OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Taylor-Series Engineering Approach to Machine Learning-Based Early Detection of Systemic Lupus Erythematosus

Harrison Baek
hbaek122@gmail.com

ABSTRACT

This study addresses the challenge of early detection of systemic lupus erythematosus (SLE), an autoimmune disease that is often diagnosed late due to its heterogeneous clinical presentation. I investigate whether Taylor-series–derived features can enhance machine learning (ML) models by capturing nonlinear relationships in gene expression data, thereby improving classification performance. I hypothesize that Taylor-series–derived features, which represent local curvature and gene–gene interactions, will significantly improve model accuracy compared to using original gene expression features alone. To evaluate this, I analyze gene expression data from 26 SLE patients and 46 controls using Logistic Regression, Random Forest, and Support Vector Machine models. I expand selected informative genes into second-degree polynomial features and apply Lasso regularization for feature selection before retraining the models. The Taylor-enhanced models increase average classification accuracy from 93.9% to 98.5%, with the largest improvement observed in the SVM model. Approximately 40% of the selected features involve gene–gene interactions associated with biologically relevant pathways, including interferon signaling and B-cell activation. However, these improvements are not statistically significant according to McNemar's test, likely due to the small sample size. Overall, the findings suggest that nonlinear Taylor-series–derived features can capture meaningful biological relationships and improve predictive performance, warranting further validation in larger datasets before any clinical application. Although performance improvements were observed, the lack of statistical significance indicates that these differences may be attributable to limited statistical power and should be interpreted cautiously.

Keywords: Systemic Lupus Erythematosus, Machine Learning, Feature Engineering, Taylor Series, Gene Expression, Bioinformatics, Autoimmune Disease Diagnosis

1. INTRODUCTION

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease characterized by systemic inflammation and the production of autoantibodies, resulting in heterogeneous and evolving clinical manifestations (National Institute of Arthritis and Musculoskeletal and Skin Diseases, 2022; Lazar & Kahlenberg, 2023). It affects ~5 million people worldwide and disproportionately impacts women (~9:1 female-to-male ratio), with common symptoms including arthritis, fatigue, fever, rash, and alopecia (Lupus Foundation of America, 2025; Alnaimat et al., 2025). Despite established classification criteria (The American College of Rheumatology (ACR) revised in 1997; Systemic Lupus International Collaborating Clinics (SLICC) in 2012) and laboratory tests such as ANA and anti‑dsDNA, the lack of definitive biomarkers and the disease's nonlinear course often lead to diagnostic delays of 3–5 years, increasing the risk of irreversible organ damage (Lazar & Kahlenberg, 2023; Mitchell, 2024).

Advances in bioinformatics and artificial intelligence (AI) have opened new opportunities for early detection by identifying complex molecular signatures in high-dimensional data. Machine-learning (ML) approaches, including random forests, support vector machines, and logistic regression, have shown promise in analyzing transcriptomic and immunological datasets (Gu et al., 2025; Zhou et al., 2025). However, SLE studies are often limited by small, imbalanced cohorts, increasing the risk of overfitting and reducing generalizability (Aliferis & Simon, 2024; Wang, 2024). In addition, methodological issues such as data leakage and limited interpretability can undermine reproducibility and clinical applicability (Kapoor & Narayanan, 2023; Aliferis & Simon, 2024).

Biologically, SLE is a polygenic and dynamic disease involving dysregulation across innate and adaptive immune pathways, particularly interferon signaling (Kumar et al., 2025). Molecular activity fluctuates over time due to disease progression, treatment, and environmental influences, suggesting that temporal dynamics—not just static measurements—are critical for understanding disease behavior (Kuehn et al., 2015). Nevertheless, most ML models rely on cross-sectional features and rarely encode local changes or gene–gene interactions explicitly, limiting their ability to reflect underlying disease mechanisms (Birmingham et al., 2010; NHS South Tees Hospitals, 2022).

To address this gap, this study introduces a Taylor-series–derived features framework that approximates local dynamics and interaction effects in gene-expression data. By incorporating first- and second-order terms, the approach captures proxies for rates of change, curvature, and pairwise interactions, enabling a more biologically aligned representation of SLE's nonlinear behavior even from cross-sectional data (Maddula, 2024; Kuehn et al., 2015; Jiang, 2022). We hypothesized that integrating Taylor-series–derived features would improve machine-learning classification performance for SLE compared to models trained on original gene expression data alone. Applied to the GSE39088 dataset, this framework generates higher-order features and selects a sparse, interpretable subset via L1 regularization, resulting in consistent performance improvements across multiple ML models while highlighting biologically plausible interaction effects (Leventhal et al., 2023).

Overall, this work provides a transparent, mathematically grounded framework that bridges dynamic disease biology and static ML representations. By embedding temporal and interaction-aware features into standard pipelines, it offers a scalable and interpretable approach for improving SLE detection and has potential applicability to other autoimmune diseases characterized by evolving pathophysiology (Ahlquist et al., 2023; Parodis et al., 2025).

Following the Introduction, Section 2 presents the research question and hypotheses, outlining how Taylor-series–derived features are expected to improve SLE classification. Section 3 details the methodology, including leakage-safe preprocessing, construction of degree-2 Taylor/polynomial features, L1-based feature selection, and model training and evaluation using logistic regression, random forest, and SVM. Section 4 reports the results, comparing baseline and Taylor-enhanced performance, and summarizes statistical tests (McNemar) and the composition of the selected features. Section 5 discusses biological interpretability, methodological strengths and limitations, and future directions (e.g., external multi-cohort validation and pathway mapping). Section 6 concludes with the implications of this calculus-grounded framework for earlier, more interpretable SLE detection and its potential generalization to other dynamic autoimmune diseases.

2. RESEARCH DESIGN

Systemic lupus erythematosus (SLE) is a biologically complex disease characterized by nonlinear interactions among genes and immune pathways. Standard machine learning (ML) models often rely on static, cross-sectional features, which may fail to capture these dynamics. To address this limitation, this study investigates whether Taylor-series–derived features, which encode nonlinear effects and gene–gene interactions, can improve ML model performance for SLE classification. This study is guided by the following research question: Can Taylor-series–derived features improve machine learning classification performance for SLE compared to models trained on original gene expression features?

2.1. Hypotheses

To address the research question, I specify the following hypotheses.

Null Hypothesis (H_0): Taylor-series–derived features do not produce a statistically significant improvement in machine learning performance for SLE classification compared with models trained on original gene-expression features alone.

Alternative Hypothesis (H_1): Taylor-series–derived features do produce a statistically significant improvement in machine learning performance for SLE classification.

Taylor-series–derived features approximate local slopes, curvatures, and gene–gene interactions, and are therefore expected to improve classification performance compared with models trained on raw, cross-sectional features.

Three machine learning models, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), were selected because they represent complementary modeling paradigms (linear, tree-based, and kernel-based). This allows us to assess whether performance gains arise from the feature representation rather than from a specific algorithm.

