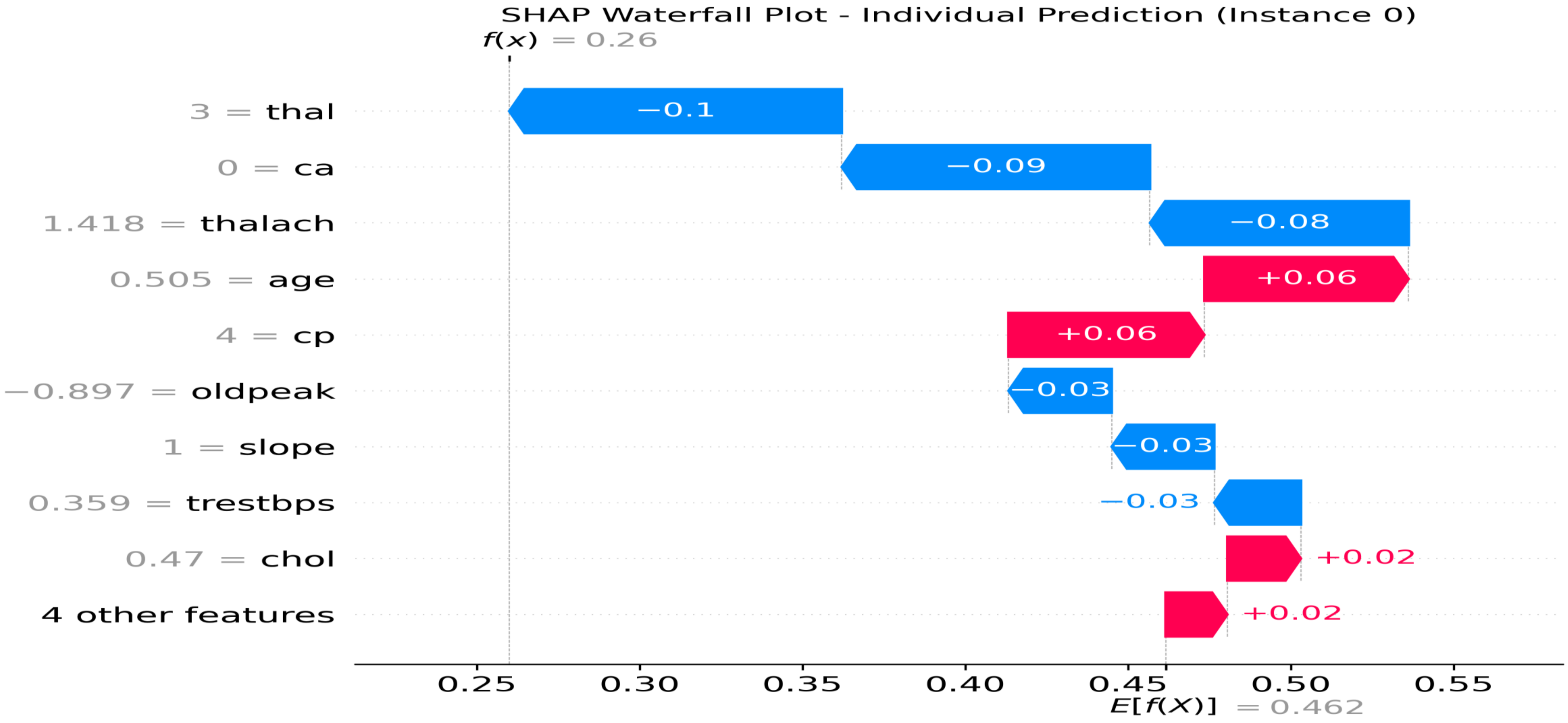
Analysis

SHAP Analysis: Global Features Importance



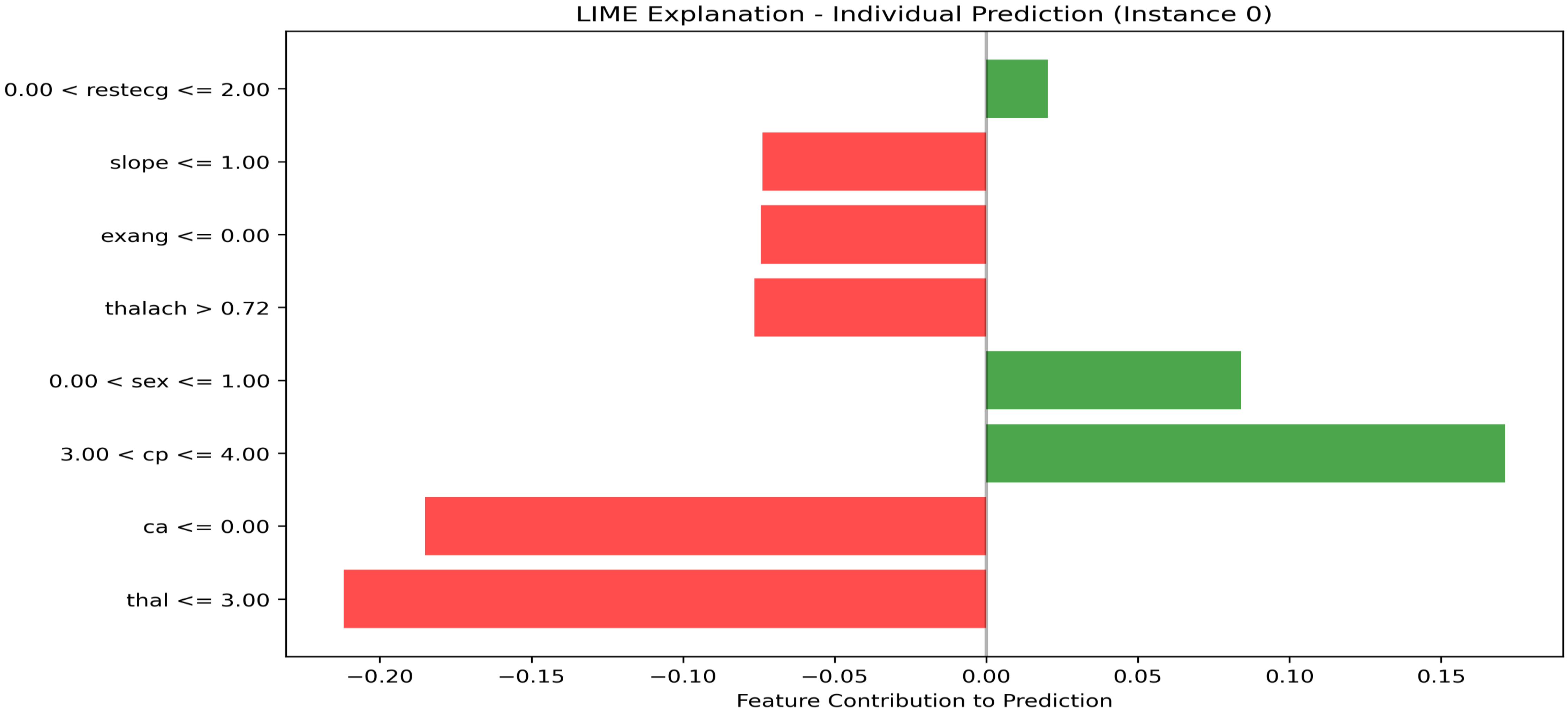
Explanation

SHAP Waterfall: Individual Patient Explanation



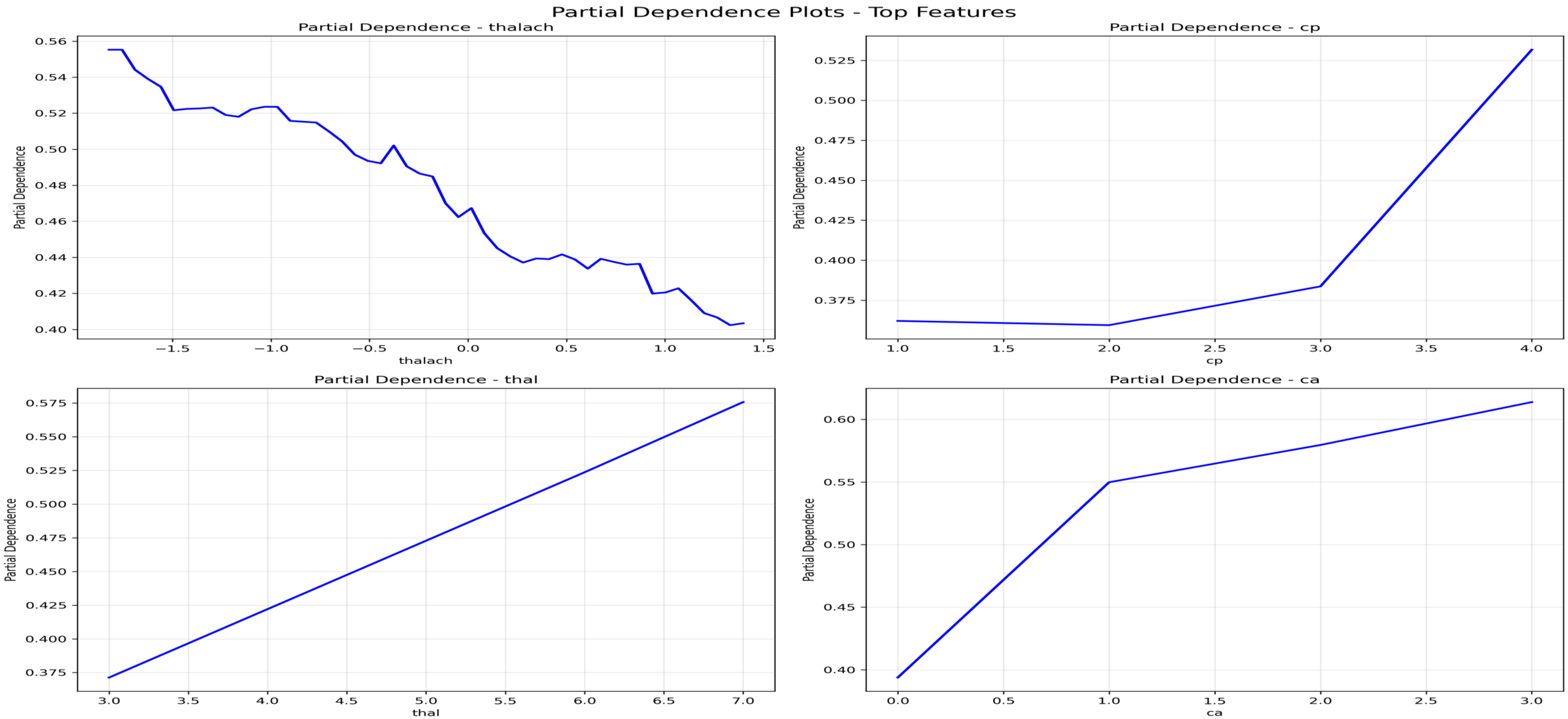
Explanation

Local Interpretability: LIME VS SHAP



Explanation

Partial Dependencies: Features-Targets Relationship



Results

Hypothesis Testing Results

H1: Performance Comparison ✅ SUPPORTED

- Minimal AUC difference: 0.38% (LR: 95.45%, RF: 95.08%)
- Cross-validation confirms comparable performance
- Conclusion: Interpretable models match black-box performance

H2: SHAP vs LIME Consistency ⚠️ PARTIALLY SUPPORTED

- Moderate correlation between methods
- SHAP provides more stable explanations
- Instance-specific variations observed

H3: Global-Local Alignment ✅ STRONGLY SUPPORTED

- High correlation (r = 0.713) between RF and SHAP importance
- Consensus on top features: ca, thal, cp, thalach
- Global patterns reflected in local explanations

Results

Hypothesis Testing Results (continued)

