Research Proposal

Title: *Predicting Structural Heart Disease from ECG Data Using Machine Learning and Explainable Recommendation System*

---

1. Introduction

Structural Heart Disease (SHD) represents a spectrum of abnormalities including valve disorders, reduced ventricular function, and wall thickening. Echocardiography (echo) remains the gold standard for diagnosis, but it is costly, requires expertise, and is not always available in resource-limited settings.

In contrast, Electrocardiography (ECG) is inexpensive, widely available, and quick to perform, yet is often underutilized in detecting SHD. This research proposes to leverage ECG-derived features with machine learning to predict echocardiogram-confirmed SHD and to design an interpretable recommendation system that can support clinical decision-making.

---

2. Dataset Description

The dataset **"Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs 1.0.0"** links ECG-derived measurements with echocardiographic outcomes. It contains:

- **Demographics:** Age, sex, acquisition year.
- **ECG Features:** Ventricular rate, atrial rate, PR interval, QRS duration, QTc interval.
- **Echocardiography Flags (binary):** Indicate moderate or greater severity of conditions such as:
    - Reduced left ventricular ejection fraction (LVEF ≤ 45%)
    - Valve stenosis or regurgitation
    - Wall thickening
    - Right ventricular dysfunction
    - Pericardial effusion
    - Pulmonary hypertension

- **Echocardiography Measurements (continuous):**
    - Ejection fraction
    - Pulmonary artery pressure

    - Valve severity levels
    - Septal and wall thickness

- **Target Label:** *shd_moderate_or_greater_flag* (0 = no SHD, 1 = SHD present).

This dataset enables both **classification** (disease presence) and **regression** (disease severity) tasks. The main focus of this research will be **binary classification**.

3. Problem Statement

While echocardiography is essential for SHD diagnosis, its limited accessibility creates barriers for early detection. ECGs are widely available but require specialized expertise to link with SHD outcomes.

There is a need for an automated, accurate, and interpretable approach that predicts SHD from ECG data, explains the underlying reasoning, and guides follow-up clinical action.

4. Objectives

- **Predictive Modeling:** Develop a binary classifier to predict SHD presence from ECG and demographic features.
- **Model Benchmarking:** Compare multiple machine learning models (Logistic Regression, Random Forest, XGBoost, Neural Networks) for predictive performance.
- **Explainability:** Identify key ECG features contributing to SHD prediction using explainable AI techniques (e.g., SHAP).
- **Recommendation System:** Build an LLM-powered system that generates natural language explanations and clinical recommendations based on model outputs.

5. Methodology

- **Data Preprocessing:** Handle missing values, encode categorical features, and standardize ECG variables.
- **Exploratory Analysis:** Analyze feature distributions and correlations between ECG features and SHD outcomes.
- **Model Development:** Train and evaluate multiple classification models.
- **Evaluation Metrics:** Use ROC-AUC, F1-score, precision, and recall.
- **Explainability Layer:** Apply feature importance analysis (e.g., SHAP, LIME).
- **LLM Recommendation System:**
    - Translate feature contributions into interpretable text.
    - Explain why SHD is predicted.
    - Provide recommendations for further clinical evaluation or lifestyle/risk management.

6. Expected Outcomes

- A robust predictive model for detecting SHD using ECG features.
- Identification of ECG markers most strongly associated with SHD.
- An interpretable recommendation system powered by LLMs that can explain predictions in natural language.
- A decision-support tool that prioritizes patients for echocardiography, improving early detection and resource allocation.

7. Novelty and Contributions

- **Scientific:** Validates the predictive capability of ECG data for SHD using machine learning.
- **Clinical:** Provides interpretable recommendations rather than black-box predictions.
- **Technological:** Integrates ML with LLMs for explainable decision support.
- **Societal:** Offers a low-cost, scalable screening solution for early SHD detection in resource-limited settings.

8. Conclusion

This research will not only build a predictive model for SHD from ECG data but also ensure interpretability through a recommendation system that explains why SHD is predicted and what steps can be taken next.

Such a system has the potential to bridge the gap between affordable ECG screening and costly echocardiographic diagnosis, contributing to timely detection and better patient care.