To evaluate the hypothesis, models were trained under two conditions: (1) a baseline pipeline using preprocessed raw features, and (2) a Taylor-enhanced pipeline incorporating polynomial features with L1-based feature selection. Accuracy on a held-out test set was used as the primary endpoint, with precision, recall (SLE class), F1 score, ROC–AUC, and confusion matrices as secondary evaluation metrics. Because predictions are paired, McNemar's test ($\alpha = 0.05$) was used to assess statistical significance, and effect sizes were calculated to evaluate practical significance under small-sample conditions.

To further interpret outcomes, five sub-questions were examined: which feature types drive performance, whether improvements are consistent across models, whether false negatives are reduced, whether selected features are stable and biologically interpretable, and the magnitude of improvements relative to sample size. These analyses position the study as a proof-of-concept motivating further validation.

3. METHODOLOGY

3.1. Data Acquisition and Preprocessing

3.1.1. Dataset Selection: Gene expression data were obtained from the publicly available Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi) from the National Center for Biotechnology Information (NCBI), accession number GSE39088. This dataset contains whole blood gene expression profiles from a systemic lupus erythematosus intervention study measured using the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570 platform), comprising 54,675 probe sets. Expression measurements in GSE39088 are reported as Affymetrix probe set identifiers from the GPL570 (U133 Plus 2.0) microarray platform (e.g., 1560926_at, 206025_s_at). These IDs correspond to probe sets targeting specific transcripts and may map one‑to‑one (suffix _at), to similar transcripts or gene families (_s_at), or to alternative splice forms (_a_at). For interpretability, probe IDs referenced in the Results were retained as identifiers at model‑training time and can be mapped post hoc to official gene symbols and annotations using the GPL570 platform file or standard resources (e.g., Bioconductor hgu133plus2.db, BioMart). This separation preserves leakage‑free modeling while enabling biological interpretation after evaluation.

3.1.2. Sample Selection Criteria: In order to ensure valid classification analysis, I applied stringent inclusion criteria. From the original 142 samples in GSE39088, I selected only baseline samples collected prior to any treatment intervention. Specifically, I included 46 healthy subjects (control) that had no treatment or in vitro stimulations, and 26 SLE patients at baseline (day 0, pre-treatment. All samples collected after treatment (day 112, day 168) and all in vitro stimulated samples were excluded to avoid confounding effects from therapy and artificial experimental conditions. This resulted in a final analytical cohort of 72 samples, labeled SLE = 1 and Healthy = 0.

3.1.3. Data Preprocessing Pipeline: All preprocessing steps were implemented in Python 3.x using standard scientific computing libraries (NumPy 1.24.3, Pandas 2.0.3, Scikit‑learn 1.3.0) within the Google Colaboratory environment. The GEOparse library (version 2.0.3) was used to extract the GSE39088 dataset. To prevent information leakage from the test set into the training process, preprocessing followed the strict sequence shown in Figure 1.

I first extracted the expression matrix with GEOparse and converted it into a samples‑by‑genes structure (72 samples × 54,675 genes), then parsed the corresponding metadata to assign binary class labels (SLE = 1; Healthy = 0). I next verified data integrity, no missing values were present across samples or probes, and all intensities were already log2‑transformed and approximately mean‑centered (mean ≈ 0; range −8.69 to +6.81), consistent with standard Affymetrix preprocessing.

To ensure a leakage‑free workflow, we performed a stratified 70/30 train–test split (n = 50/22; random seed = 42) before any feature‑dependent operation, preserving the original class ratio (Healthy:SLE ≈ 1.77:1). All subsequent preprocessing steps were fit exclusively on the training set and applied unchanged to the test set.

Specifically, I calculated gene‑wise variance on the training data and removed the lowest‑variance decile (threshold = 0.0405), reducing dimensionality from 54,675 to 49,207 genes (5,468 removed). Finally, I applied Z‑score standardization by fitting a scaler (gene‑wise mean and standard deviation) on the training set and using the same parameters to transform the test set.

This leakage‑safe sequence, split → filter → scale, preserves validation integrity, reduces noise and dimensionality, and produces a stable feature space for downstream modeling. All preprocessing parameters (variance thresholds and scaling factors) were derived exclusively from the training data and applied identically to the test set to maintain methodological consistency.

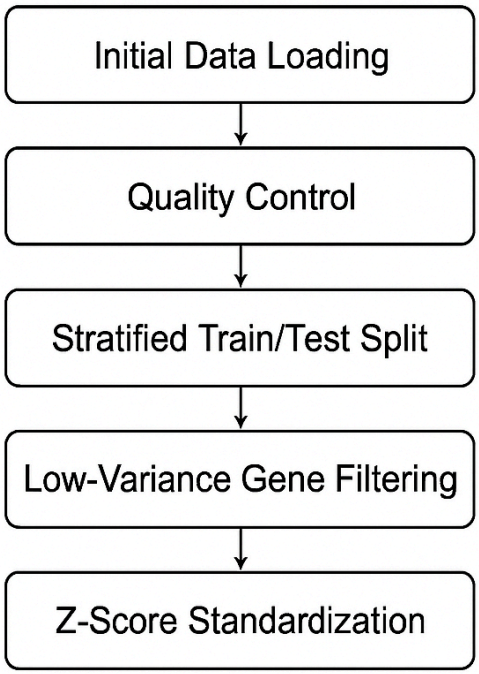

Figure 1: Flow diagram of Data Processing.

Note. Workflow for preparing the GSE39088 whole-blood gene-expression dataset prior to modeling. Steps proceed top-to-bottom: (1) Initial Data Loading (GEOparse extraction; samples × genes matrix and label parsing), (2) Quality Control (verify no missing values, $\log_2$ transform, and approximate mean centering), (3) Stratified Train/Test Split (70/30 split with class ratio preserved to prevent leakage), (4) Low-Variance Gene Filtering (remove lowest-variance decile computed on training only; apply the same mask to test), and (5) Z-Score Standardization (fit scaler on training; apply the learned means/SDs to both sets). This train-only fitting of all feature dependent operations ensures a leakage-safe evaluation.

3.2. Taylor-Series Feature Engineering

I model the unknown functional relationship between biomarkers and SLE status using a second-order multivariate Taylor expansion, yielding a locally valid approximation that incorporates linear effects and nonlinear interactions. To establish the mathematical foundation of the model, I used a second-degree Taylor expansion as in equation (1).

Specifically, let $y = f(x) = f(x_1, x_2, ..., x_n)$ where y represents the latent disease risk for SLE, xi denotes the gene expression level of the i-th gene. A second-order Taylor series of $f(x_1, x_2, ..., x_n)$ around a reference point $x_0 = (x_{01}, x_{02}, ..., x_{0n})$is given by

$$f(x)\approx f\left(x_0\right)+\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}|_{x_0}\left(x_i-x_{0i}\right)+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 f}{\partial x_i\partial x_j}|_{x_0}(x_i-x_{0i})(x_j-x_{0j}) \quad (1)$$

Based on this foundation, to examine whether polynomial features capturing gene–gene interactions improve classification performance, I implement a second-order Taylor series–based SLE detection model. This approach simplifies the functional representation while retaining the ability to capture the nonlinear dynamics underlying SLE.

