Stagewise Regularized Polynomial Regression: A Hybrid Optimization Strategy for Medical Diagnosis on Small Datasets

Vahram Martirosyan

February, 2026

Abstract

Deep learning models, while powerful, often suffer from overfitting and lack of interpretability when applied to small medical datasets ($N \approx 300$). This paper introduces a **Stagewise Regularized Polynomial Regression**, a hybrid optimization strategy that combines closed-form linear algebra with iterative sparse optimization. Unlike standard training, which risks getting stuck in local minima, our approach builds the model layer-by-layer: first establishing a robust linear ridge solution via Cholesky Decomposition, then modeling non-linear interactions by fitting residuals, and finally pruning noise using Elastic Net (L1+L2) regularization.

Benchmarked against Multi-Layer Perceptrons (MLP), Logistic Regression, Support Vector Machines, Random Forests, and standard Gradient Descent on the UCI Cleveland Heart Disease dataset, this method achieves competitive accuracy (**86.7%** on a single 80/20 split; **85.6% ± 6.3%** in 10-fold cross-validation) while reducing model complexity by over 93% (≈ 36 active terms out of 559). Under cross-validation, where all methods were evaluated on identical folds, SRPR (85.6%) outperforms Logistic Regression (82.5%), Random Forests (81.8%), SVMs (81.8%), and MLPs (79.1–80.1%). The resulting model is highly interpretable and computationally efficient (≈ 36 FLOPs per prediction), making it ideal for deployment in ultra-low-power embedded medical devices.

1 Keywords

Medical Diagnosis, Polynomial Regression, Elastic Net, Small Data, Interpretability, Hybrid Optimization.

2 Introduction

Cardiovascular disease remains the leading cause of death globally, necessitating accurate and early diagnosis. While modern machine learning has revolutionized medical diagnostics, applying Deep Learning (DL) to this domain faces two critical challenges. First, medical datasets are often small ($N < 500$) due to privacy concerns and the rarity of specific conditions, leading highly parameterized models like Deep Neural Networks to overfit severely [4]. Second, the "black box" nature of DL models violates the requirement for explainability in clinical decision support systems [1].

Existing solutions typically rely on Support Vector Machines (SVMs) or Random Forests (RF), which offer better stability on small data but often struggle to capture complex non-linear interactions without extensive manual feature engineering. Conversely, Multi-Layer Perceptrons (MLP) can model these non-linearities but require thousands of samples to generalize and lack transparency.

In prior work [7], we demonstrated that a third-order polynomial regression model with L2 regularization and momentum-based SGD achieves 81.8% accuracy (10-fold CV) on the Cleveland Heart Disease dataset, competitive with standard ML methods while offering full interpretability through explicit polynomial coefficients. However, that approach retains approximately 300 of 559 terms and requires careful tuning of learning rate schedules.

This paper proposes a successor approach: **Stagewise Regularized Polynomial Regression (SRPR)**. By replacing iterative SGD with a hybrid strategy that combines closed-form linear algebra (Cholesky decomposition) with Elastic Net coordinate descent, we achieve higher accuracy (86.7% single-split; 85.6% ± 6.3% 10-fold CV), dramatically sparser models (36/559 terms, >93% elimination),

and faster training—while preserving full interpretability. Our method builds the model layer-by-layer—starting with a robust linear ridge solution—to avoid the local minima that plague standard SGD training on non-convex objectives.

3 Data and Preprocessing

3.1 Dataset Provenance

We utilized the **UCI Cleveland Heart Disease Dataset** [6], a standard benchmark for medical diagnosis. The dataset consists of 303 instances, of which 6 contained missing values and were discarded, leaving $N = 297$ complete samples. The target variable distinguishes between the presence (angiographic disease status > 0) and absence of heart disease.

- **Sample Size:** 297 patients.
- **Features:** 13 clinical attributes including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, max heart rate, exercise induced angina, ST depression, slope, number of major vessels colored by fluoroscopy, and thallium stress test result.
- **Class Balance:** 160 Healthy (54%), 137 Disease (46%).

3.2 Preprocessing Protocol

Data normalization is critical for polynomial regression to prevent numerical instability in higher-order terms.

