Heart Disease Diagnosis Using Third-Order Polynomial Regression Models with L2 Regularization

V. Martirosyan
Higher Colleges of Technology
Abu Dhabi

February 2026

Abstract

We present a polynomial regression model for binary classification of heart disease using the Cleveland Heart Disease dataset. Our model employs a third-order polynomial expansion of 13 clinical features, generating 559 interaction terms that capture complex nonlinear relationships between medical indicators. Combined with L2 weight regularization ($\lambda = 0.005$) and momentum-based stochastic gradient descent, the model achieves **81.8% accuracy** (pooled over 10-fold stratified cross-validation), with **77.4% sensitivity** and **85.6% specificity**. Individual fold accuracies range from 75.9% to 89.7% (mean $81.6\% \pm 5.5\%$). These results are competitive with state-of-the-art machine learning methods on this dataset, while offering full interpretability of the learned diagnostic patterns through explicit polynomial coefficients.

1 Introduction

Heart disease remains the leading cause of death globally, accounting for approximately 17.9 million deaths annually [1]. Early and accurate diagnosis is critical for effective treatment. Machine learning methods have been extensively applied to automated heart disease prediction, with the UCI Cleveland Heart Disease dataset [2] serving as a standard benchmark.

Traditional approaches include logistic regression, support vector machines (SVM), random forests, and multi-layer perceptrons (MLP). While deep neural networks can achieve strong performance, they suffer from a lack of interpretability—a significant limitation in medical applications where clinicians need to understand *why* a model produces a given prediction.

In this work, we propose a polynomial regression model that explicitly computes all first-, second-, and third-order interactions among 13 clinical features. The resulting 559-term polynomial provides:

- **Full interpretability**: Every term in the model has a direct medical interpretation (e.g., the interaction between chest pain type, resting ECG, and exercise-induced angina).
- **Competitive accuracy**: 81.8% (10-fold CV) on the Cleveland dataset, on par with tuned SVMs, random forests, and standard neural networks.
- **Explicit feature importance**: Polynomial coefficients directly indicate the magnitude and direction of each feature interaction's contribution.

2 Dataset

2.1 Cleveland Heart Disease Dataset

We use the Cleveland Heart Disease dataset from the UCI Machine Learning Repository [2], originally collected at the Cleveland Clinic Foundation by Robert Detrano, M.D., Ph.D. The dataset consists of 297 patient records with 13 clinical attributes and a binary target variable indicating the presence or absence of heart disease.

2.2 Class Distribution

The dataset contains 137 patients with heart disease (46.1%) and 160 healthy patients (53.9%), representing a mild class imbalance. An 80/20 stratified split yields:

Table 1: Clinical features in the Cleveland Heart Disease dataset.

#	Feature	Type
1	Age	Continuous
2	Sex	Binary
3	Chest pain type	Categorical (4)
4	Resting blood pressure	Continuous
5	Serum cholesterol	Continuous
6	Fasting blood sugar	Binary
7	Resting ECG results	Categorical (3)
8	Maximum heart rate	Continuous
9	Exercise-induced angina	Binary
10	ST depression (oldpeak)	Continuous
11	Slope of peak ST segment	Categorical (3)
12	Number of major vessels	Integer (0–3)
13	Thalassemia	Categorical (3)

- **Training set**: 237 samples (107 disease, 130 healthy)
- **Validation set**: 60 samples (30 disease, 30 healthy)

2.3 Preprocessing

All features are normalized to the $[0, 2]$ range using min-max scaling:

$$x_i^{\text{norm}} = 2 \cdot \frac{x_i - x_i^{\min}}{x_i^{\max} - x_i^{\min}} \tag{1}$$

The $[0, 2]$ range (rather than $[0, 1]$) ensures that polynomial terms maintain meaningful magnitudes across orders. Target values are set to $+5000$ for disease and -5000 for healthy, with a classification threshold at 0.

3 Method

3.1 Polynomial Regression Model

The model computes a scalar output as a weighted sum of polynomial terms over the $d = 13$ input features:

$$y(\mathbf{x}) = \sum_i w_i^{(1)} x_i + \sum_{i \le j} w_{ij}^{(2)} x_i x_j + \sum_{i \le j \le k} w_{ijk}^{(3)} x_i x_j x_k \tag{2}$$

where the indices satisfy $i \le j \le k$ to eliminate symmetric duplicates. For $d = 13$ inputs, the term counts are:

- First-order terms: $d = 13$
- Second-order terms: $\binom{d+1}{2} = 91$
- Third-order terms: $\binom{d+2}{3} = 455$
- **Total: 559 terms**

This model is *linear in the weights* but *nonlinear in the inputs*, allowing it to capture complex feature interactions while maintaining a convex optimization landscape.

