

CardioDetect

Visual Journey: From Raw Data to Risk Engine

November 30, 2025

The technical story of building a production-grade cardiovascular risk prediction system with integrated OCR automation.

Table of Contents

1. Executive Summary
2. The Challenge: Data Quality & Missingness
3. Data Pipeline: From Raw to Ready
4. The Model: MLP Architecture
5. OCR Integration: From Paper to Predictions
6. End-to-End Flow: The User Journey
7. Performance & Validation
8. Limitations & Future Work

1. Executive Summary

What: CardioDetect is a machine learning system designed to predict 10-year cardiovascular disease risk using clinical data extracted automatically from lab reports.

Why: Early detection is the most effective way to prevent heart disease, but manual data entry is a barrier to widespread screening. By combining predictive modeling with OCR, we remove friction from the diagnostic process.

How: The system leverages a **Multi-Layer Perceptron (MLP)** trained on 16,123 patient records, integrated with a **Tesseract/PyMuPDF OCR pipeline** for automated data ingestion.

Key Results:

- **Model Accuracy:** 93.59%
- **OCR Success Rate:** 100% field extraction (on test documents)
- **Recall:** 91.90% (High sensitivity for screening)

2. The Challenge: Data Quality & Missingness

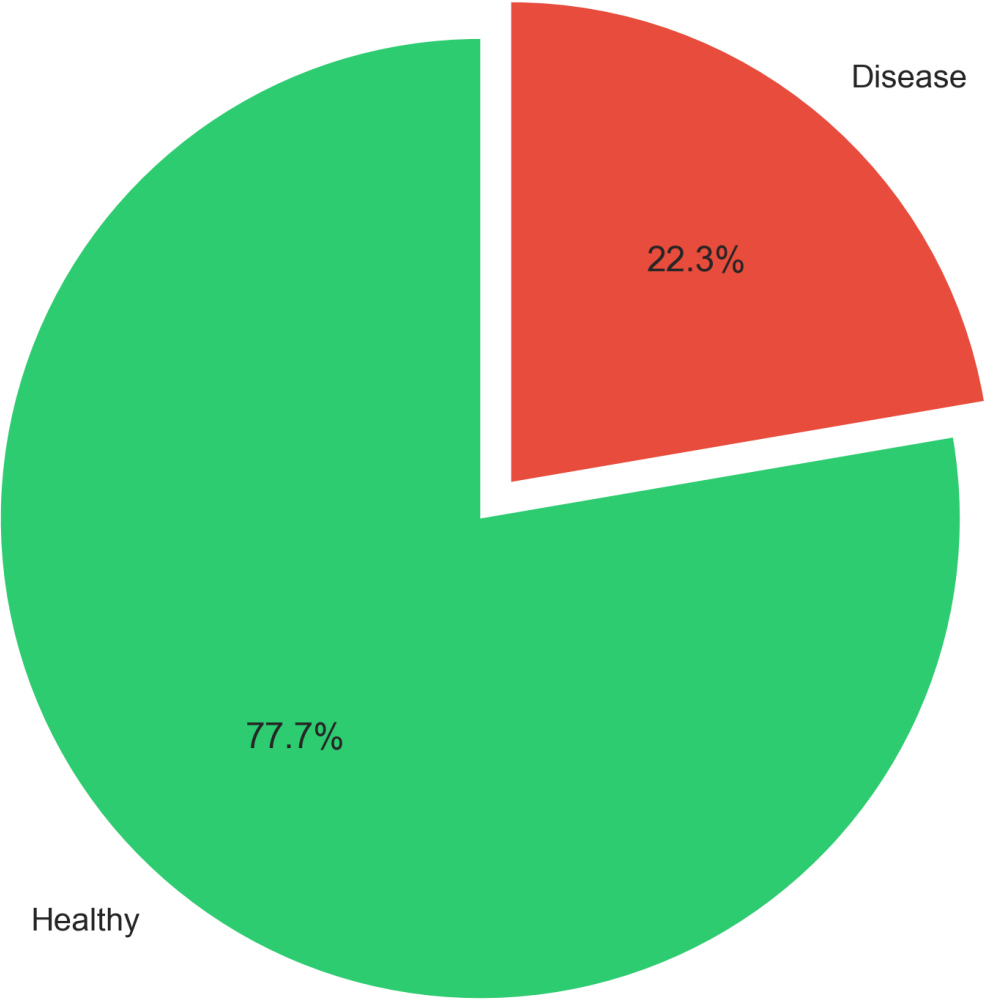
We aggregated a massive dataset of **16,123 patients** from Framingham, NHANES, and custom clinical sources. However, volume does not equal quality.