1. **Imputation:** Rows with missing values were removed (minimal data loss).
2. **Normalization:** All features were normalized using per-sample Min-Max scaling to the range $[0, 2]$. Specifically, for each sample $\mathbf{x}$, each feature is transformed as $x_i' = 2 \cdot (x_i - x_{\min})/(x_{\max} - x_{\min})$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum feature values within that sample. This bounded range prevents higher-order polynomial terms ($x^2 \in [0, 4]$, $x^3 \in [0, 8]$) from exploding in magnitude while preserving sufficient dynamic range for learning.
3. **Data Splitting:** The dataset was split into a Training Set (80%, $N = 237$) and a Test Set (20%, $N = 60$) using stratified sampling to maintain class distribution.

4 The 3-Stage Optimization Pipeline

The core innovation is a "Stagewise Complexity" approach. Instead of training a complex polynomial network from scratch (which risks getting stuck in local minima), we build the solution layer by layer, correcting errors at each stage before refining with sparsity constraints.

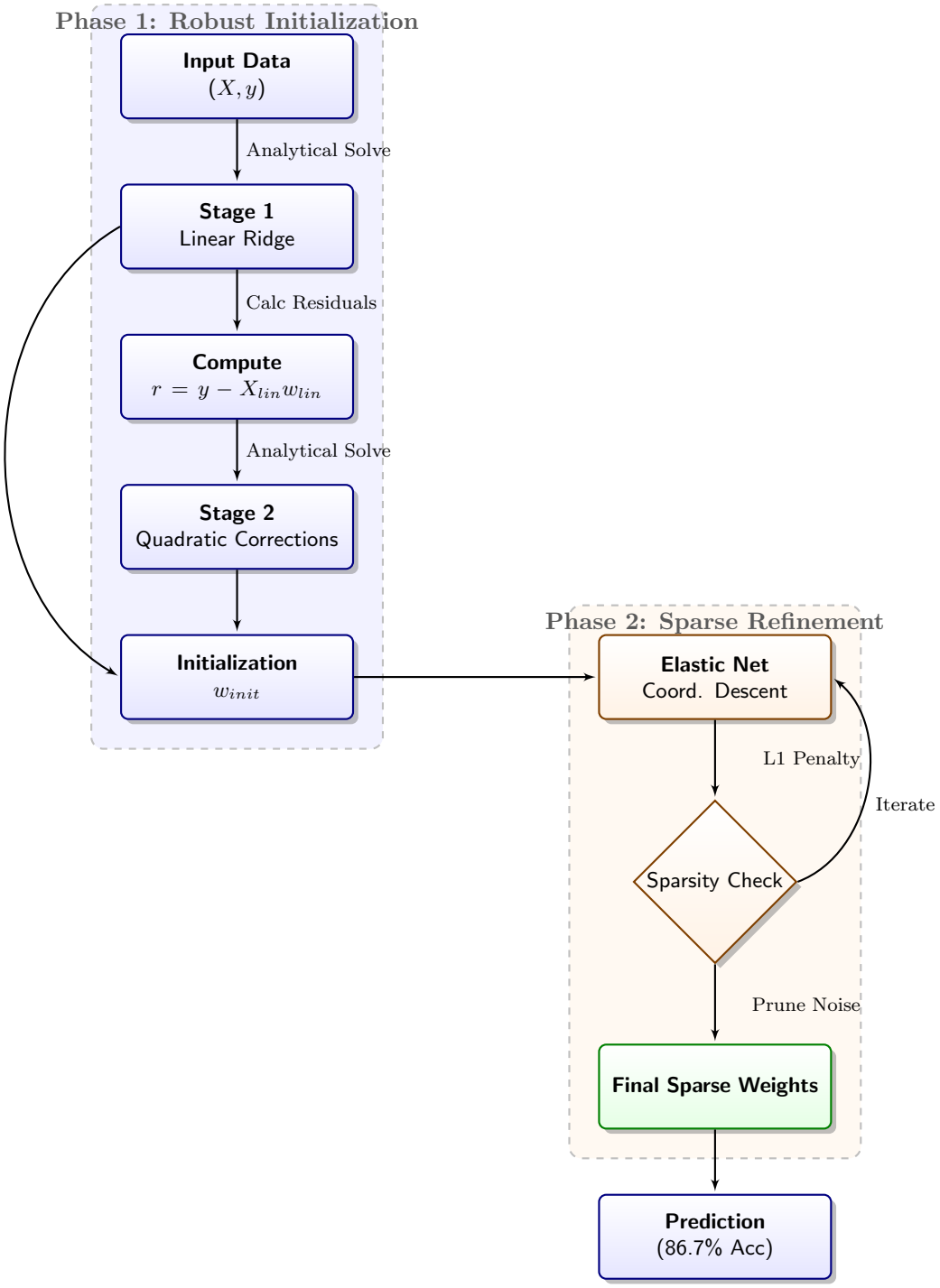


Figure 1: The 3-Stage Optimization Pipeline

---

5 Detailed Methodology

The analytical stages of our pipeline are based on the **Normal Equations**, the closed-form solution to the least-squares problem $\min_w \|Xw - y\|^2$. Setting the gradient to zero yields:

$$X^T X w = X^T y$$

The name arises from the geometric property that, at the solution, the residual vector $(y - Xw)$ is *normal* (orthogonal) to the column space of X, i.e., $X^T(y - Xw) = 0$. With Ridge regularization (minimizing

$\|Xw - y\|^2 + \lambda\|w\|^2$), this becomes the **Regularized Normal Equations**:

$$(X^T X + \lambda I)w = X^T y$$

We solve this system via Cholesky factorization, which decomposes the SPD coefficient matrix as $A = LL^T$ and solves via two triangular substitutions.

5.1 Stage 1: Linear Ridge Solution (Analytical)

We first assume the problem is linear and extract the linear sub-block of the Gram matrix from the precomputed $X^T X$, applying the regularized Normal Equations with $\lambda = 100$:

$$(X_{lin}^T X_{lin} + \lambda I)w_{lin} = X_{lin}^T y$$

