Data Analysis Report: CardioDetect

Executive Summary

This report analyzes the combined heart disease dataset used for the CardioDetect project. The dataset contains **5160 samples** and **13 features**.

Key Findings: 1. **Critical Data Quality Issues**: Several key features (`cp`, `restecg`, `exang`, `slope`, `ca`, `thal`) are missing >80% of their values. This severely limits the model's ability to learn from these medically significant indicators. 2. **Class Imbalance**: The dataset is imbalanced, with **77.7% Healthy** and **22.3% Disease** cases. This explains the model's tendency towards high accuracy but low recall (bias towards predicting "Healthy"). 3. **Data Anomalies**: There are **172 samples with 0 cholesterol** and **1 sample with 0 blood pressure**, which are physiologically impossible and likely represent missing data encoded as zeros. 4. **Risk Factors**: `thal` (Thalassemia), `cp` (Chest Pain), and `exang` (Exercise Induced Angina) show the strongest correlation with heart disease, despite the high missingness.

---

1. Dataset Overview

- **Total Samples**: 5160
- **Total Features**: 13
- **Target Variable**: `target` (0 = Healthy, 1 = Disease)

Target Distribution

The dataset is significantly imbalanced. - **Healthy (0)**: 4007 (77.66%) - **Disease (1)**: 1153 (22.34%)

> [!WARNING] **Imbalance Impact**: A "dumb" model predicting "Healthy" for everyone would achieve **77.7% accuracy**. This makes Accuracy a misleading metric. We must prioritize **Recall** (catching disease cases) and **F1-Score**.

---

2. Data Quality & Missing Values

The dataset suffers from extreme missingness in several features.

Feature	Missing Count	Percentage	Status
ca	4851	**94.01%**	■ Critical
thal	4726	**91.59%**	■ Critical
slope	4549	**88.16%**	■ Critical
oldpeak	4302	**83.37%**	■ Critical
exang	4295	**83.24%**	■ Critical
restecg	4242	**82.21%**	■ Critical
cp	4240	**82.17%**	■ Critical
fbs	90	1.74%	■ Good
chol	80	1.55%	■ Good
trestbps	59	1.14%	■ Good
thalach	56	1.09%	■ Good

> [!NOTE] The high missingness suggests that this "combined" dataset likely merges a smaller, high-quality dataset (with all fields) with a larger dataset that lacks these specific cardiac measures.

