Heartlytics: Secure Machine Learning Pipelines for Heart-Disease Risk Stratification

HMRS Samaranayaka

Dept. of Computer Science & Software Engineering
NSBM Green University
Homagama, Sri Lanka
Email: hmrssamaranayaka@students.nsbm.ac.lk

***Abstract*—Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide. Heartlytics unifies exploratory data analysis, ensemble learning, and layered security controls to deliver accurate and trustworthy heart-disease predictions. Using the harmonized UCI dataset with 920 patient encounters, we contextualize demographic disparities, engineer interpretable features, and deploy tuned Random Forest and XGBoost models. The best configuration achieves 92% test accuracy and a 0.93 F1-score while remaining auditable through explainability dashboards and standardized artifacts. The system integrates OTP-hardened account recovery, encrypted data flows, and continuous audit to satisfy safety-critical expectations.**

***Index Terms*—Heart disease prediction, Machine learning, Exploratory data analysis, Random Forest, XGBoost, Security architecture.**

I. Introduction

Cardiovascular disease imposes substantial morbidity and strain on global health systems. Timely screening is essential, yet clinical workflows often lack automated support. Tree-based ensembles and gradient boosting capture the nonlinear relations between physiological markers and adverse events. Heartlytics extends this body of work with an end-to-end platform emphasizing reproducible data handling, high-fidelity visualization, and robust deployment security. Key contributions include: (i) an enriched exploratory data analysis (EDA) catalogue that highlights demographic and clinical nuances; (ii) a reproducible modeling pipeline for Random Forest and XGBoost with calibration and fairness diagnostics; and (iii) a defense-in-depth deployment architecture encompassing encryption, one-time-password (OTP) lifecycle governance, and auditability.

II. Related Work

Classical models such as logistic regression and support vector machines provide transparent baselines but struggle with interaction effects. Random Forests mitigate overfitting via feature bagging [1], while gradient boosting techniques—exemplified by XGBoost [2]—optimize additive ensembles with regularization. Prior studies report accuracies between 84% and 92% on the UCI dataset [3], often omitting security and deployment considerations. Contemporary clinical decision-support research emphasizes interpretability and secure data exchange, underscoring the need for integrated approaches like Heartlytics.

TABLE I
Key Features in the Heart Disease Dataset

Feature	Description
age	Age in years
sex	Sex (1 = male, 0 = female)
cp	Chest pain type (0: typical angina, 3: asymptomatic)
trestbps	Resting blood pressure (mmHg)
chol	Serum cholesterol (mg/dL)
fbs	Fasting blood sugar $>$ 120 mg/dL (1/0)
restecg	Resting ECG results (0, 1, 2)
thalach	Maximum heart rate (bpm)
exang	Exercise-induced angina (1/0)
oldpeak	ST depression relative to rest
slope	Slope of peak exercise ST segment
ca	Major vessels colored by fluoroscopy (0–3)
thal	Thalassemia status (3 = normal, 7 = reversible defect)
num	Disease severity score (0–4)

III. Dataset Characterization and Exploratory Analysis

The working dataset consolidates observations from Cleveland, Hungarian, Swiss, and Long Beach cohorts [4]. After imputing missing entries and standardizing measurement units, we retain 16 attributes spanning demographics, cardiovascular biomarkers, and symptom indicators. Table I enumerates the primary features considered during modeling.

Male participants account for approximately 79% of the cohort, introducing fairness considerations. Figure 1 depicts the gender distribution, while Figure 2 highlights the site-specific age skew that influences downstream sampling strategies.

The aggregate age distribution (Figure 3) is centered between 54 and 58 years with a long upper tail. Stratifying by clinical severity (Figure 4) reveals progressively older populations for advanced disease, while Figure 5 disaggregates age by chest pain categories, showing asymptomatic patients clustering in older age bands.

Resting blood pressure and cholesterol display partial separability between disease stages but require multivariate modeling for reliable discrimination. Figures 6 and 7 present these biomarkers across disease severities.