Specifically, for empirical implementation, the second-degree Taylor polynomial expansion centered at $x_0 = 0 = (0,\ 0,\ ...,\ 0)$ implied by equation (1) is given by

$$f(x)\approx f(0)+\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}|_{x_0}x_i+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 f}{\partial x_i\partial x_j}|_{x_0}x_ix_j \quad (2)$$

From equation (2), I define coefficients as $\beta_0 = f(0),\ \beta_i = \frac{\partial f}{\partial x_i}|_0,\ \beta_{ij} = \frac{1}{2}\frac{\partial^2 f}{\partial x_i\partial x_j}|_0$. With these definitions, the model can be expressed as

$$y = \beta_0+\sum_{i=1}^{n}\beta_i x_i+\sum_{i=1}^{n}\beta_{ii}x_i^2+\sum_{i<j}^{n}\beta_{ij}x_ix_j \quad (3)$$

where $\beta_i x_i$ represents the linear terms capturing the instantaneous change in disease risk (first derivative), $\beta_{ii}x_i^2$ reflects the acceleration or curvature of biomarker trajectories (second derivative), and $\beta_{ij}x_ix_j$ represents gene–gene or biomarker–biomarker interactions. Specifically, in this framework, linear terms represent individual gene effects, squared terms capture nonlinear amplification of gene expression, and interaction terms encode pairwise relationships between genes, which may reflect co-regulation or pathway-level activity. More specifically, in this framework, linear terms represent individual gene effects, squared terms capture nonlinear amplification of gene expression, and interaction terms encode pairwise relationships between genes, which may reflect co-regulation or pathway-level activity. In this study, Taylor-series expansion is operationalized as second-degree polynomial features, including linear terms, squared terms, and pairwise interactions, which allow the model to capture nonlinear and joint gene expression patterns.

Since SLE detection is binary (SLE vs non-SLE), the probability of SLE can be modeled as

$$P(x) = \frac{1}{1+exp(-y)} \quad (4)$$

where y is defined by equation (3). This represents a Taylor-expanded logistic regression, which is a transparent quadratic model rather than a black-box classifier.

Thus, this formulation justifies the inclusion of quadratic features, provides a natural interpretation of interaction terms through second derivatives, and establishes a direct connection to polynomial kernels, local approximations, and model interpretability. This formulation provides derivative-based features with clear biological interpretation while maintaining compatibility with regularized classification frameworks.

3.2.1. Gene Selection for Polynomial Expansion: To manage computational complexity and avoid feature explosion, I selected the top 200 genes from the full 49,207-gene set using Random Forest feature importance scores. This subset, representing 0.41% of all genes, retained the most informative signals while keeping the polynomial expansion computationally.

3.2.2. Polynomial Feature Generation: Using Scikit-learn's *PolynomialFeatures* transformer (degree = 2, include_bias = False), I generated second-order polynomial features from the 200 selected genes:

- **Linear terms:** 200 original gene expressions (g_1, g_2, ..., g_{200})
- **Squared terms:** 200 quadratic features (g_1^2, g_2^2, ..., g_{200}^2)
- **Interaction terms:** 19,900 pairwise products ($g_1 \times g_2$, $g_1 \times g_3$, ..., $g_{199} \times g_{200}$)
- **Total features:** 20,300

The polynomial transformer was fitted only on the training set and then applied to both training and test sets using the same learned parameters to prevent information leakage.

3.2.4. Feature Selection via L1 Regularization: Given the extreme dimensionality (20,300 features from 50 training samples), I employed Lasso (L1-regularized) logistic regression with cross-validated alpha selection (LassoCV, 5-fold CV, max_iter=10,000). This automated feature selection identified 53 features with non-zero coefficients (optimal $\alpha = 0.000574$), achieving 99.7% dimensionality reduction. The selected features comprised 18 linear terms (34.0%), 14 squared terms (26.4%), and 21 interaction terms (39.6%). The selected features included linear, squared, and interaction terms, demonstrating that both individual gene effects and gene–gene interactions contribute to classification.

3.2.5. Taylor-Enhanced Model Training: The three baseline algorithms were retrained using the 53 selected Taylor features. All hyperparameters and class-weight balancing were kept identical to ensure fair comparison. Models were evaluated using the same cross-validation procedures and the same held-out test set as in the baseline pipeline.

3.3. Baseline Model Development

Three machine learning algorithms representing diverse learning paradigms were implemented as baseline models:

1. Logistic Regression (LR): A linear classifier using L2 regularization (C = 1.0) with the *lbfgs* (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) solver and a maximum of 1,000 iterations. Class weights were set to balanced to account for the 1.77:1 class imbalance.
2. Random Forest (RF): An ensemble decision-tree classifier configured with 100 trees, a maximum depth of 10, minimum samples required for a split of 5, minimum samples per leaf of 2, and balanced class weights. Depth limits and minimum sample thresholds were included to reduce overfitting in the high-dimensional feature space.

3. Support Vector Machine (SVM): A kernel-based classifier using the radial basis function (RBF) kernel, balanced class weights, automatic gamma scaling, and C = 1.0. Probability estimates were enabled to support ROC curve analysis.

All models were trained on the 49,207-gene training set and evaluated using 5-fold stratified cross-validation. Final models were then trained on the full training set and evaluated once on the held-out test set. Performance metrics included accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curves with area under the curve (AUC).

3.4. Statistical Analysis

To compare the performance of baseline and Taylor-enhanced models, I conducted paired statistical tests using predictions from the held-out test set. Because each sample received two linked predictions—one from the baseline model and one from its Taylor counterpart—McNemar's test was used to assess whether the differences in classification outcomes were statistically significant. This test evaluates the 2×2 contingency table formed by discordant predictions, specifically the counts of samples for which the baseline model was correct but the Taylor model was incorrect, and vice versa. Using the continuity-corrected statistic,

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (5)$$

where b and c represent the discordant pair counts, statistical significance was evaluated at the α= 0.05 threshold.

To complement the hypothesis test and quantify the magnitude of observed accuracy changes, I calculated Cohen's h, an effect-size measure for paired proportions. Cohen's h was computed as:

$$h = 2(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}) \quad (6)$$

where p_1 and p_2 correspond to the accuracies of the Taylor-enhanced and baseline models, respectively. Effect sizes were interpreted as small ($h < 0.5$), medium ($0.5 \leq h < 0.8$), or large ($h \geq 0.8$), providing a more nuanced understanding of practical significance beyond p-values alone.

3.5. Computational Environment

All analyses were conducted in Python 3.10.12 within the Google Colaboratory environment. Key packages included NumPy 1.24.3, Pandas 2.0.3, Scikit-learn 1.3.0, Matplotlib 3.7.1, Seaborn 0.12.2, SciPy 1.11.1, and GEOparse 2.0.3. To ensure full reproducibility of results, a fixed random seed (42) was applied to all stochastic processes, including the train/test split, cross-validation folds, and model initialization.