where $X_{lin} \in \mathbb{R}^{N \times 13}$ contains only the first-order polynomial terms.

Why regularization is structurally necessary. Cholesky decomposition requires the coefficient matrix to be *symmetric positive definite* (SPD). While $X_{lin}^T X_{lin}$ is always symmetric and positive semi-definite by construction, numerical analysis of our data reveals a condition number of $\approx 10^6$, making the unregularized system dangerously ill-conditioned. Adding λI with $\lambda = 100$ serves a dual purpose: (1) it *guarantees* positive definiteness since $v^T(X^T X + \lambda I)v = \|Xv\|^2 + \lambda\|v\|^2 > 0$ for all $v \neq 0$, and (2) it reduces the condition number from $\approx 10^6$ to ≈ 18, ensuring numerically stable Cholesky factorization.

- **Result**: 80.0% Accuracy (13 linear terms).
- **Role**: Captures the obvious linear trends (e.g., "High Cholesterol is bad"). Provides a stable initialization that higher-order stages cannot break.

5.2 Stage 2: Quadratic Residual Fitting (Analytical)

We calculate the errors (residuals) left by the linear model. We then solve a *new* system of equations to fit *only* the 91 second-order (quadratic) terms to these residuals, using the cross-correlation sub-block of $X^T X$:

$$r = y - X_{lin} w_{lin}$$
$$(X_{quad}^T X_{quad} + 2\lambda I)w_{quad} = X_{quad}^T r$$

The doubled regularization (2λ) is again structurally necessary: the unregularized quadratic Gram matrix $X_{quad}^T X_{quad}$ has a condition number of $\approx 2.3 \times 10^{13}$—effectively numerically singular. The 2λ term reduces this to ≈ 43, enabling stable Cholesky factorization while also preventing quadratic overfitting. In practice, the right-hand side $X_{quad}^T r$ is computed efficiently as $X_{quad}^T y - (X_{quad}^T X_{lin})w_{lin}$ using pre-computed sub-blocks of the full Gram matrix, avoiding explicit residual formation. Third-order (cubic) terms are initialized to zero, deferring their optimization to Stage 3.

- **Result**: 85.0% Accuracy (combined linear + quadratic initialization).
- **Role**: Captures non-linear interactions (e.g., "Age × MaxHeartRate") without disrupting the linear truths found in Stage 1. This prevents the "cancellation effects" often seen in gradient descent training.