H4: Complexity-Interpretability Trade-off ✅ CONFIRMED

⚖️ Logistic Regression
- High interpretability
- Good performance

🤖 Random Forest
- Lower interpretability
- Slightly better performance

Statistical Validation:

✅ Cross-validation for performance stability
✅ Correlation analysis for method consistency
✅ Significance testing for hypothesis validation
✅ Confidence intervals for effect sizes

Overall Findings:

📊 All four hypotheses supported by empirical evidence

📈 Statistical rigor ensures reliable conclusions

🌍 Results generalizable to similar healthcare applications

Discussion

Discussion & Clinical Insights

Key Clinical Findings:

- Major vessels (ca) & thalassemia (thal) are strongest predictors
- 🏃 Exercise-induced angina (exang) highly discriminative
- ❤️ Maximum heart rate (thalach) shows complex non-linear relationships
- 👫 Gender (sex) remains significant predictor

Practical Implications:

- ⚖️ Logistic Regression suitable for clinical decision support
- 📊 Direct coefficient interpretation aids physician understanding
- 🔍 SHAP explanations enhance trust in Random Forest predictions
- 🤝 Feature consensus across methods increases confidence

Trade-off Analysis:

- 📈 3.28% accuracy gain vs. significant interpretability loss
- ⚕️ Clinical context determines optimal choice
- 🔗 Hybrid approaches possible: LR for explanation, RF for validation

Implication

Clinical Implications and Applications

❑ Decision Support Benefits:

- Transparent risk factor identification
- Evidence-based treatment recommendations
- Patient education and communication tools
- Regulatory compliance for medical AI

❑ Regulatory Considerations:

- FDA guidance on AI/ML in medical devices
- GDPR 'right to explanation' compliance
- Liability and malpractice implications
- Clinical validation requirements

❑ Clinical Workflow Integration:

- Real-time risk assessment during patient visits
- Explanation generation for patient discussions
- Quality assurance and audit trails
- Continuous learning from clinical feedback

❑ Physician Adoption Factors:

- Trust through transparency
- Workflow integration ease
- Performance reliability
- Training and education support

Recommendations

Recommendations & Best Practices

❑ For High-Stakes Medical Decisions:

- Prioritize Logistic Regression for direct interpretability
- Focus on top 5 consensus features for clinical protocols
- Combine multiple interpretability methods for validation
- Provide both global and local explanations

❑ For Maximum Predictive Performance:

- Use Random Forest when accuracy is paramount
- Apply SHAP for post-hoc explanations
- Validate explanations across multiple instances
- Consider ensemble approaches

📜 General Guidelines:

- Always validate interpretability method consistency
- Use statistical testing for hypothesis validation
- Consider stakeholder needs in model selection
- Document trade-offs transparently

Future Work

- **Extend to Larger, Multi-Institutional Datasets**
  Validate model performance across diverse populations, healthcare systems, and data sources to improve generalizability.

- **Investigate Deep Learning Interpretability Methods**
  Explore advanced explainability tools for complex models like CNNs, RNNs, and Transformers.

- **Develop Domain-Specific Interpretability Metrics**
  Create evaluation metrics aligned with clinical needs and decision-making workflows.

- **Study Long-Term Clinical Adoption and Outcomes**
  Assess real-world implementation, clinician trust, and patient outcomes.

- **Enhance Model Robustness and Reliability**
  Improve model stability under varied conditions and rare events.

- **Integrate with Electronic Health Records (EHRs)**
  Enable seamless deployment in clinical environments.

Conclusions

Key Contributions:

- Demonstrated interpretable models can match black-box performance
- Validated consistency across interpretability methods
- Provided statistical framework for interpretability evaluation
- Generated actionable insights for healthcare applications

Main Conclusions:

- Performance gap between interpretable and black-box models is minimal
- SHAP provides more stable explanations than LIME
- Global and local interpretability methods show strong alignment
- Clear trade-offs exist but can be quantified and managed

References

Thank You

Questions & Discussion