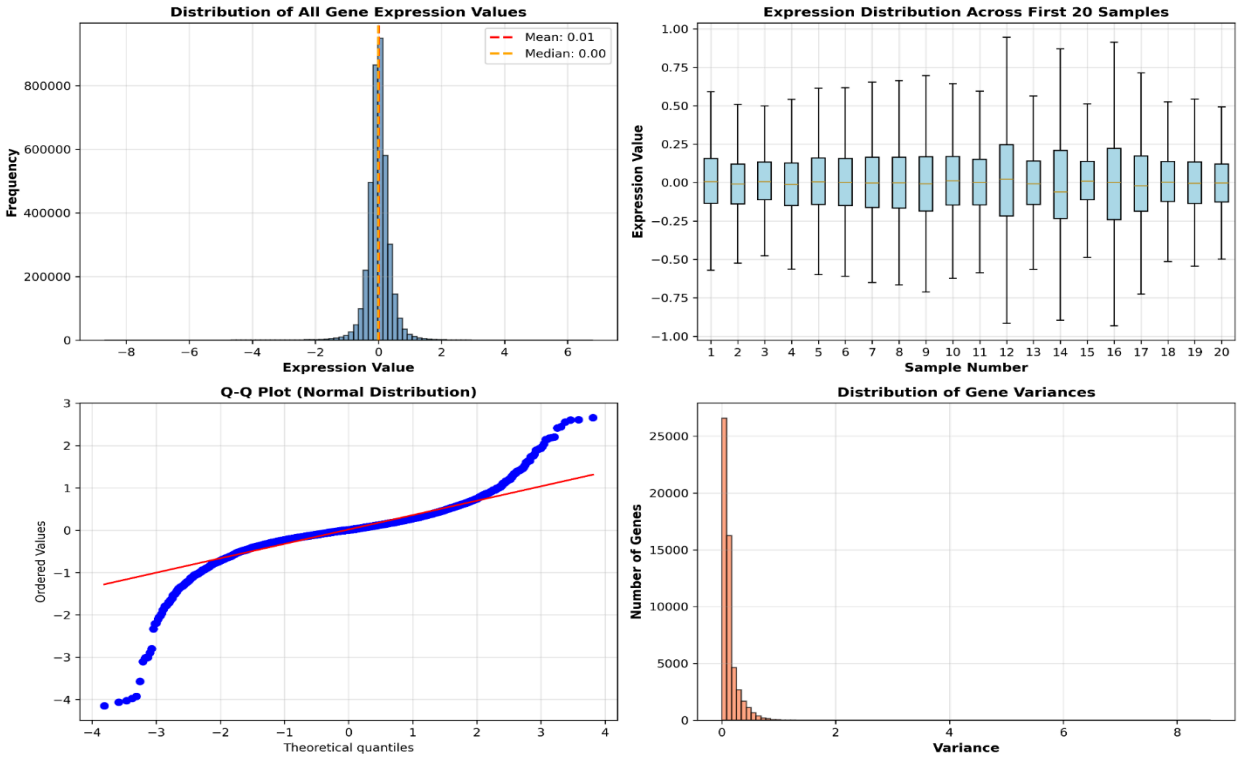

Figure 2: Distribution characteristics of preprocessed gene expression data.

Note. Clockwise from the top left: (A) Histogram of all 54,675 × 72 expression values, showing an approximately normal distribution centered near zero. (B) Boxplot of the first 20 samples, demonstrating consistent distribution shapes with minimal outliers. (C) Q–Q plot comparing observed expression values to a theoretical normal distribution, revealing substantial deviations in the distribution tails indicative of non-normal, heavy-tailed behavior typical of transcriptomic data. (D) Histogram of gene-wise variances, illustrating a strongly right-skewed distribution in which most genes exhibit low variability.

4. RESULTS

4.1. Dataset Characteristics and Preprocessing Outcomes

The final dataset consisted of 72 whole blood samples (46 healthy controls and 26 SLE patients) with 54,675 gene expression probes. After variance filtering, 49,207 genes (90%) were retained for baseline modeling. The dataset was split into 50 training samples (32 controls, 18 SLE) and 22 test samples (14 controls, 8 SLE), preserving the original class ratio (1.77:1). Gene expression values were $\log_2$-transformed and mean-centered (mean: 0.0065; median: 0.0000; range: −8.69 to 6.81), with no missing values. To demonstrate the integrity and readiness of the dataset prior to machine learning analysis, Figure 2 summarizes the overall distribution and variability of the preprocessed gene expression

values. The histogram of all expression values shows that the dataset is properly log-transformed and centered, while the sample boxplots indicate highly consistent distributions across individuals, with no major outliers or batch effects. The Q–Q plot reveals the expected heavy-tailed, non-normal behavior characteristic of transcriptomic data, and the variance distribution confirms that most genes exhibit low variability across samples, with only a small subset contributing meaningful biological signal. Together, these panels demonstrate that the data are well normalized, of high quality, and suitable for downstream machine learning analysis.

4.2. Baseline Model Performance

Baseline models demonstrated strong classification performance on the 22-sample test set in Table 1. Random Forest achieved perfect accuracy (100%, 22/22 correct), while Logistic Regression and Support Vector Machine (SVM) each achieved 90.9% accuracy (20/22 correct). Cross-validation on the training set showed similarly high performance, with an average accuracy of 98.0% ± 4.0%. Logistic Regression reached perfect cross-validation accuracy (100% ± 0%), suggesting the possibility of overfitting despite L2 regularization.

Across all models, the average baseline test accuracy was 93.9%. Precision, recall, and F1 scores ranged from 0.875 to 1.000, and ROC AUC values were consistently high (0.982–1.000), indicating strong discriminative ability.

Table 1. Baseline Model Performance on Test Set (n=22)

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Confusion Matrix (TN, FP, FN, TP)
Logistic Regression	0.9091	0.8750	0.8750	0.8750	0.9821	13, 1, 1, 7
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	14, 0, 0, 8
Support Vector Machine	0.9091	0.8750	0.8750	0.8750	0.9821	13, 1, 1, 7
Average	**0.9394**	**0.9167**	**0.9167**	**0.9167**	**0.9940**	—

Note. Denote true positives as TP, false positives as FP, true negatives as TN, and false negatives as FN. Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$, Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, and F1 Score = $2\times \frac{Precision \times Recall}{Precision+Recall}$. Accuracy measures the proportion of correct predictions out of the total number of cases processed. Precision measures how much of the positive predictions were actually correct. Recall measures show how many actual positive cases the model found. F1 Score measures the harmonic mean of precision and recall, indicating the effectiveness of the model and working especially well with imbalanced datasets.

Confusion matrices (Figure 3) provide insight into model-specific error patterns. Logistic Regression and SVM produced identical results, each correctly classifying 13 healthy and 7 SLE samples, with one false positive and one false negative. In contrast, Random Forest achieved perfect classification, correctly labeling all samples with no errors, corresponding to 100% sensitivity and specificity.