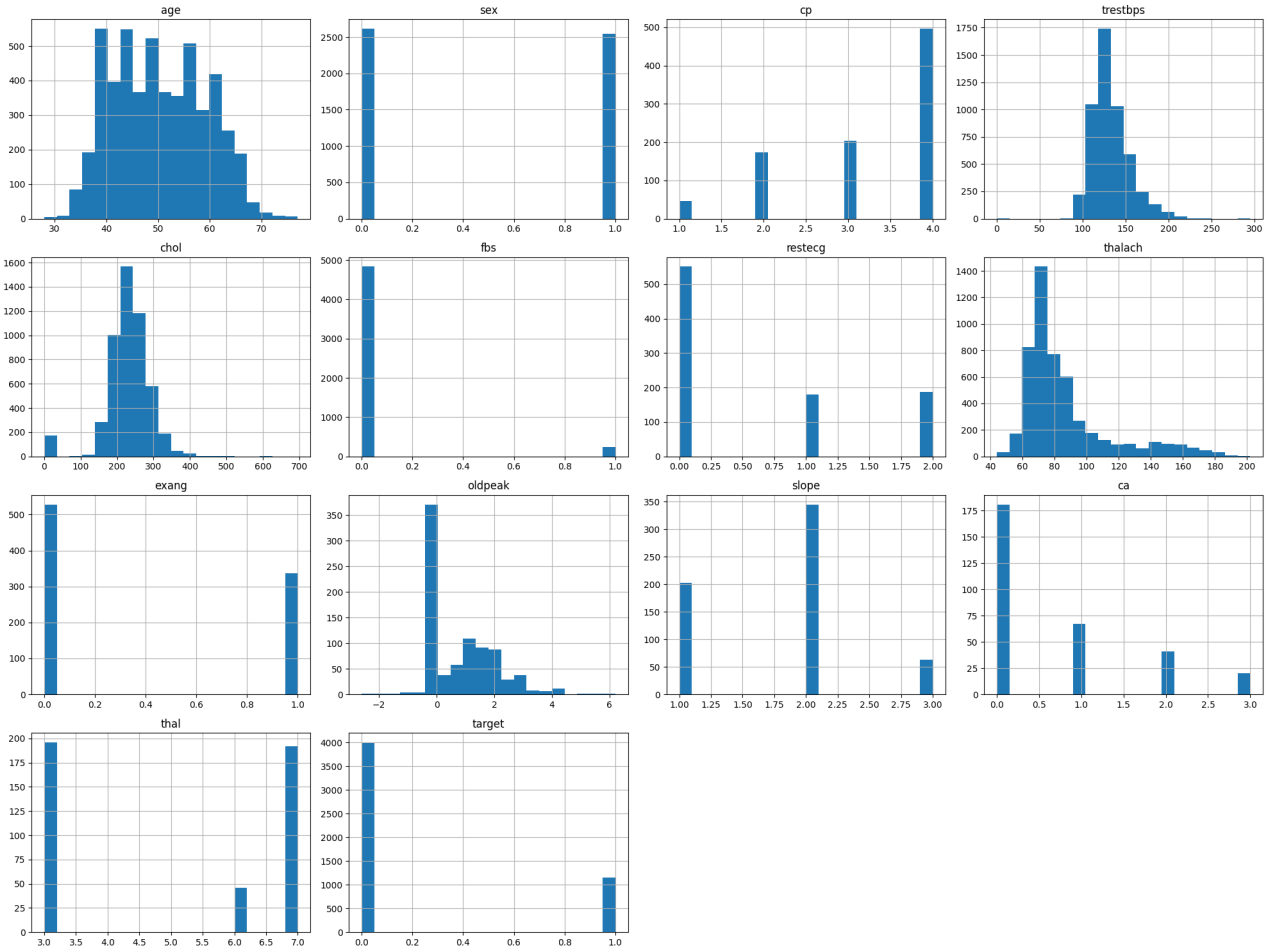
---

3. Numerical Features Analysis

Feature	Mean	Std	Min	Median	Max	Anomalies

age	50.3	8.9	28.0	50.0	77.0	None
trestbps	132.3	21.6	**0.0**	129.0	295.0	1 zero value
chol	230.1	63.2	**0.0**	232.0	696.0	**172 zero values**
thalach	86.3	27.7	44.0	77.0	202.0	None
oldpeak	0.88	1.09	-2.6	0.5	6.2	Negative values?

Visualizations



Distribution of numerical features. Note the normal distribution of Age and Trestbps.