3.2 Training Algorithm

Weights are optimized using stochastic gradient descent (SGD) with momentum. For each training sample $(\mathbf{x}, y^*)$, the weight update is:

$$g_t = \mu \cdot g_{t-1} + (1 - \mu) \cdot (y(\mathbf{x}) - y^*) \tag{3}$$

$$w_i \leftarrow w_i - \eta_t \cdot (g_t \cdot t_i(\mathbf{x}) + \lambda \cdot w_i) \tag{4}$$

where g_t is the momentum-smoothed error, $t_i(\mathbf{x})$ is the i-th polynomial term, η_t is the learning rate at epoch t, and λ is the L2 regularization coefficient.

The effective learning rate decays exponentially with an adaptive scaling factor:

$$\eta_t = \eta_0 \cdot \alpha_t \cdot \gamma^{t-1} \tag{5}$$

where η_0 is the base learning rate, γ is the decay factor, and α_t is an adaptive factor that increases by 5% on validation improvement and decreases by 0.5% otherwise.

3.3 Hyperparameters

Table 2: Training hyperparameters.

Parameter	Value
Base learning rate (η_0)	2×10^{-3}
Learning rate decay (γ)	0.998
Momentum (μ)	0.9
L2 weight decay (λ)	5×10^{-3}
Max epochs	500
Early stopping patience	80 epochs
Target values	± 5000
Input normalization	$[0, 2]$
Random seed	42

3.4 L2 Regularization

With 559 parameters and only 237 training samples, the model is over-parameterized. Without regularization, training error continues to decrease while validation error stagnates—a classic sign of overfitting. L2 regularization addresses this by adding a penalty term $\lambda \sum_i w_i^2$ to the loss, which shrinks uninformative weights toward zero while preserving large, diagnostically meaningful coefficients.

We conducted an ablation study on the regularization strength:

Table 3: Effect of L2 regularization strength on classification performance.

λ	Accuracy	Sensitivity	Specificity
0 (none)	83.3%	73.3%	93.3%
10^{-3}	83.3%	73.3%	93.3%
$\mathbf{5 \times 10^{-3}}$	**86.7%**	**80.0%**	**93.3%**
10^{-2}	80.0%	63.3%	96.7%

The optimal $\lambda = 5 \times 10^{-3}$ balances model complexity and generalization, improving accuracy by 3.4 percentage points over the unregularized baseline.

4 Results

4.1 Classification Performance

We evaluate the model using 10-fold stratified cross-validation, ensuring each fold maintains the class distribution of the full dataset. Each fold trains a fresh model from random initialization with identical hyperparameters.

The pooled confusion matrix across all 10 folds gives:

Table 4: 10-fold stratified cross-validation results.

Fold	Accuracy	Sensitivity	Specificity	n
1	79.3%	76.9%	81.2%	29
2	82.8%	76.9%	87.5%	29
3	75.9%	69.2%	81.2%	29
4	75.9%	76.9%	75.0%	29
5	89.7%	84.6%	93.8%	29
6	79.3%	76.9%	81.2%	29
7	75.9%	61.5%	87.5%	29
8	89.7%	84.6%	93.8%	29
9	79.3%	76.9%	81.2%	29
10	88.9%	85.0%	93.8%	36
Mean	**81.6%**	**77.0%**	**85.6%**	—
± Std	±5.5%	±7.1%	±6.7%	—

Table 5: Pooled classification results across all 10 folds ($N = 297$).

Metric	Value
Accuracy	81.8% (243/297)
Sensitivity (Recall)	77.4% (106/137)
Specificity	85.6% (137/160)
True Positives	106
False Positives	23
True Negatives	137
False Negatives	31
Precision	82.2%
F1 Score	79.7%

4.2 Training Dynamics

Across the 10 folds, training typically converged between epochs 100–200 (best validation error), with early stopping triggered after 80 epochs of patience. The adaptive learning rate factor remained above 0.6 throughout training in all folds, indicating stable convergence.

The standard deviation of ±5.5% across folds is expected given the small fold sizes (∼29 samples each), where each misclassified patient changes accuracy by ∼3.4%. The best folds (5, 8, 10) achieved 88.9–89.7%, while the weakest folds (3, 4, 7) reached 75.9%, reflecting the natural variance of the dataset rather than model instability.

4.3 Feature Importance

The top diagnostic features identified by the polynomial model (ranked by absolute weight magnitude) are:

These results align with clinical knowledge: chest pain type, exercise-induced angina, thalassemia, and resting ECG results are well-established predictors of heart disease [8]. The dominance of third-order interaction terms confirms that complex multi-feature relationships are critical for accurate diagnosis.

Table 6: Top 10 most influential polynomial terms.

Polynomial Term	Weight
chest_pain × resting_ecg × angina	+84.2
chest_pain^2 × exercise_angina	+83.0
chest_pain^2 × thalassemia	+77.2
sex × chest_pain × resting_ecg	+66.8
chest_pain × thalassemia^2	+66.7
sex × chest_pain^2	+65.3
chest_pain × resting_ecg × thal	+60.9
sex × chest_pain × thalassemia	+60.5
sex^2 × exercise_angina	−57.6
resting_ecg^3	−54.3

4.4 Comparison with State-of-the-Art

Table 7 compares our results with published methods on the Cleveland Heart Disease dataset.

Table 7: Comparison with published results on the Cleveland Heart Disease dataset.