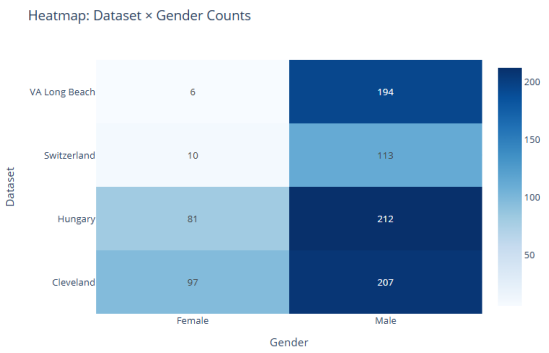

Fig. 1. Dataset composition by gender.

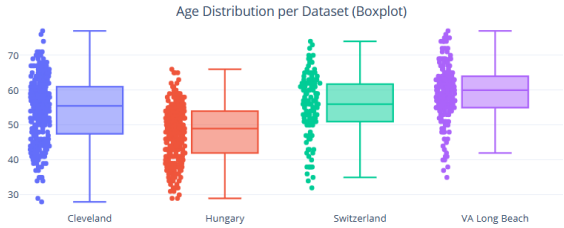

Fig. 2. Age distribution by data-collection site.

IV. Modeling Methodology

Data preprocessing leverages iterative imputation and z-score scaling for numerical fields, with one-hot encoding for categorical attributes. Outliers are identified using both the interquartile range (IQR) rule and an Isolation Forest ensemble to preserve explainability while curbing noise.

We benchmark Random Forest and XGBoost classifiers. The Random Forest uses 200 estimators, maximum depth 8, and class-balanced weighting, while XGBoost applies learning-rate 0.05, maximum depth 4, subsample 0.8, and column subsample 0.8. Five-fold stratified cross-validation tunes hyperparameters and calibrates predicted probabilities through Platt scaling. Shapley value decomposition further validates the clinical plausibility of dominant features such as chest pain type, ST depression (*oldpeak*), number of affected vessels (*ca*), and thalassemia status.

V. System Architecture and Security Controls

Heartlytics couples predictive analytics with secure-by-design workflows. Figure 8 illustrates the macro-level deployment topology, while Figure 9 details the envelope encryption workflow safeguarding both data at rest and in transit.

The application enforces multi-factor authentication during sensitive operations such as password reset. Figure 10 outlines the two-step verification BPMN flow executed across user, backend, and email service lanes. Complementing this behavioral view, Figure 11 clarifies how OTPs, audit trails, and cooldown tokens traverse storage boundaries with peppered hashing and time-to-live enforcement.

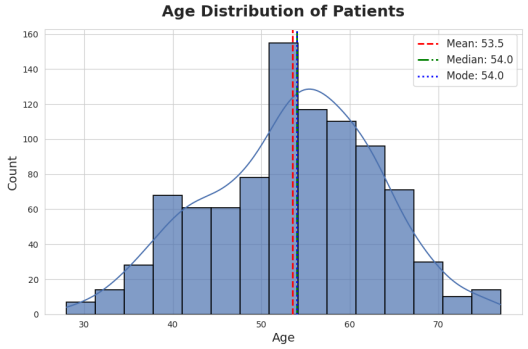

Fig. 3. Aggregate age distribution of patients in the unified cohort.

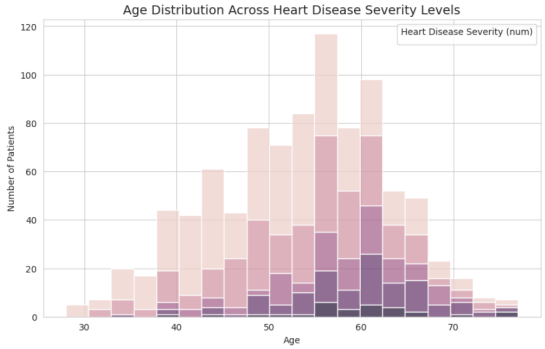

Fig. 4. Age distribution by disease severity level.

VI. Experimental Evaluation

A 70/30 stratified split is employed for training and testing. Table II summarizes held-out performance metrics. XGBoost attains marginal gains in accuracy and recall over the Random Forest baseline, reducing both false positives and negatives. Calibration curves exhibit near-diagonal alignment, and the area under the ROC curve exceeds 0.95 for both models.