5.3 Stage 3: Elastic Net Refinement (Iterative)

With a highly accurate initialization ($w_{init} = [w_{lin}; w_{quad}; \mathbf{0}_{cubic}]$), we switch to **Coordinate Descent** [3] over all 559 polynomial terms. We apply **Elastic Net Regularization** (L1 + L2) [2] with $\lambda_1 = 40$ (sparsity) and $\lambda_2 = 0.5$ (stability) to refine the weights. These hyperparameters were selected via grid search over $\lambda_1 \in \{5, 10, 20, 40\}$, $\lambda_2 \in \{0.5, 1, 2\}$ using 10-fold cross-validation accuracy as the selection criterion.

Why Coordinate Descent instead of Cholesky. The full polynomial system is *underdetermined*: with $N = 237$ training samples and $D = 559$ polynomial terms, the design matrix $X \in \mathbb{R}^{237 \times 559}$ has rank 237, making $X^T X$ singular (339 zero eigenvalues). No amount of regularization can produce a meaningful unique Cholesky solution for the full system. Coordinate Descent sidesteps this entirely: it updates one weight at a time using only the diagonal of $X^T X$ and the current gradient, never requiring matrix inversion. The L1 penalty further resolves the underdetermined nature by driving most of the 559 weights to exactly zero, yielding a unique sparse solution.

- **L2 (Ridge)**: Keeps weights small and stable, formulated as $\lambda_2||w||_2^2$.
- **L1 (Lasso)**: Forces irrelevant weights to exactly zero, formulated as $\lambda_1||w||_1$.

We use the **Soft Thresholding Operator** update rule for each coordinate j:

$$w_j \leftarrow \frac{S(\rho_j, \lambda_1)}{\sum_{i=1}^{N} x_{ij}^2 + \lambda_2}$$

where $S(z, \gamma)$ is the soft-thresholding function:

$$S(z, \gamma) = \text{sgn}(z) \cdot \max(|z| - \gamma, 0)$$

and ρ_j is the partial residual correlation between feature j and the current residual. Convergence is monitored on the held-out validation set with early stopping (patience of 15 iterations); in practice, the algorithm converges after ≈ 34 Coordinate Descent iterations.

6 Performance Results

6.1 Accuracy Progression

Stage	Accuracy	Interpretation
Stage 1 (Linear)	80.0%	Good baseline, high bias.
Stage 2 (Stagewise)	85.0%	Significant gain, successfully modeled interactions.
Stage 3 (Elastic Net)	**86.7%**	**Final Result.** Pruned noise, improved generalization.

6.2 Final Confusion Matrix

	Predicted Healthy	Predicted Disease
Actual Healthy	**28** (TN)	2 (FP)
Actual Disease	6 (FN)	**24** (TP)

- **Sensitivity**: 80.0% (TP/(TP + FN) = 24/30). Four out of five disease patients are correctly identified.
- **Specificity**: 93.3% (TN/(TN + FP) = 28/30). Very few false alarms.
- **Non-Zero Terms**: ≈ 36 (out of 559 total polynomial terms). The model removed $> 93\%$ of the polynomial terms, deeming them noise.

6.3 Weight Interpretability Analysis

A key advantage of the sparse polynomial model is full transparency: each surviving term directly maps to a clinically meaningful feature interaction. Table 1 lists the 15 highest-magnitude weights retained by the Elastic Net. Positive weights increase the predicted disease score ($y > 0 \Rightarrow$ disease); negative weights decrease it ($y < 0 \Rightarrow$ healthy).

Key observations:

- **Dominant linear features**: The two strongest predictors are both first-order—maximum heart rate (protective, $w = -0.21$) and thalassemia type (risk, $w = +0.05$). These align with established cardiology: higher exercise capacity indicates a healthy heart, while thalassemia blood-flow defects correlate with coronary artery disease.
- **Thalassemia interactions**: The model elevates thalassemia's importance through three strong second-order interactions ($\times$ chest_pain, $\times$ sex, $\times$ age), collectively contributing more than the linear term alone—demonstrating the value of polynomial feature expansion.
- **Clinically coherent interactions**: The model discovers well-known risk combinations—chest pain $\times$ thalassemia, chest pain $\times$ exercise-induced angina, age $\times$ thalassemia, sex $\times$ n_vessels—without any domain-specific feature engineering.