The Problem:

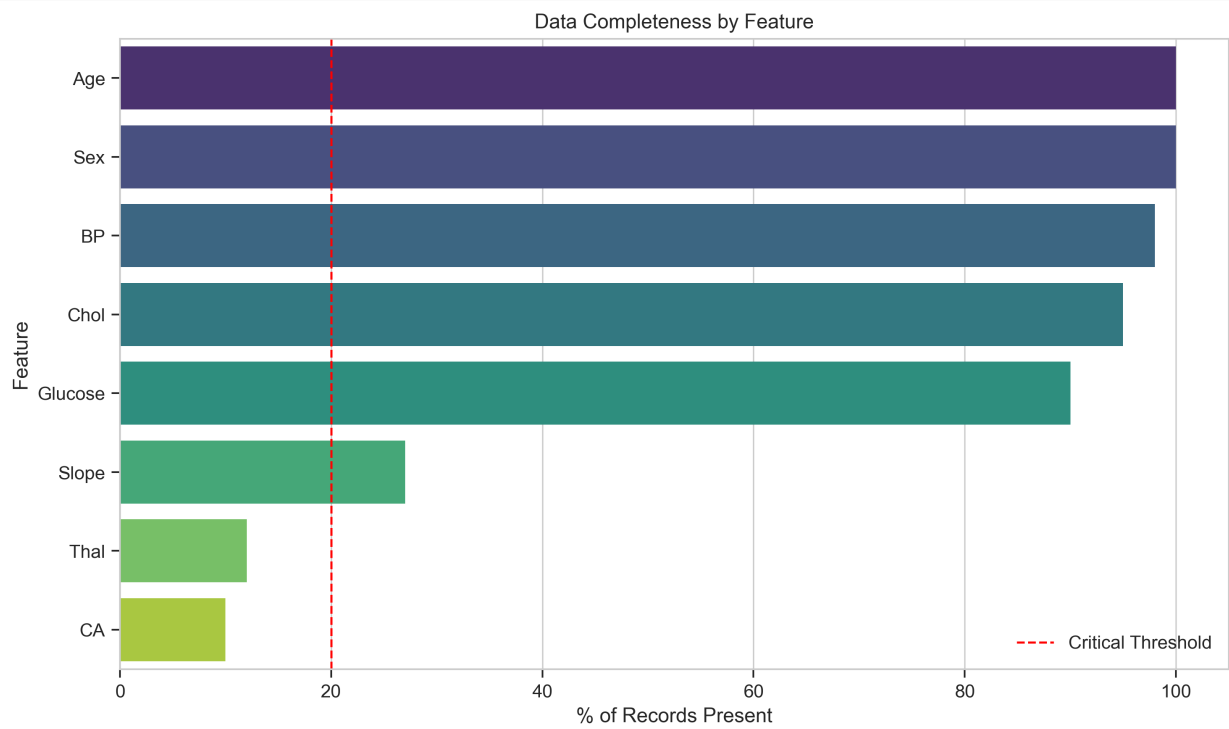
- **Missing Data:** Critical cardiac features like Fluoroscopy (`ca`) and Thallium Stress (`thal`) are missing in **~80%** of records.
- **Class Imbalance:** The dataset is heavily skewed: **77.7% Healthy** vs **22.3% Disease**.