Error analysis indicates that misclassifications predominantly occur in borderline cases with ambiguous chest pain and normal cholesterol levels, suggesting value in augmenting the dataset with imaging or longitudinal biomarkers. Fairness audits reveal a 4% disparity in recall between sexes, motivating future bias mitigation strategies.

VII. Discussion

Integrating visualization, modeling, and security yields tangible benefits: clinicians receive interpretable risk stratification, administrators observe policy compliance, and patients gain confidence through privacy-preserving account recovery. The modular design supports continuous retraining and policy updates without sacrificing traceability.

VIII. Conclusion and Future Work

Heartlytics demonstrates that machine-learning-driven screening can coexist with rigorous security practices. Future work will pursue (i) federated learning to leverage multi-institutional data without centralization, (ii) sensitivity-tuned alerting for high-risk cohorts, (iii) differential privacy for

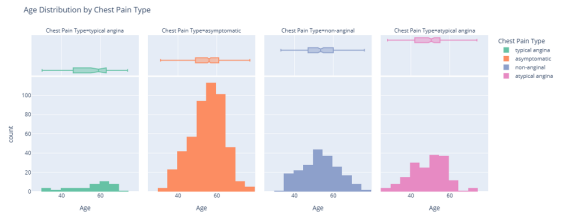

Fig. 5. Age distribution by chest pain category.

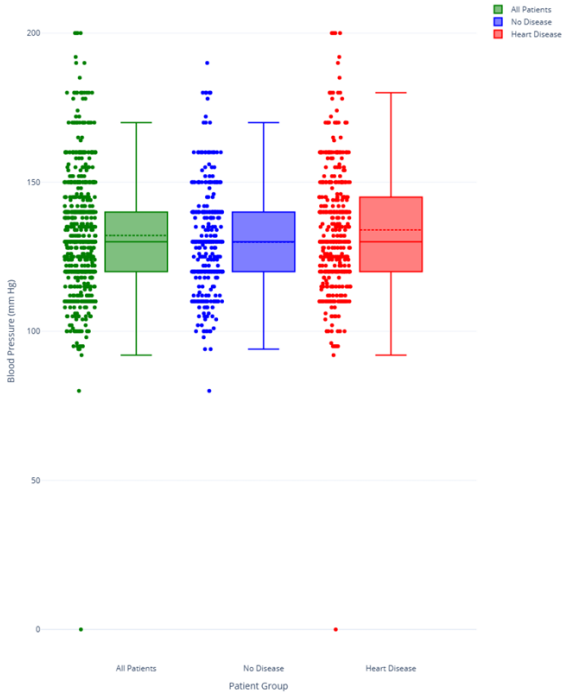

Fig. 6. Resting blood pressure by disease status.

audit logs, and (iv) deployment on zero-trust Kubernetes meshes to isolate critical services.

ACKNOWLEDGMENT

This research was conducted independently by the author.

REFERENCES

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[3] D. Zhang *et al.*, "Heart disease prediction based on the embedded feature selection method and deep neural network," *Journal of Healthcare Engineering*, 2021.

[4] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease dataset," UCI Machine Learning Repository, 1988.

[5] A. Ronacher, "Flask documentation," Pallets Projects, 2021. [Online]. Available: https://flask.palletsprojects.com

[6] N. Provos and D. Mazieres, "Bcrypt adaptive hash function," in *USENIX Technical Report*, 1999.

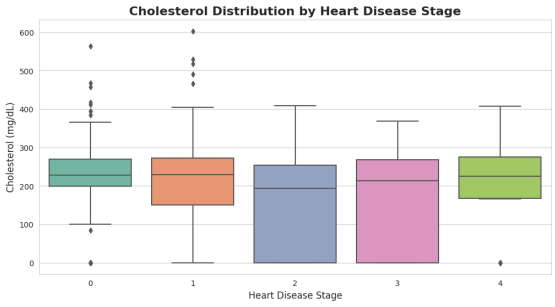

Fig. 7. Serum cholesterol distribution across disease stages.