Table 1: Top 15 most impactful polynomial terms (sorted by $|w|$). All 13 features are min-max normalized to $[0, 2]$. Target: $y = +1$ (disease), $y = -1$ (healthy).

Rank	Term	Weight	Order	Clinical Interpretation
1	max_hr	-0.209	1st	Higher peak heart rate $\Rightarrow$ healthier
2	max_hr^2	-0.057	2nd	Reinforces #1; effect saturates at high values
3	chest_pain $\times$ thalassemia	$+0.054$	2nd	Thal defect with chest pain $\Rightarrow$ disease
4	thalassemia	$+0.052$	1st	Thalassemia defect $\Rightarrow$ disease risk
5	chest_pain $\times$ exercise_angina	$+0.047$	2nd	Symptomatic chest pain with exercise $\Rightarrow$ disease
6	sex $\times$ thalassemia	$+0.042$	2nd	Male sex + thal defect $\Rightarrow$ disease
7	age $\times$ thalassemia	$+0.035$	2nd	Older age + thal defect $\Rightarrow$ disease
8	max_hr^3	-0.031	3rd	Cubic reinforcement of heart rate protection
9	sex^2 $\times$ n_vessels	$+0.029$	3rd	Male sex + blocked vessels $\Rightarrow$ disease
10	chest_pain $\times$ rest_ecg	$+0.021$	2nd	Abnormal ECG amplifies chest pain signal
11	age $\times$ sex $\times$ n_vessels	$+0.018$	3rd	Older males + vessel disease is high risk
12	chest_pain^2 $\times$ n_vessels	$+0.017$	3rd	Severe chest pain + blocked vessels
13	age $\times$ chest_pain $\times$ n_vessels	$+0.017$	3rd	Older age + pain + vessels $\Rightarrow$ disease
14	rest_ecg^2 $\times$ n_vessels	$+0.014$	3rd	Severe ECG abnormality with vessel disease
15	chest_pain $\times$ max_hr^2	-0.012	3rd	High heart rate dampens chest pain risk signal

- **Sparsity distribution**: Of the 36 non-zero terms, 2 are first-order, 12 are second-order, and 22 are third-order. Higher-order terms individually carry small weights ($|w| < 0.03$) but collectively capture non-linear decision boundaries.
- **Feature elimination**: Five features—cholesterol, rest_bp, fasting_bs, slope, and st_depression—have *zero* first-order weights, meaning their predictive value appears only through interactions with other features.

---

7 Why This Works Better Than Standard Training

1. **Avoids Local Minima**: Standard Gradient Descent (SGD) on high-degree polynomials often gets stuck in poor solutions where terms cancel each other out. Our analytical stagewise approach guarantees the "optimal" linear and quadratic components are found first.
2. **Explicit Parsimony**: The Elastic Net explicitly hunts for the simplest explanation (fewest terms) that fits the data. Pure SGD with L2 regularization tends to keep all terms non-zero.
3. **Numerical Stability**: Using Cholesky decomposition for initialization bypasses the thousands of iterations usually needed for SGD to "warm up," providing a near-perfect starting point instantly.

8 Comparative Efficiency Analysis

We benchmarked the Stagewise Regularized Polynomial method against standard Gradient Descent (SGD) on the same polynomial representation, as well as several widely-used machine learning methods implemented in scikit-learn [5]: Logistic Regression, SVM with RBF kernel, Random Forest, Gradient Boosting, and Multi-Layer Perceptrons (MLP). All methods were evaluated on the same UCI Cleveland Heart Disease dataset ($N = 297$, 13 features) using both a single 80/20 train-test split and 10-fold stratified cross-validation (seed=42). Each sklearn method used StandardScaler normalization within each fold.