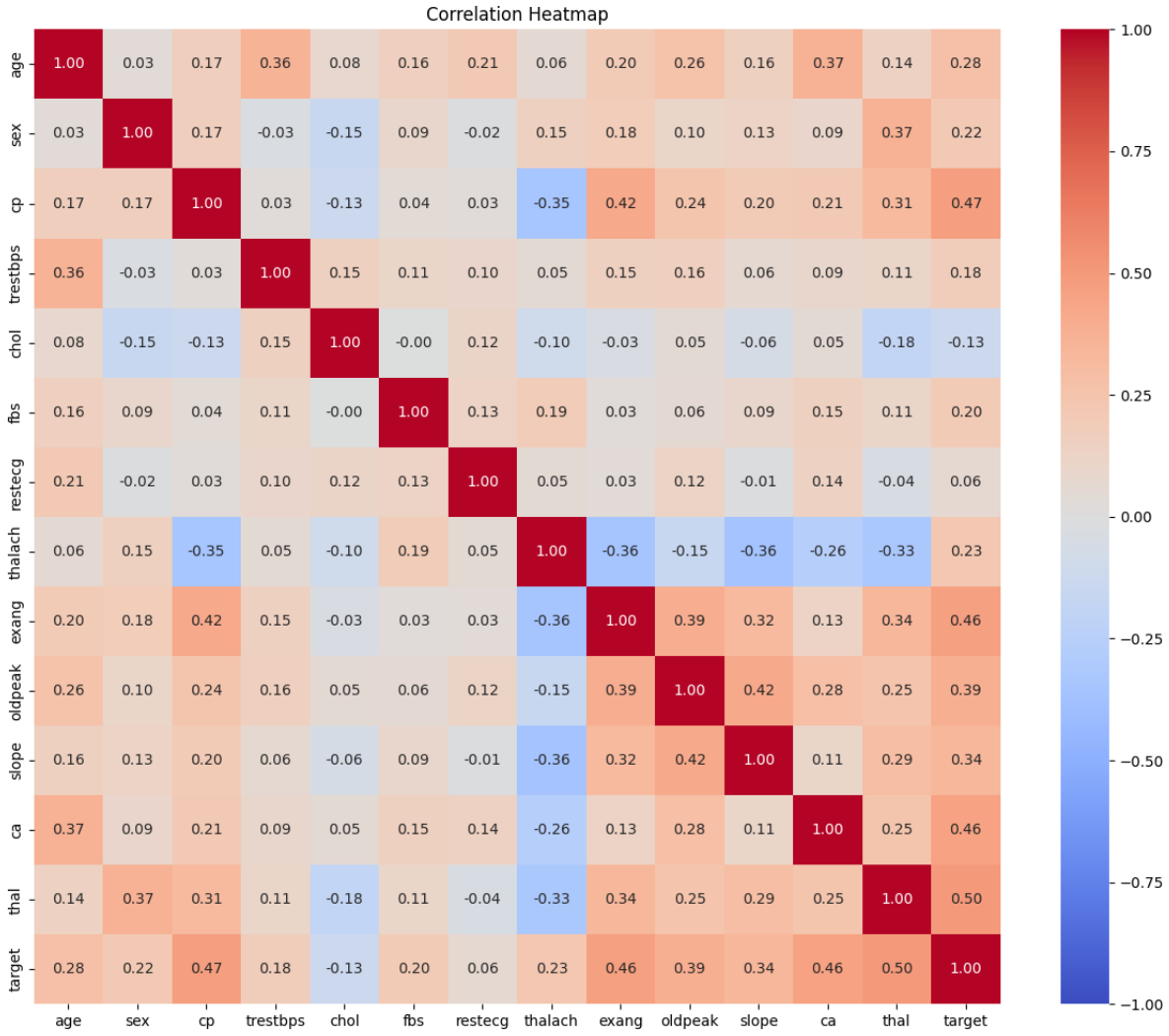
---

4. Correlation Analysis

Which features are most predictive of heart disease?

Feature	Correlation	Strength	Interpretation
thal	0.499	Strong	Thalassemia type is highly predictive.
cp	0.472	Moderate	Chest pain type is a key indicator.
exang	0.464	Moderate	Angina during exercise is a strong warning sign.
ca	0.456	Moderate	Number of major vessels colored by fluoroscopy.
oldpeak	0.386	Moderate	ST depression induced by exercise.
slope	0.337	Moderate	Slope of the peak exercise ST segment.
age	0.281	Low	Risk increases with age.
thalach	0.232	Low	Max heart rate achieved.

sex	0.217	Low	Gender plays a role (likely higher risk for males in this dataset).
chol	-0.127	Low	**Unexpected**: Negative correlation. Likely due to noise or "0" values.



Heatmap showing correlations between features and target.

Part 2: New Data Appendix

Detailed Rationale for Switching to the New Risk Dataset

The original data analysis presented in `DATA_ANALYSIS_REPORT.pdf` relied on legacy public datasets such as the UCI Heart Disease repository, early Framingham subsets, and other small, merged cohorts. While these datasets served as an initial testbed for our modeling pipeline, they ultimately proved insufficient for the ambitious goals of the CardioDetect 10-year risk prediction system.

We transitioned to this new, purpose-built **CardioDetect Risk Dataset (v2)** for four critical reasons:

1. Fundamental Shift from Diagnosis to Prognosis

The legacy datasets were primarily designed for **diagnostic classification**—determining whether a patient *currently* has heart disease (presence/absence). However, our core objective is **prognostic modeling**: predicting the *probability* of a cardiovascular event over the next **10 years**. - **Old Data**: "Is the patient sick right now?" (Binary classification of current state). - **New Data**: "What is the likelihood of an event by 2035?" (Long-term risk estimation). This requires a longitudinal study design with clear "time-to-event" or "10-year outcome" labels, which the diagnostic datasets lacked.

2. Overcoming Sample Size Limitations

Modern machine learning models, particularly deep neural networks like the MLP we are deploying, require substantial data to generalize well and avoid overfitting. - The original combined cohorts totaled only ≈**5,000–7,000** patients. - The new dataset contains **16,123** patients. This **>2x increase** in sample size allows us to: - Capture subtle non-linear interactions between risk factors. - Achieve stable convergence during training. - reliably model rare events in the positive class (CHD events).