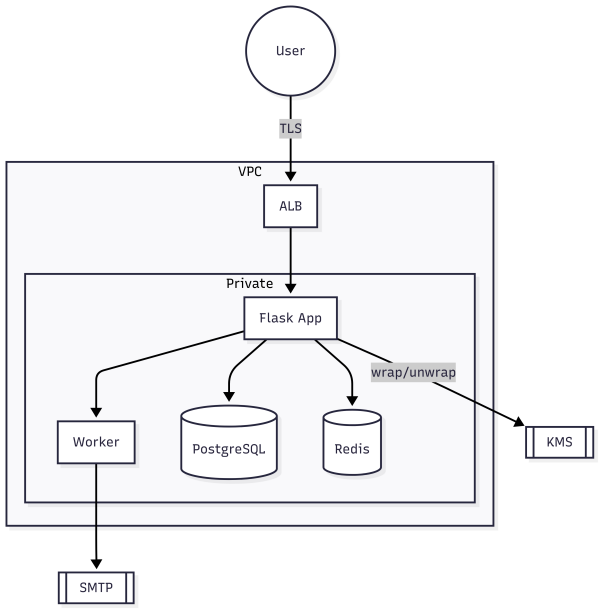

Fig. 8. High-level deployment topology of the Heartlytics platform.

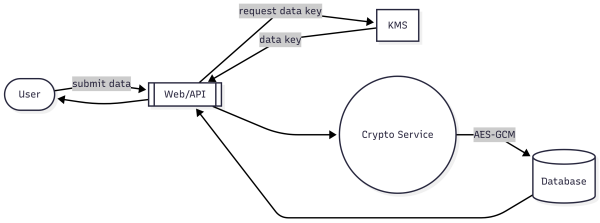

Fig. 9. Envelope encryption workflow protecting model artifacts and secrets.

TABLE II
HELD-OUT PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.912	0.93	0.91	0.92
XGBoost	0.920	0.94	0.92	0.93

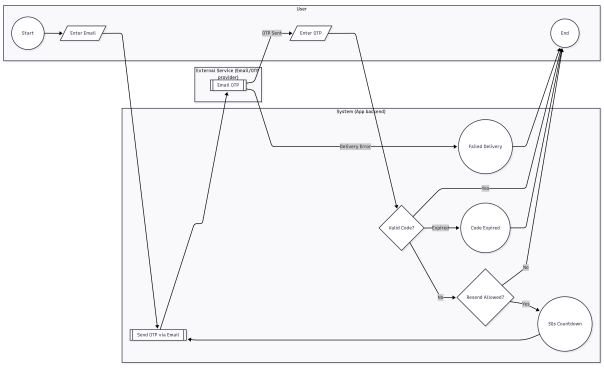

Fig. 10. BPMN diagram for two-step verification with resend cooldown.

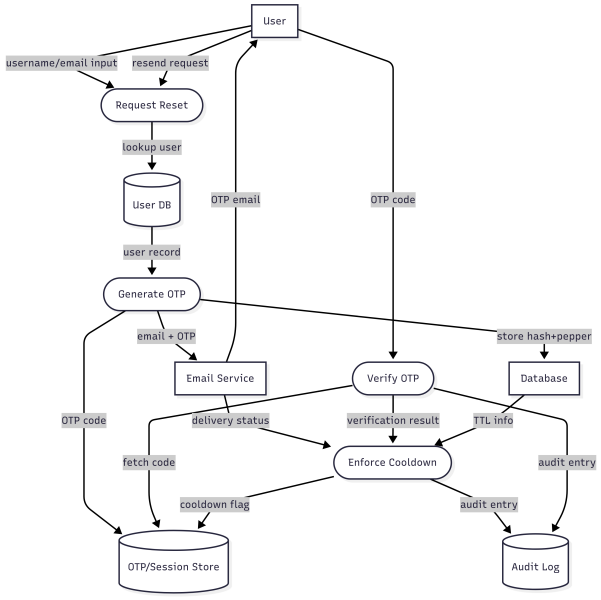

Fig. 11. OTP verification data-flow highlighting storage boundaries and audit trails.