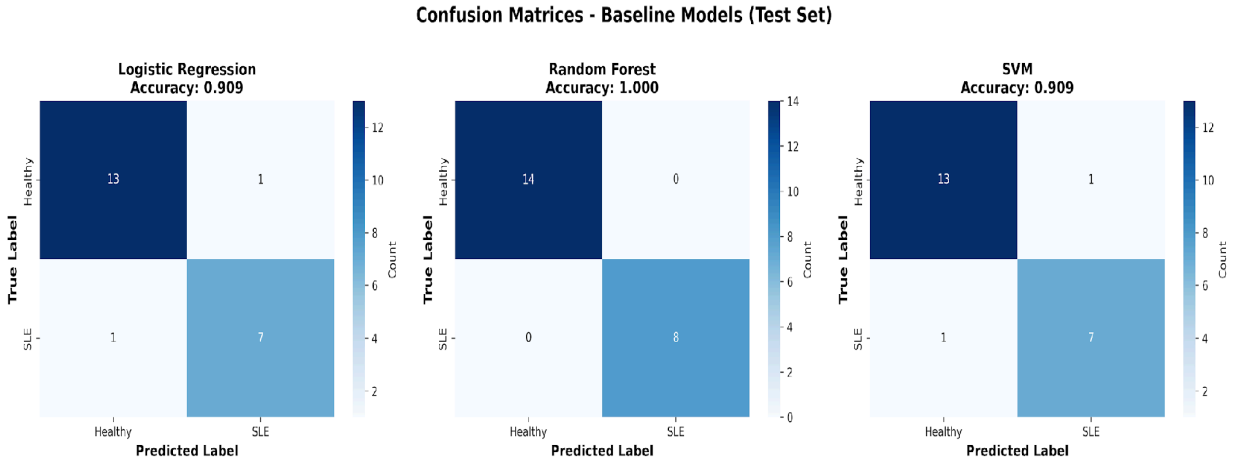

Figure 3: Confusion matrices for baseline models

ROC curves (Figure 4) further illustrate model performance across classification thresholds. Random Forest achieved an AUC of 1.000, indicating perfect separation, while Logistic Regression and SVM each achieved an AUC of 0.982, reflecting strong performance despite minor classification errors.

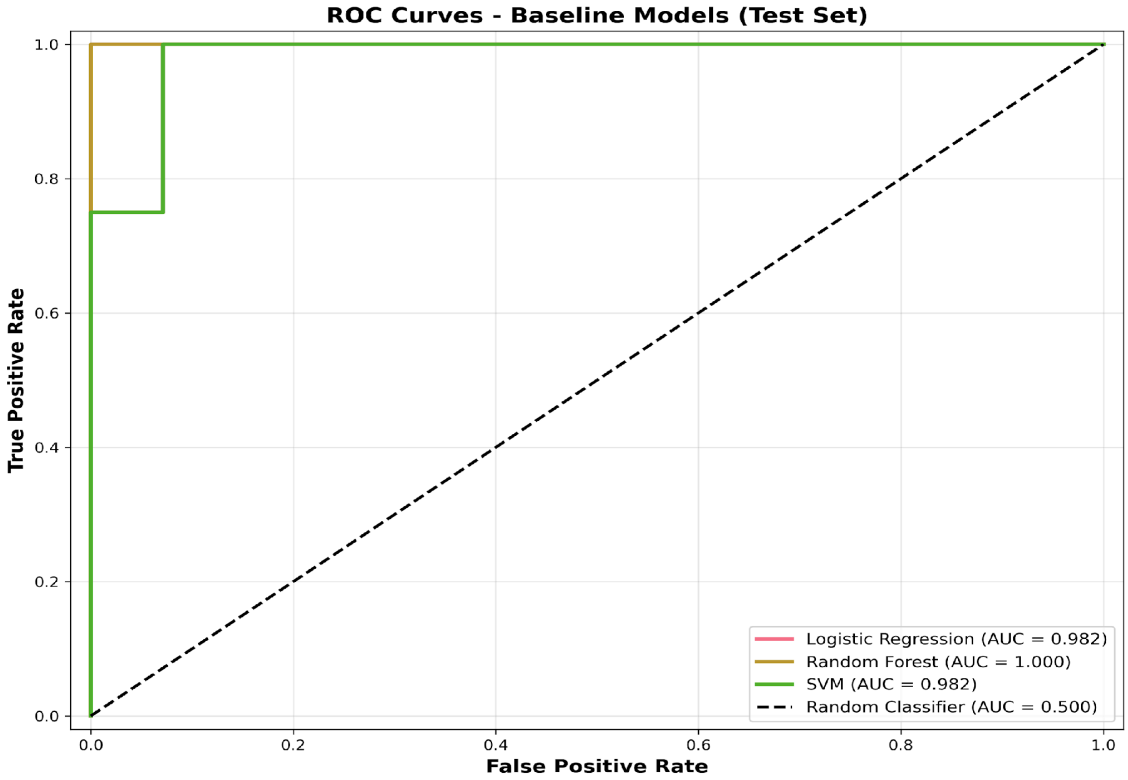

Figure 4: ROC curves for baseline models with area under curve (AUC) values.

Note. All three models demonstrate excellent discrimination (AUC ≥ 0.98).

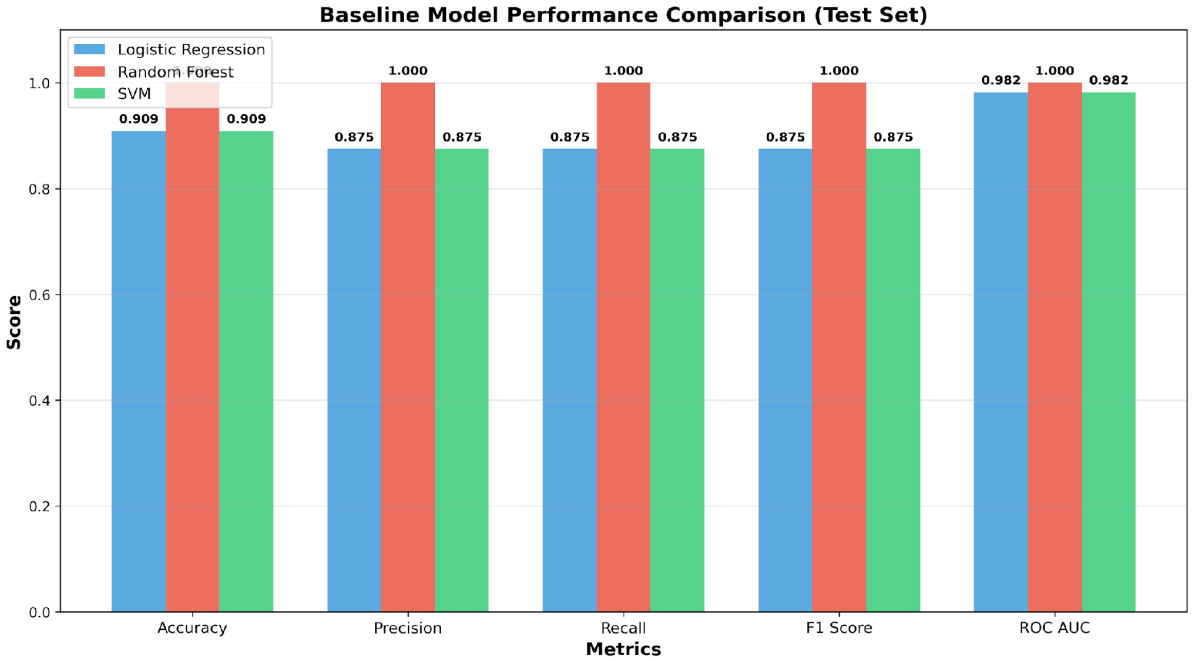

Figure 5: Baseline Model Performance on Test Set (total sample *(n)* =22).