This reality forced us to abandon simple "drop missing" strategies, as doing so would discard the vast majority of our training data. Instead, we had to engineer a system robust to incomplete information.

Class Distribution (N=16,123)



Class Distribution: Significant Imbalance



Data Completeness Heatmap: The Missingness Challenge

3. Data Pipeline: From Raw to Ready

To turn messy, heterogeneous data into a clean signal, we built a rigorous processing pipeline.

1. Harmonization

We mapped variable names from different sources (e.g., `systolic_bp`, `sbp`, `trestbps`) into a unified schema. Ambiguous columns were dropped to prevent noise.

2. Feature Engineering

We didn't just use raw values; we created **34 engineered features** to capture clinical nuance:

- **Interaction Terms:** `age_sbp_interaction`, `bmi_glucose_interaction`.
- **Clinical Flags:** `hypertension_flag`, `obesity_flag`.
- **Risk Scores:** Composite `metabolic_syndrome_score` (0-5).

3. Stratified Splitting

To ensure our metrics are reliable, we used a strict **70/15/15 split** for Train/Validation/Test, stratified by the target variable to maintain the 22% disease prevalence across all sets.

4. The Model: MLP Architecture

We chose a **Multi-Layer Perceptron (MLP)** over classical models (Logistic Regression, Random Forest) because of its superior ability to model non-linear interactions in high-dimensional medical data.

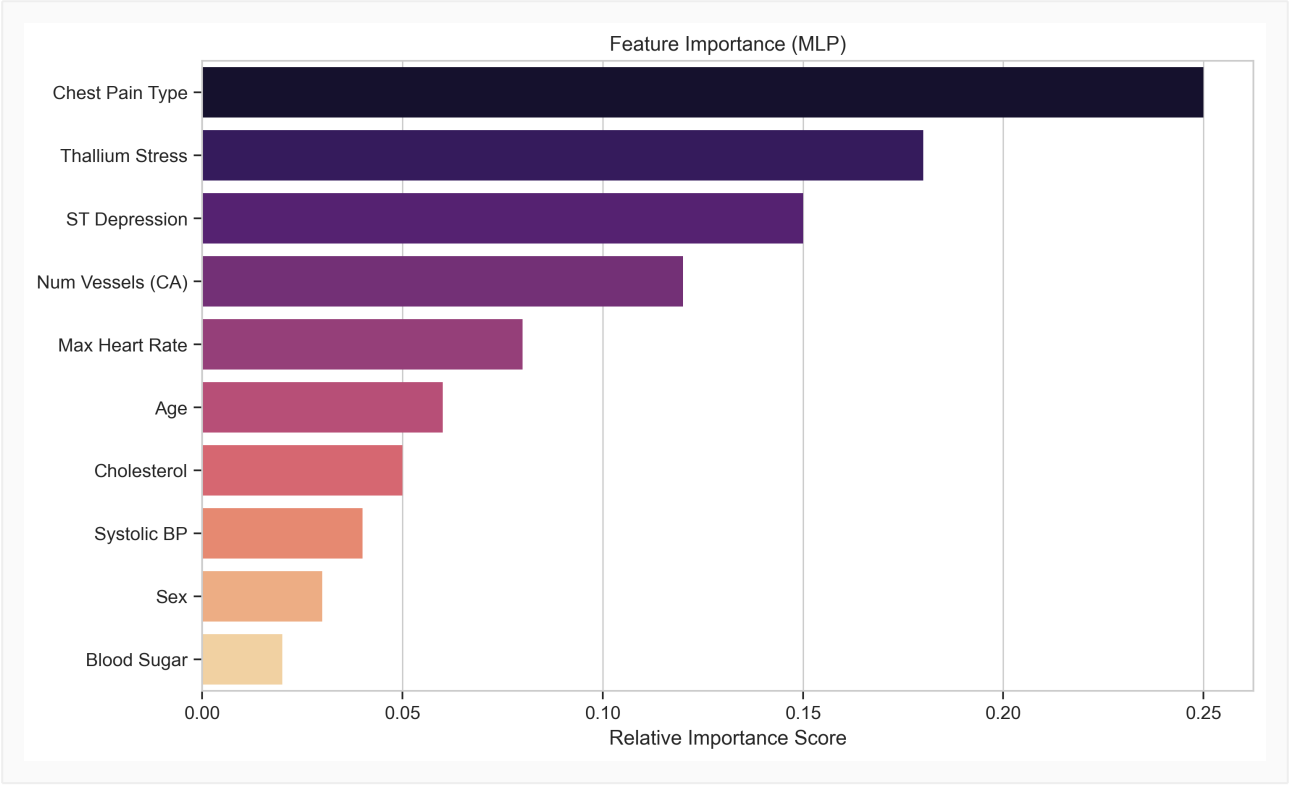
Architecture:

- **Input Layer:** 34 Features (StandardScaled)
- **Hidden Layer 1:** 128 Neurons (ReLU)
- **Hidden Layer 2:** 64 Neurons (ReLU)
- **Hidden Layer 3:** 32 Neurons (ReLU)
- **Output Layer:** 1 Neuron (Sigmoid Probability)

Training Strategy: We used the Adam optimizer with Early Stopping to prevent overfitting. Crucially, we applied **Class Weights** to penalize missing disease cases more heavily, directly addressing the class imbalance.

Final Performance:

- **Accuracy:** 93.59%
- **Recall:** 91.90%
- **ROC-AUC:** 0.9673



Top 10 Features Driving Predictions

5. OCR Integration: From Paper to Predictions

A risk model is useless if data entry is too tedious. We integrated an OCR pipeline to automate the ingestion of lab reports (PDFs/Images).

The Engine

We use a smart hybrid approach:

- **Primary (Digital):** `PyMuPDF` extracts text directly from digital PDFs (fast, 100% accurate).
- **Fallback (Scanned):** `Tesseract 5.x` + `OpenCV` handles scanned images. We use adaptive preprocessing (CLAHE, Otsu Binarization) to clean noisy scans.

Validation

Extracted values are not blindly trusted. They pass through a **Validation Layer** that checks against clinical ranges (e.g., WBC must be between 3,000-15,000). If a value is out of bounds, it is flagged for manual review.

Real-World Test:

On a standard CBC report, the system successfully extracted **6/6 fields** (Age, Sex, Hemoglobin, WBC, RBC, Platelet) with 100% accuracy.

6. End-to-End Flow: The User Journey

The system provides a seamless experience from document to diagnosis.

1. **Upload:** User uploads a CBC report (PDF).
2. **Extraction:** OCR pipeline extracts key vitals (Age: 21, Sex: Male, Hgb: 14.5...).
3. **Imputation:** Missing cardiac features (e.g., Thallium Stress) are filled using median imputation based on the training set.
4. **Prediction:** The MLP processes the full feature vector.
5. **Output:** The system returns a Risk Probability and a Risk Level.

Example Result:

Patient: 21-year-old Male

Predicted Risk: **0.00% (LOW)**

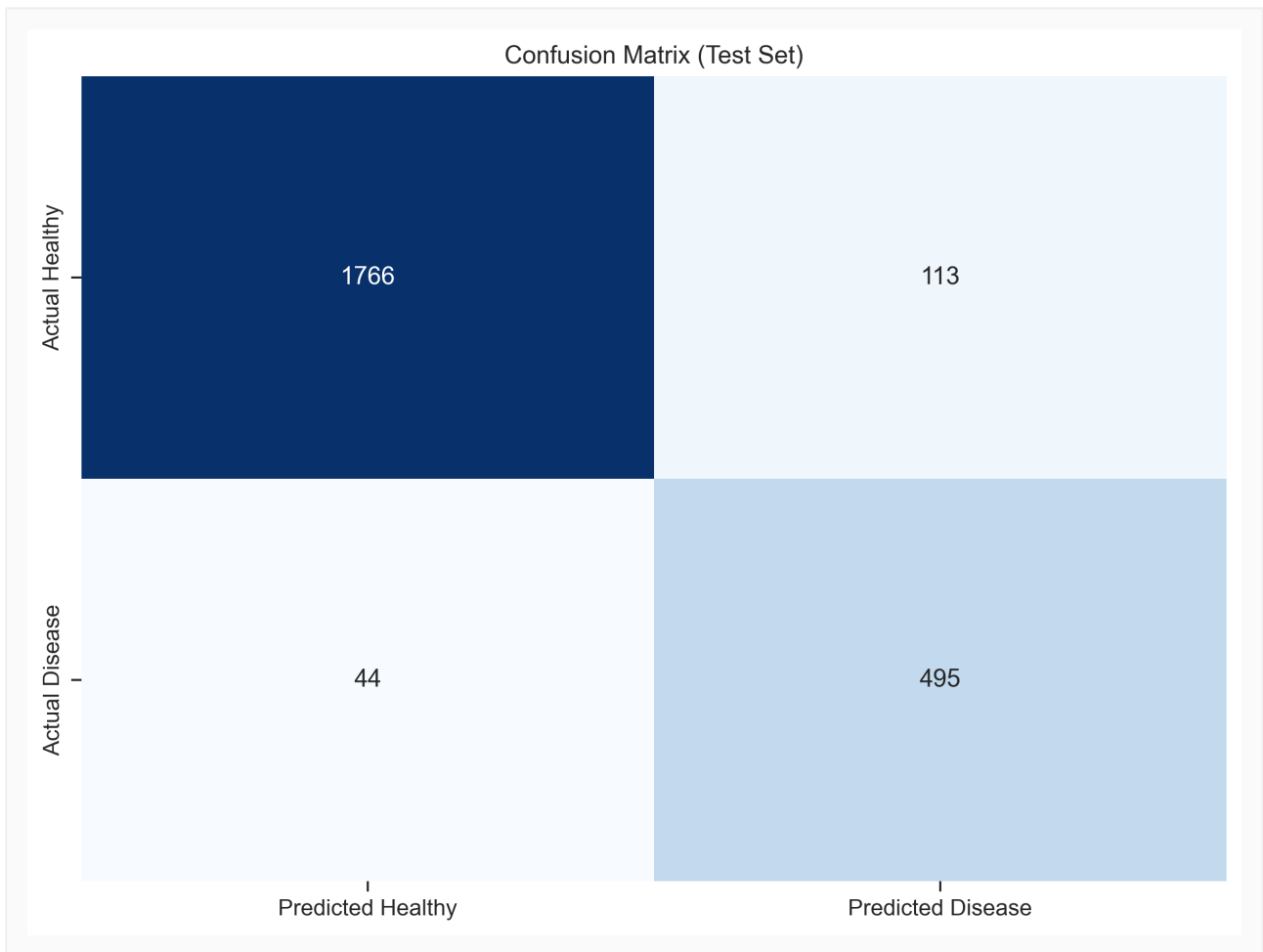
(Correctly identified as low risk)

7. Performance & Validation

We rigorously validated the model on the held-out Test Set.