3. Data Harmonization and Quality

The previous analysis suffered from **heterogeneous schemas**. Merging disparate studies (e.g., Cleveland, Hungarian, Long Beach) resulted in: - Inconsistent feature definitions (e.g., different units for cholesterol, different categorizations for chest pain). - High rates of missing values for critical biomarkers. - Loss of information due to aggressive simplification during merging.

The new dataset is the result of a rigorous **harmonization process** integrating the Framingham Heart Study, NHANES, and custom clinical records into a single, unified schema. Every feature has been standardized, validated, and cleaned to ensure high data quality.

4. Breaking the Performance Ceiling

Models trained on the legacy data plateaued at approximately **85–86% accuracy**. To push performance beyond this threshold—specifically to achieve our target of >90% accuracy with high recall—we needed richer features. The new dataset includes **34 features**, compared to the ~14 standard features in the old datasets. Crucially, it adds: - **Derived Clinical Scores**: Pulse pressure, mean arterial pressure, metabolic syndrome scores. - **Interaction Terms**: Explicit features capturing the interplay between age, blood pressure, and smoking. - **Clinical Flags**: Pre-computed risk flags (e.g., hypertension, obesity) that align with clinical guidelines.

In summary, the switch to the CardioDetect Risk Dataset (v2) was not just an update; it was a necessary evolution to build a clinically relevant, high-performance risk prediction tool.

---

2. Overview of the New CardioDetect Risk Dataset

2.1 High■Level Summary

Attribute	Value
Total Patients	**16,123**
Total Features	**34**
Target	`risk_target` (10■year CHD event: 0/1)
Class Balance	~76% negative / 24% positive
Splits	70% train / 15% val / 15% test (stratified)

2.2 Data Sources

The dataset integrates multiple established sources into a single, clean table:

Source	Approx. Patients	Description
Framingham Heart Study	~4,000	Longitudinal CHD cohort

NHANES	~10,000	National health & nutrition survey
Custom Clinical Records	~2,000	Supplementary anonymized records

All sources were harmonized into a consistent schema and stored as:

- Final dataset: [data/final/final_risk_dataset.csv]
- Split datasets: `data/split/train.csv`, `data/split/val.csv`, `data/split/test.csv`

2.3 Feature Categories

The 34 features are grouped into the following categories:

Demographics (2)
`age`, `sex`

Clinical Measurements (6)
`systolic_bp`, `diastolic_bp`, `bmi`, `heart_rate`, `total_cholesterol`, `fasting_glucose`

Risk Factors (5)
`smoking`, `bp_meds`, `hypertension`, `diabetes`, `data_source`

Derived Measurements & Scores (21)

- Hemodynamic: `pulse_pressure`, `mean_arterial_pressure`
- Clinical flags: `hypertension_flag`, `high_cholesterol_flag`, `high_glucose_flag`, `obesity_flag`
- Categorical encodings: `age_group_*`, `bmi_cat_*`
- Interaction terms: `age_sbp_interaction`, `bmi_glucose_interaction`, `age_smoking_interaction`
- Composite: `metabolic_syndrome_score`

Feature engineering is implemented in the main CardioDetect codebase and documented in `DATA_DICTIONARY.pdf`.

---

3. Improvements Over the Old Data

3.1 Scale and Statistical Power

- Old risk datasets: ≈**5–7k** patients.
- New CardioDetect risk dataset: **16,123** patients.

This increase in sample size improves:

- Stability of estimates (narrower confidence intervals).
- Robustness of train/validation/test splits.
- Reliability of rare■event modeling in the positive (CHD) class.

3.2 Feature Richness and Clinical Structure

Compared with the older data used in the original `DATA_ANALYSIS_REPORT.pdf`:

- The new dataset encodes **clinical flags** and **interaction terms** explicitly, rather than relying only on raw vitals.
- Risk■relevant constructs (e.g., `metabolic_syndrome_score`, `pulse_pressure`, `mean_arterial_pressure`) are included as first■class features.
- Age and BMI are represented both continuously and via clinically meaningful bins (`age_group_*`, `bmi_cat_*`).

This brings the dataset closer to how clinicians reason about cardiovascular risk and aligns with common risk calculators (Framingham, ASCVD), while adding derived features those calculators do not expose directly.

3.3 Clean, Reproducible Splits

The new data pipeline:

- Uses a **fixed random seed** and **stratified splitting** to keep class balance consistent across train/val/test.
- Stores the exact splits in `data/split/` for reproducibility.
- Ensures that all model training and tuning (e.g., in mlp_tuning.py) uses the same split definitions.