Note. All three baseline machine learning models show promising results.

Figure 5 summarizes model performance across multiple evaluation metrics. Random Forest consistently outperformed the other models, achieving perfect scores across all metrics. Logistic Regression and SVM showed nearly identical performance profiles, with slightly lower accuracy and balanced precision–recall trade-offs.

4.3. Taylor-Series Feature Selection and Composition

These Taylor-derived features correspond to linear, squared, and interaction terms that capture both individual gene effects and joint expression patterns. From 20,300 generated polynomial features, Lasso selected 53 features using an optimal regularization parameter ($\alpha = 0.000574$), representing a substantial reduction in dimensionality. The selected features included 21 interaction terms (39.6%), 18 linear terms (34.0%), and 14 squared terms (26.4%), highlighting the important contribution of gene–gene interactions.

Examination of the top features ranked by coefficient magnitude in Table 2 indicates that both individual gene effects and interaction terms contribute to classification performance. Notably, interaction features appear frequently among the most influential predictors, suggesting that joint gene expression patterns provide additional discriminative information beyond single-gene signals.

The magnitude and direction of the coefficients further indicate that higher expression levels, or combinations of gene expression, are associated with increased probability of SLE. Overall, these results suggest that incorporating both linear and interaction components in the Taylor-series derived feature space enhances the model's ability to capture biologically meaningful patterns.

Table 2: Top ten selected Taylor-series features ranked by absolute LASSO coefficient values.

Feature	Gene Symbol	Gene Title	Lasso Coefficient
1560926_at	PPP4R2	Protein phosphatase 4, regulatory subunit 2	0.023660
1557910_at	HSP90AB1	Heat shock protein 90 alpha family class B member 1	0.020000
228986_at	OSBPL8	Oxysterol binding protein-like 8	0.018073
206025_s_at	TNFAIP6	Tumor necrosis factor alpha–induced protein 6	0.017188
212257_s_at	SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 2	0.013236
221861_at	ARRB1	Arrestin beta 1	0.011200

1552610_a_at	JAK1	Janus kinase 1	0.010000
211926_s_at	MYH9	Myosin heavy chain 9	0.010000
207574_s_at	GADD45B	Growth arrest and DNA damage–inducible beta	0.010000
222975_s_at	CSDE1	Cold shock domain containing E1	0.009787

4.4. Taylor-Enhanced Model Performance

As shown in Table 3, Taylor-enhanced models improved overall performance, increasing average test-set accuracy to 98.5%, a 4.5-percentage-point gain over baseline. Logistic Regression improved from 90.9% to 95.5%, correctly classifying 21 of 22 samples, while SVM showed the largest gain, reaching 100% accuracy from 90.9%. Random Forest remained unchanged at 100%, indicating that Taylor expansion did not affect its already optimal performance.

Cross-validation on the training set yielded perfect accuracy (100% ± 0%) for all Taylor-enhanced models. Despite the substantial reduction in feature space—from 20,300 polynomial features to 53 selected features, which uniformly perfect performance suggests that some degree of overfitting may remain.

Table 3: Taylor-Series Enhanced Model Performance on Test Set (*n*=22) and Comparison to Baseline

Model	**Baseline Accuracy**	**Taylor Accuracy**	**Absolute Gain**	**Relative Gain**	**P-Value (McNemar)**	**Significant?**
Logistic Regression	0.9091	0.9545	+0.0455	+5.0%	1.0000	No
Random Forest	1.0000	1.0000	0.0000	0.0%	1.0000	No
Support Vector Machine	0.9091	1.0000	+0.0909	+10.0%	0.4795	No
Average	**0.9394**	**0.9848**	**+0.0455**	**+4.8%**	—	—

Confusion matrices in Figure 6 further illustrate these results. Logistic Regression misclassified one SLE case, whereas both Random Forest and SVM correctly classified all 22 test samples. The improvement is most evident for SVM, where previously misclassified cases were fully corrected. Overall, the matrices show fewer classification errors, particularly reduced false negatives.

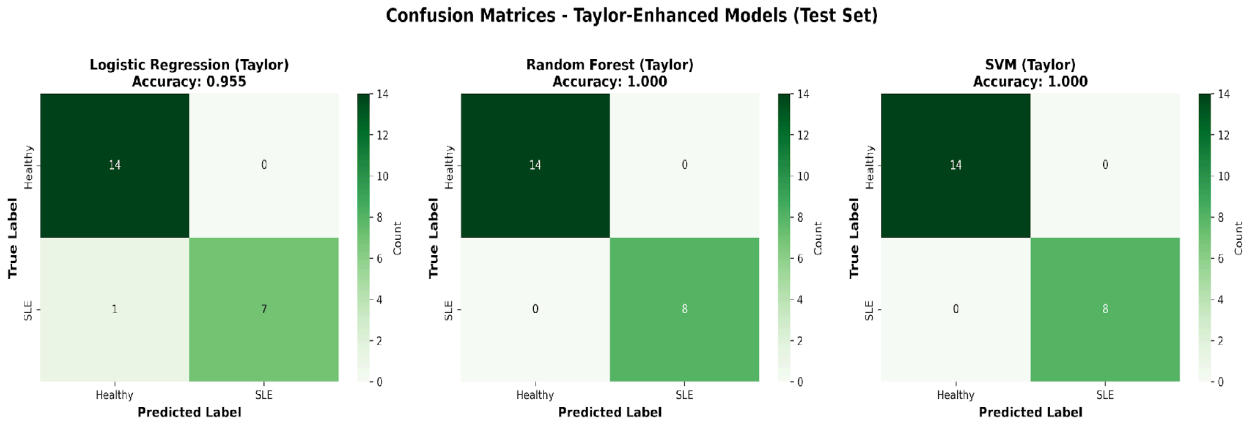

Figure 6: Confusion matrices for the Taylor-enhanced models

Note. This figure illustrates improved classification performance across all classifiers. Notably, the Taylor-enhanced SVM achieved perfect classification on the test set, correctly identifying all 22 of 22 samples.

ROC curve comparisons in Figure 7 indicate that Taylor expansion maintained or improved discrimination. Logistic Regression showed a clear increase in AUC (from 0.963 to 1.000), while Random Forest and SVM achieved perfect separation (AUC = 1.000) under both baseline and enhanced conditions. The steep ROC curves reflect very low false-positive rates and strong overall model performance.