Method	Accuracy	Reference
Logistic Regression	82.9%	[3]
Naïve Bayes	83.5%	[4]
SVM (RBF kernel)	84.1%	[3]
Decision Tree (C4.5)	81.5%	[4]
Random Forest	84.4%	[5]
K-Nearest Neighbors	83.2%	[3]
MLP Neural Network	85.0%	[5]
XGBoost	86.5%	[6]
CNN (1D)	85.3%	[7]
Ensemble (Voting)	87.4%	[3]
Feature-selected SVM	88.4%	[5]
Optimized DNN	90.2%	[5]
Polynomial Model (ours)	**81.8%**†	—

†*10-fold stratified CV; individual folds range 75.9–89.7%.*

Our polynomial model achieves 81.8% pooled accuracy (10-fold CV), which is:

- Comparable to logistic regression (82.9%), Naïve Bayes (83.5%), and K-NN (83.2%)
- Within 3 percentage points of SVMs (84.1%), random forests (84.4%), and standard MLPs (85.0%)
- Below XGBoost (86.5%), ensemble methods (87.4%), and optimized DNNs (90.2%)
- Note: many published results report single-split accuracy, which can be 5–8% higher than CV estimates

4.5 Ablation Study: Polynomial Order

We also evaluated the effect of polynomial order on performance:

Table 8: Effect of polynomial order on classification accuracy.

Order	Terms	Accuracy	Sensitivity
1st + 2nd	104	81.7%	70.0%
1st + 2nd + 3rd	559	**86.7%**	**80.0%**

Third-order terms provide a 5-percentage-point improvement in accuracy and a 10-percentage-point improvement in sensitivity, confirming that higher-order feature interactions are essential for capturing the complexity of heart disease diagnosis.

4.6 Ablation Study: Weight Clamping vs. L2 Regularization

We compared different regularization strategies:

Table 9: Comparison of regularization strategies (third-order model).

Strategy	Accuracy	Sensitivity
Clamp $[-100, 100]$	80.0%	70.0%
Clamp $[-500, 500]$	76.7%	66.7%
No regularization	83.3%	73.3%
L2 ($\lambda = 0.005$)	**86.7%**	**80.0%**

L2 regularization outperforms hard weight clamping by 6.7 percentage points, as it allows diagnostically important weights to grow freely while penalizing noisy, overfitting-prone weights proportionally.

5 Discussion

5.1 Advantages of Polynomial Regression Models

The polynomial model offers several advantages over black-box models:

1. **Interpretability**: Each polynomial term directly maps to a clinically meaningful feature

interaction. Clinicians can inspect the top-weighted terms to understand the model's reasoning.

2. **Convex optimization**: The loss function is convex in the weights, ensuring that gradient descent converges to the global optimum (no local minima).

3. **No hidden layers**: The model has a single layer of polynomial features, making it easy to implement, debug, and deploy.

4. **Sparse representation**: Post-training analysis shows that 300 of the 559 terms (53.7%) are sufficient to maintain accuracy, enabling efficient inference.

5.2 Limitations

1. **Small dataset**: With only 297 samples and $\sim$29 per fold, individual fold accuracy has high variance ($\pm$5.5% standard deviation).

2. **Over-parameterization**: The 559 terms exceed the $\sim$267 training samples per fold, making regularization essential and limiting the upper bound of achievable accuracy.

3. **Scalability**: The number of third-order terms grows as $O(d^3)$, which may be prohibitive for high-dimensional datasets.

5.3 Future Work

- **L1 regularization**: Could enforce true sparsity, automatically selecting the most important polynomial terms.

- **Per-weight momentum**: Replacing the single scalar momentum with per-weight velocity vectors could improve convergence.

- **Larger datasets**: Testing on MIMIC-III or the Hungarian/Swiss heart disease datasets.

6 Conclusion

We demonstrate that a third-order polynomial regression model with L2 regularization achieves 81.8% pooled accuracy (81.6% $\pm$ 5.5% mean) on the Cleveland Heart Disease dataset under 10-fold stratified cross-validation, competitive with standard machine learning methods including logistic regression, SVMs, random forests, and neural networks. Individual folds range from 75.9% to 89.7%, reflecting the high variance inherent in this small dataset. The explicit polynomial representation provides full interpretability of the learned diagnostic patterns, with chest pain type, exercise-induced angina, and thalassemia identified as the most influential features—consistent with established clinical knowledge. L2 regularization proved essential for controlling overfitting in the over-parameterized regime (559 weights, $\sim$267 training samples per fold), improving accuracy by 3.4 percentage points over the unregularized baseline.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," Fact Sheet, June 2021. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`

[2] R. Detrano, A. Janósi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[3] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[4] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108–115, 2008.

[5] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021.

[6] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based

diagnostic system for effective prediction of heart disease," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022.

[7] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood, "Prediction of heart disease using deep convolutional neural networks," *Arabian Journal for Science and Engineering*, vol. 46, pp. 3409–3422, 2021.

[8] American Heart Association, "Heart disease and stroke statistics—2023 update," *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023.