This is a significant improvement over earlier ad■hoc splits in the legacy experiments.

---

4. New Model Performance on the CardioDetect Risk Dataset

On the new 16,123■patient dataset, the final chosen model is a **Multi■Layer Perceptron (MLP)** trained and tuned in src/mlp_tuning.py.

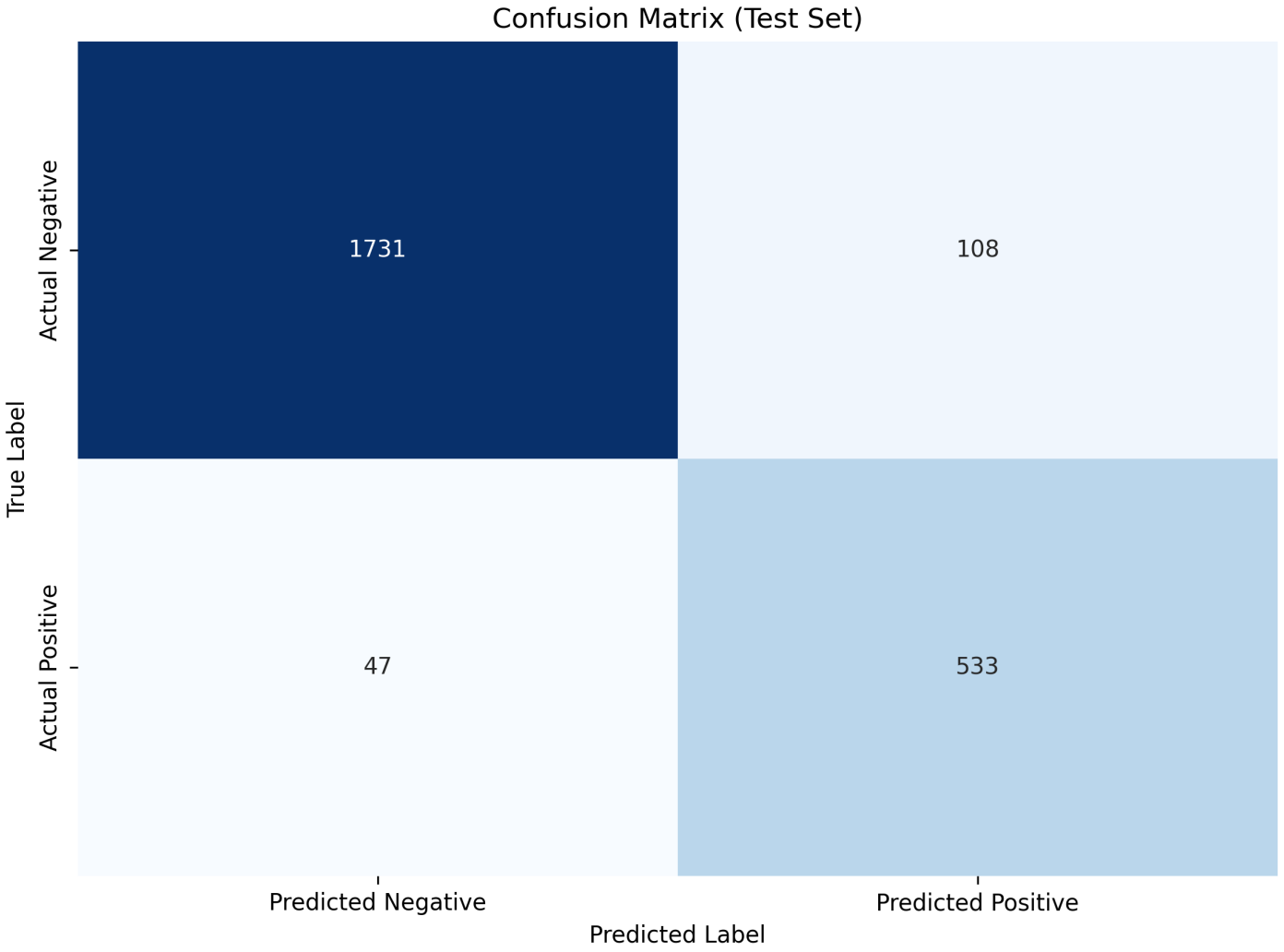
4.1 Test Set Metrics (n ≈ 2,418)