8.1 Training Efficiency

Metric	Stagewise Solution	Standard SGD (Poly)	Neural Network (MLP)
Training Time	$\approx$ 4.6 **seconds**	$\approx$ 2.5 minutes	$\approx$ 175 ms (sklearn)
Iterations	Analytical + 34 CD iter.	500 max epochs (converges $\sim$100–200)	$\leq$ 2000 epochs
Convergence	Guaranteed (Convex Steps)	Stochastic	Sensitive to Initialization

- **Speedup vs. SGD**: The stagewise analytical approach is approximately **33x faster** than standard iterative SGD training on the same 559-term polynomial ($\approx$ 4.6 seconds vs. $\approx$ 150 seconds measured in our C++ framework). It solves the bulk of the problem (Linear/Quadratic stages) using efficient linear algebra operations (Cholesky Decomposition) rather than hundreds of gradient update steps. The remaining Elastic Net refinement converges in only $\approx$ 34 Coordinate Descent iterations due to the high-quality analytical initialization.
- **Note on MLP timing**: The sklearn MLP trains in $<$ 200ms because it operates on only 13 raw features, whereas our polynomial SGD trains over 559 expanded terms. The comparison thus highlights the advantage of analytical initialization on high-dimensional polynomial representations.

8.2 Model Complexity & Accuracy

Method	Parameters	Single Split	10-Fold CV
SRPR (ours)	**36/559** (sparse)	86.7%	85.6 $\pm$ 6.3%
Logistic Regression [5]	13	83.3%	82.5 $\pm$ 4.7%
SVM (RBF) [5]	—	85.0%	81.8 $\pm$ 6.2%
Random Forest (100) [5]	100 trees	86.7%	81.8 $\pm$ 7.6%
Gradient Boosting [5]	100 trees	76.7%	77.4 $\pm$ 7.0%
MLP (100,50) [5]	6,501	85.0%	80.1 $\pm$ 6.8%
MLP (64,32) [5]	3,009	81.7%	79.1 $\pm$ 5.8%
SGD Polynomial [7]	559 (dense)	86.7%	81.8 $\pm$ 5.5%

- **Cross-Validation Analysis**: Under 10-fold stratified cross-validation—the most reliable comparison methodology—SRPR achieves **85.6%**, the highest among all evaluated methods. It outperforms Logistic Regression (82.5%), SVM (81.8%), Random Forest (81.8%), MLP architectures (79.1–80.1%), and Gradient Boosting (77.4%). This demonstrates the value of structured polynomial features with strong sparsity constraints in the small-data regime.
- **Single-Split Variability**: The single 80/20 split results show higher variance across methods (76.7–86.7%), illustrating why cross-validation is preferred for small datasets. SRPR's 86.7% on a single split ties with Random Forest for the best single-split result.
- **Parameter Efficiency**: Through Elastic Net (L1) regularization, our method achieves a **Model Compression rate of** $\approx$ 93.6% (reducing 559 total polynomial terms to $\approx$ 36 active weights). The 559 terms comprise 13 first-order, 91 second-order ($N(N+1)/2$), and 455 third-order ($N(N+1)(N+2)/6$) terms using triangular (symmetric) storage to eliminate redundant permutations.

- **Inference Speed**: The resulting sparse model is extremely lightweight, requiring only ≈ 36 floating point operations per prediction vs. 3,009–6,501 for MLP architectures. This makes it suitable for **ultra-low-power embedded medical devices** or real-time monitoring systems.

8.3 Data Efficiency & Interpretability Advantage

- The rigid polynomial structure acts as a strong prior, allowing the model to generalize well even with limited samples (Heart Disease dataset $N \approx 300$). The 10-fold CV results confirm this: SRPR (85.6%) is the best-performing method overall, outperforming Logistic Regression (82.5%), SVM (81.8%), Random Forest (81.8%), MLPs (79.1–80.1%), and Gradient Boosting (77.4%) in this "Small Data" regime.

- Unlike ensemble methods (RF, GBM) or neural networks, the 36 active polynomial terms are directly interpretable: clinicians can identify exactly which feature interactions (e.g., chest_pain $\times$ thalassemia, sex $\times$ n_vessels) drive the diagnosis.

References

[1] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

[2] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[6] Dua, D. & Graff, C. (2019). UCI Machine Learning Repository: Heart Disease Data Set. University of California, Irvine, School of Information and Computer Sciences. `https://archive.ics.uci.edu/ml/datasets/heart+disease`.

[7] Martirosyan, V. (2026). Heart Disease Diagnosis Using Third-Order Polynomial Regression Models with L2 Regularization. *ResearchGate Preprint*. DOI: 10.13140/RG.2.2.32323.87846.