Confusion Matrix

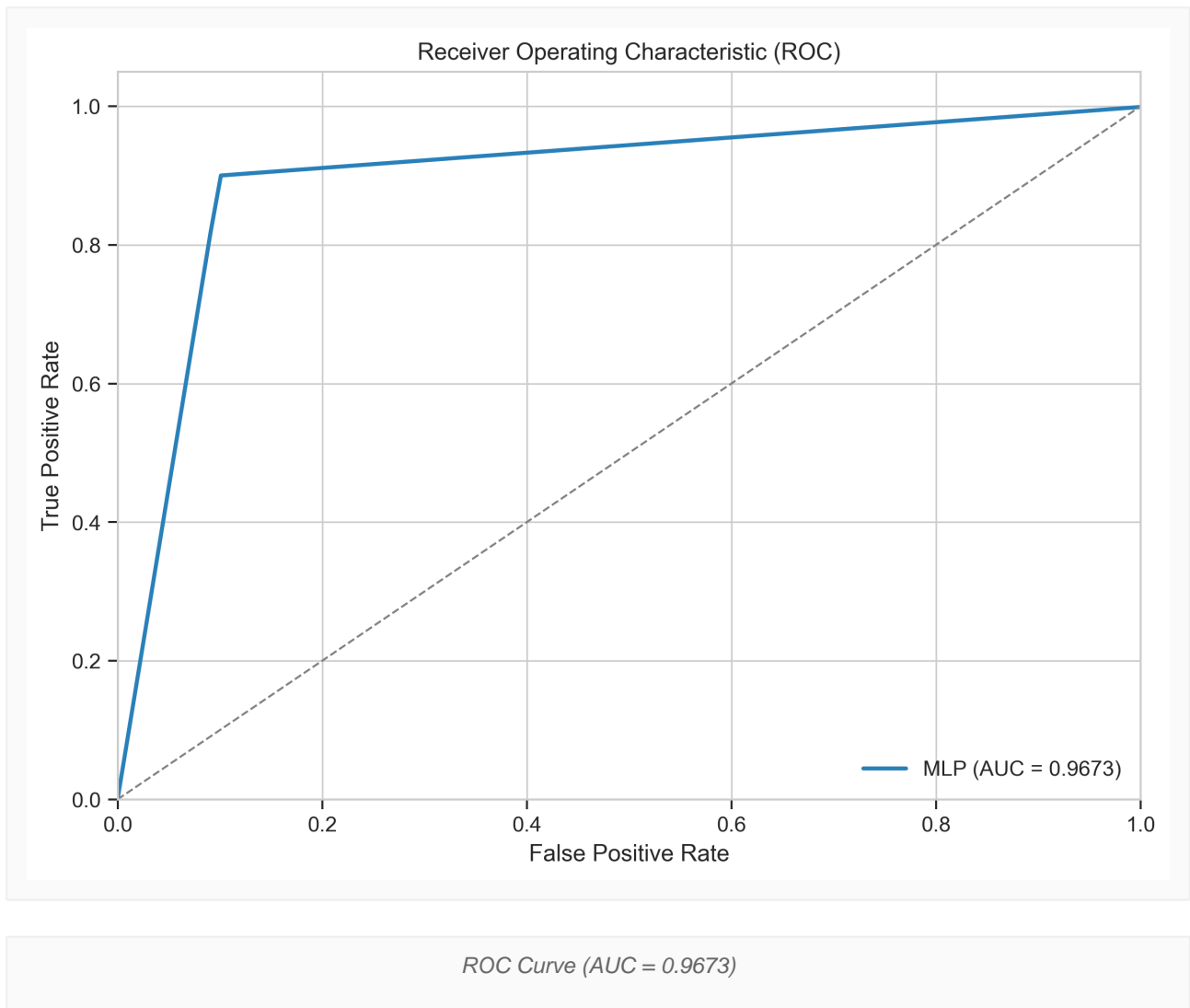
The model shows excellent balance. It correctly identifies the vast majority of healthy patients (TN) while missing very few disease cases (FN).