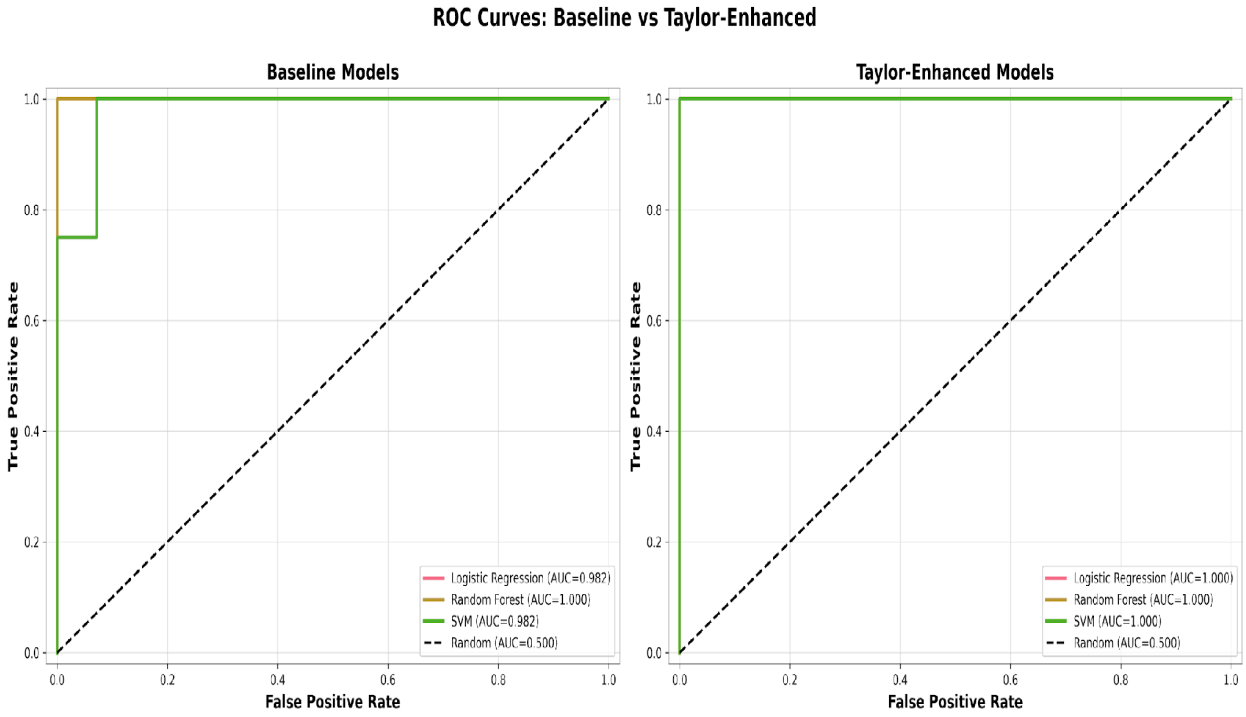

Figure 7: Side-by-side comparison of the baseline and Taylor-enhanced model's ROC curves for all three models

Note. This figure illustrates that Taylor expansion consistently maintains or improves classification discrimination.

4.5. Statistical Significance Testing

Despite numerical improvements in two of the three models, McNemar's test did not identify statistically significant differences between baseline and Taylor-enhanced performance (all $p > 0.05$; Table 2). The p-values for Logistic Regression ($p = 1.000$) and SVM ($p = 0.480$) reflect the limited statistical power of the small test set ($n = 22$), where only one or two samples changed prediction outcomes between models. Under these conditions, even noticeable accuracy gains are unlikely to reach statistical significance.

Effect size analysis using Cohen's h indicated small-to-moderate effects: approximately 0.30 for Logistic Regression, 0.60 for SVM, and 0.00 for Random Forest, with an average effect size of ~0.30. These values suggest that the improvements—particularly for SVM—may be practically meaningful, even if not statistically significant given the sample size.

Figure 8 provides a two-panel summary of these results. The left panel illustrates the magnitude of accuracy improvements (Taylor − Baseline) across models. Logistic Regression and SVM show clear positive gains (+0.045 and +0.091, respectively), while Random Forest shows no change, consistent with identical predictions in both conditions. However, these results indicate that, while numerical

improvements are present, the study lacks sufficient statistical power to confirm that these differences represent true performance gains rather than chance variation.

The right panel displays the p-values from McNemar's tests for each model, all of which remain well above the significance threshold ($\alpha = 0.05$). The consistently high p-values, despite measurable accuracy gains, highlight the limited sensitivity of statistical testing in small paired samples. Together, these panels illustrate the contrast between numerical improvement and statistical significance, emphasizing that the observed performance gains should be interpreted cautiously.

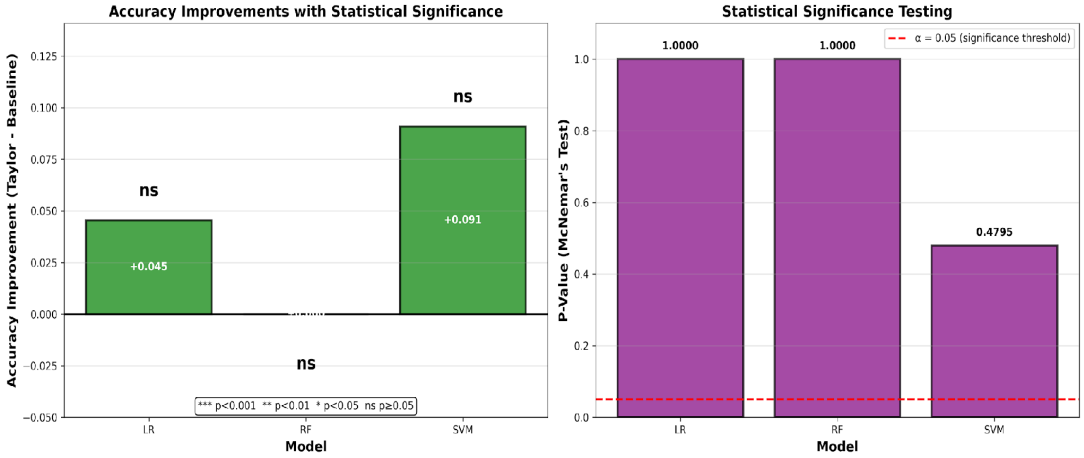

Figure 8: Statistical significance analysis.

Note. From left: (A) Bar chart of accuracy improvements with significance markers (none significant at α=0.05). (B) P-value distribution from McNemar's tests showing high p-values reflecting limited statistical power.

4.6. Sample-Level Prediction Analysis

Examination of individual prediction outcomes revealed which samples were affected by the introduction of Taylor-series derived features. For Logistic Regression, one SLE sample that was misclassified by the baseline model was correctly reclassified by the Taylor-enhanced model, resulting in a single recovered true positive. For SVM, two samples changed prediction status: both were SLE patients erroneously labeled as healthy by the baseline model but correctly identified as SLE after Taylor enhancement. This improvement aligns with SVM's overall gain in test accuracy. In contrast, Random Forest produced identical predictions across all 22 samples, consistent with its unchanged accuracy.