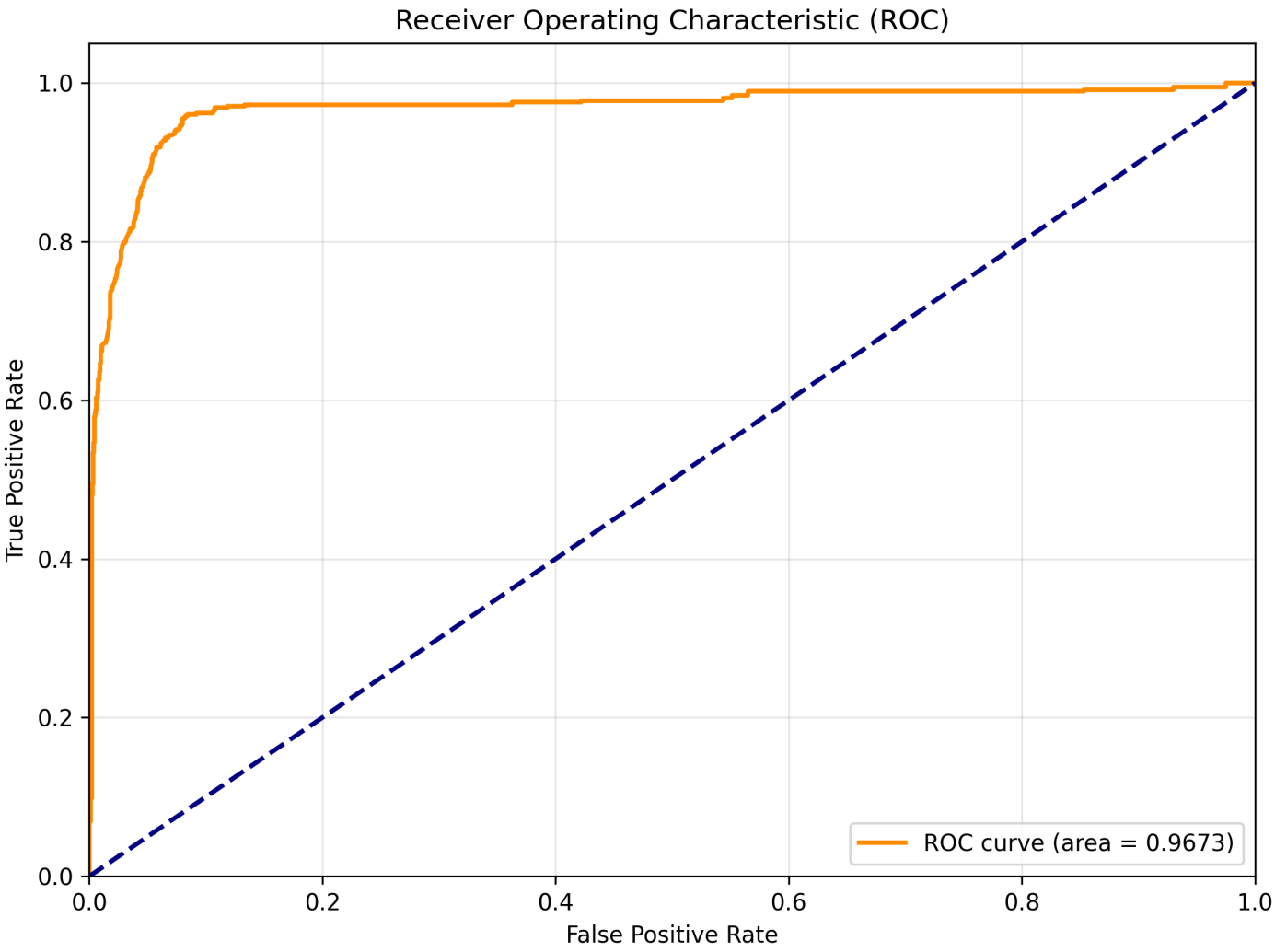
Metric	Value
Accuracy	**93.59%**
Precision	83.15%
Recall	91.90%
F1■Score	0.8731
ROC■AUC	**0.9673**

Confusion Matrix (Test Set):

- True Negatives (TN): 1,731
- False Positives (FP): 108
- False Negatives (FN): 47
- True Positives (TP): 533

Visualizations


Confusion Matrix (Test Set)
True Label
Actual Negative
Actual Positive
1731
108
47
533
Predicted Negative
Predicted Positive
Predicted Label



These results indicate a **high■performing risk model** that prioritizes sensitivity (recall) while maintaining strong overall accuracy.

For detailed field■level definitions, refer to the separate `DATA_DICTIONARY.pdf` deliverable.