**Methodology:**

**"Adaptive Ensemble Feature Selection and Genetic Algorithm-Tuned Ensemble Model for Robust Heart Disease Prediction"**

**Step 1: Dataset Collection & Integration**

**1.1 Select Global Datasets**

* Source from:
  + **UCI** (Cleveland, Statlog, Hungarian, Long Beach VA)
  + **Kaggle** (e.g., Framingham, Heart Failure, Cardiovascular Disease)
* Ensure:
  + Common features (e.g., age, cholesterol, chest pain, blood pressure)
  + Combine into a **unified dataset** with harmonized feature names

**1.2 Preprocessing**

* Handle **missing values**:
  + KNN Imputer for numerical
  + Mode or correlation-based imputation for categorical
* Detect & remove **outliers** using Isolation Forest or Z-score
* Normalize features (MinMaxScaler or RobustScaler)

**Step 2: Ensemble Feature Selection (EFS)**

**2.1 Apply Multiple Feature Selection Techniques**

* **Filter methods**: Information Gain, Chi-Square, Fisher Score
* **Embedded methods**: LASSO, Random Forest Importance
* **Wrapper methods**: Recursive Feature Elimination (RFE)

**2.2 Ensemble the Feature Ranks**

* Normalize scores (0–1 scale)
* Use **mean rank aggregation** or **majority voting**
* Select top *k* features (experiment with 5, 7, 10)

**Step 3: Handle Class Imbalance**

**3.1 Analyze Class Distribution**

* Calculate class imbalance ratio

**3.2 Apply Adaptive Sampling**

* If imbalance ratio < 1:3 → use **SMOTE + Tomek Links**
* If severe imbalance → use **ADASYN**
* Embed sampling within CV folds to prevent leakage

**Step 4: Model Design**

**4.1 Base Classifiers**

* Random Forest
* XGBoost
* LightGBM
* Logistic Regression

**4.2 Genetic Algorithm for Hyperparameter Optimization**

* Use **GA to tune parameters**:
  + Define a **fitness function**: e.g., F1-score or AUC on validation set
  + Apply:
    - Selection: Tournament
    - Crossover: Uniform or one-point
    - Mutation: Random bit or Gaussian
* Perform **k-fold CV** in each fitness evaluation

**Step 5: Ensemble Learning**

**5.1 Build Final Ensemble**

* **Soft voting** or **stacking**:
  + Meta-model (e.g., Logistic Regression) for stacking
* Compare ensemble vs base classifiers

**Step 6: Evaluation Metrics**

**6.1 Model Performance**

* Accuracy
* Precision
* Recall (Sensitivity)
* Specificity
* F1-Score
* ROC-AUC
* PR-AUC (especially important for imbalance)

**6.2 Robustness Testing**

* Inject Gaussian noise (simulate sensor noise)
* Evaluate feature importance stability using SHAP

**Step 7: Explainability Analysis**

**7.1 SHAP Analysis**

* Global feature importance
* Per-instance explanation (why this patient was predicted as high-risk)

**Step 8: Cross-Dataset Generalization (Optional but Valuable)**

* Train on Dataset A (e.g., UCI)
* Test on Dataset B (e.g., Kaggle or holdout from merged dataset)
* Evaluate generalization (transferability)