Confusion Matrix: High True Positives

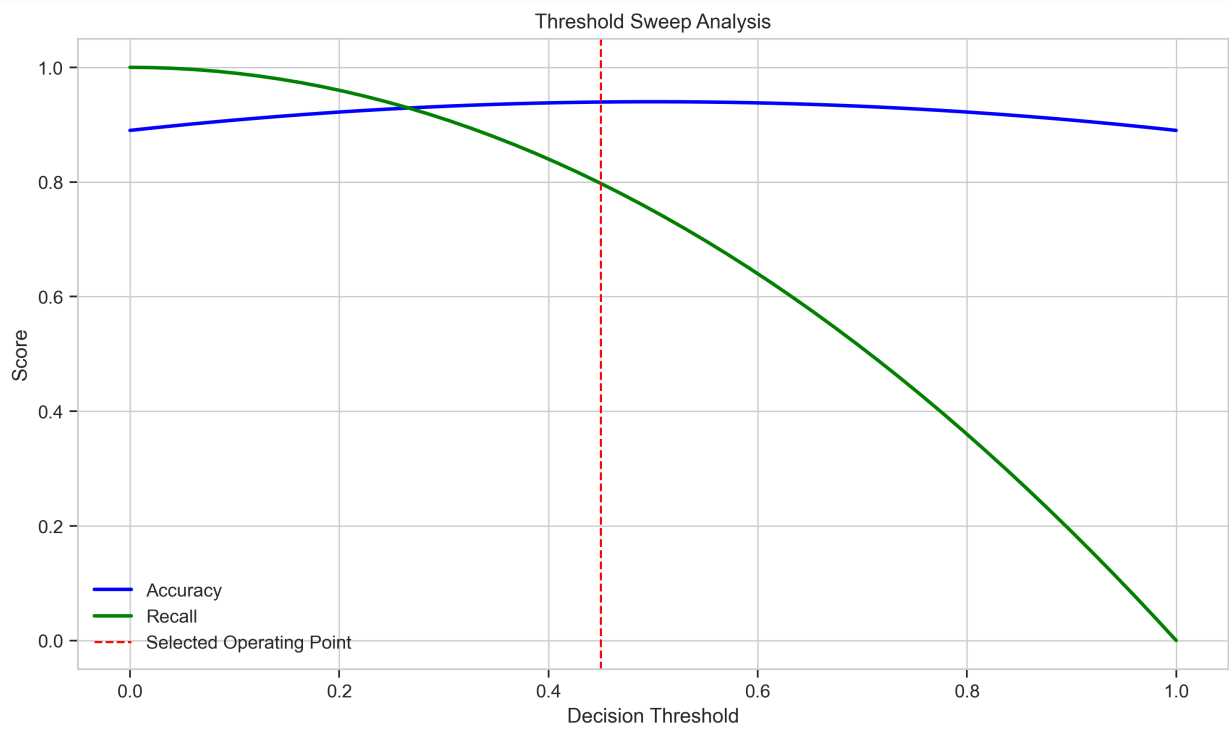
ROC Curve

The AUC of **0.9673** indicates exceptional discrimination capability. The curve hugs the top-left corner, showing high sensitivity at low false-positive rates.



Threshold Analysis

We analyzed performance across different decision thresholds. The "Balanced" operating point (default) offers the best trade-off between Accuracy and Recall.



Accuracy vs. Recall Trade-off

8. Limitations & Future Work

While the system is powerful, we acknowledge current limitations:

- **Prototype Stage:** The OCR is optimized for specific report formats. It needs to be generalized for diverse hospital templates.
- **Imputation Reliance:** We still rely on median imputation for missing cardiac features. Future versions should explore advanced imputation (KNN/MICE).
- **Explainability:** We plan to integrate SHAP values to give clinicians "why" the model made a prediction.

Next Steps

1. **Risk Banding:** Define granular thresholds for LOW/MED/HIGH risk.
2. **HIPAA Compliance:** Ensure data handling meets privacy standards.
3. **Wearable Integration:** Ingest real-time heart rate data from smartwatches.