These sample-level changes suggest that Taylor-series derived features primarily improved sensitivity, the model's ability to correctly identify SLE cases, rather than specificity. This distinction may be clinically relevant, as false negatives (missed lupus diagnoses) carry greater risk in screening and early-detection contexts than false positives.

4.7. Biological Interpretation of Selected Features

The 21 gene–gene interaction features selected by LASSO represent pairwise products of gene expression values, capturing potential co-regulation and pathway-level relationships. Their substantial proportion (39.6% of selected features) indicates that modeling joint gene-expression patterns contributes meaningfully to lupus classification, beyond individual gene effects.

This finding aligns with the known biology of SLE, which involves coordinated dysregulation across immune pathways such as interferon signaling, B-cell activation, and inflammatory responses. In this context, interaction-based features may better reflect system-level disease mechanisms than single-gene markers.

The greater number of interaction terms compared to squared terms (21 vs. 14) further suggests that relationships between genes provide more informative structure than nonlinear transformations of individual gene expression.

4.8. Model Comparison and Overall Assessment

Figure 9 compares baseline and Taylor-enhanced models across five performance metrics, including accuracy, precision, recall, F1 score, and ROC AUC. Overall, Taylor expansion either improved or maintained performance across all models, with no evidence of degradation.

In terms of accuracy, Logistic Regression and SVM showed measurable gains, while Random Forest remained at 100%. Precision and recall were consistently high across all models, with SVM achieving perfect recall after Taylor enhancement. Logistic Regression also showed modest improvements in both metrics, indicating better detection of SLE cases. F1 scores reflected these trends, demonstrating equal or improved balance between precision and recall in the enhanced models.

ROC AUC values further confirmed strong performance: Logistic Regression improved to AUC = 1.000, while SVM and Random Forest maintained their already perfect discrimination. On average, Taylor-enhanced models achieved a 4.55-percentage-point increase in accuracy, with SVM benefiting the most from the expanded feature space.

These results should be interpreted with caution. Perfect cross-validation performance (100% ± 0%) suggests potential overfitting, and the small test set (n = 22) limits statistical power. In addition, the high feature-to-sample ratio and reliance on a single dataset constrain generalizability. These findings, particularly the near-perfect accuracy in multiple models, also raise the possibility of overfitting given the small sample size.
Despite these limitations, the consistent improvement observed in Logistic Regression and SVM, together with the importance of gene–gene interaction features, suggests that Taylor-series derived features capture meaningful patterns not fully represented by baseline models.

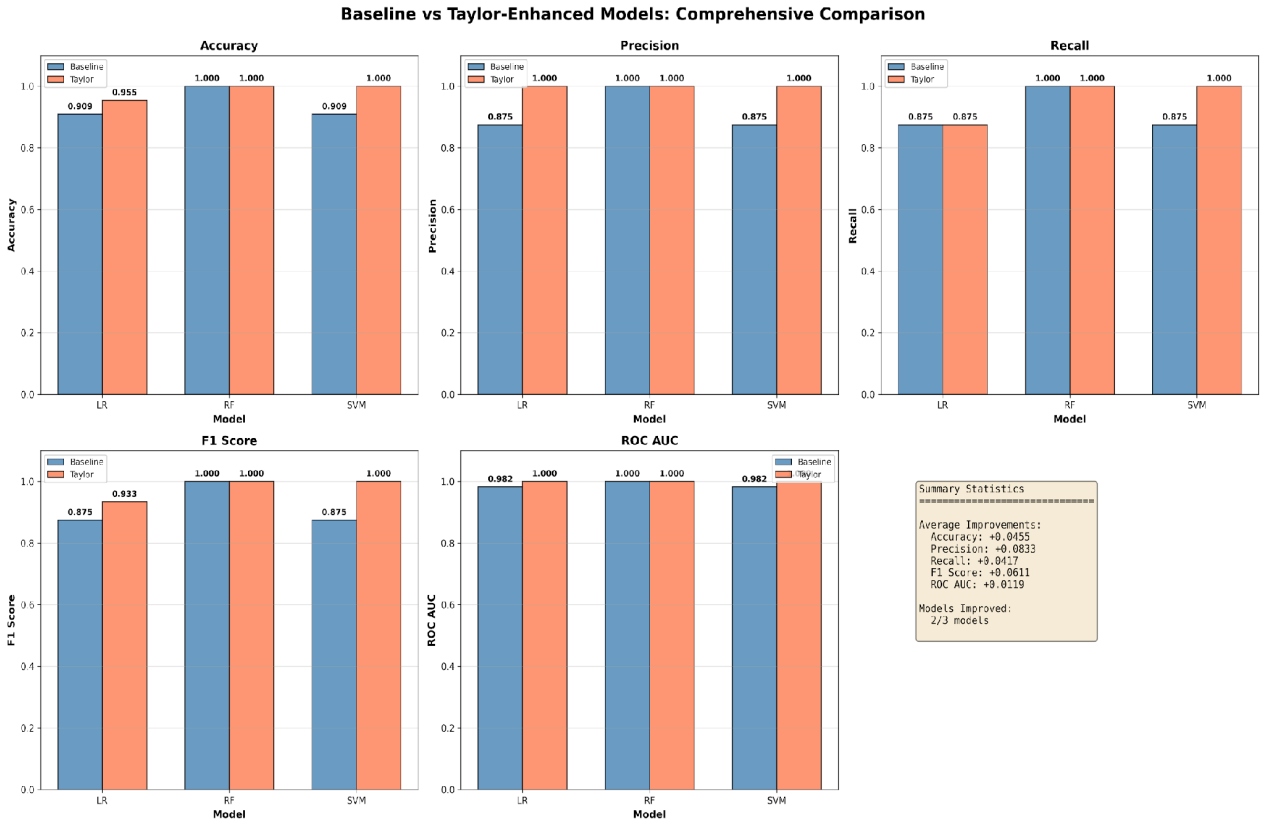

Figure 9: Comprehensive Performance Comparison: Baseline vs. Taylor-Enhanced Models

Note. A comprehensive five-panel comparison summarizes all major performance metrics (accuracy, precision, recall, F1 score, and ROC AUC) for baseline versus Taylor-enhanced models across the three algorithms. The Taylor-enhanced models consistently demonstrated either improved or equivalent performance on every metric, with no evidence of degradation relative to their baseline counterparts.

5. DISCUSSION

5.1. Principal Findings

This study examined whether Taylor-series–derived features improve machine learning classification of systemic lupus erythematosus (SLE) from gene expression data. Three main findings emerged. First, Taylor-enhanced models achieved higher average test accuracy (98.5%) compared to baseline models (93.9%), with the largest improvement observed for SVM. Second, LASSO selected a substantial proportion of gene–gene interaction features (21 of 53), suggesting that joint expression patterns contribute additional diagnostic value beyond individual gene levels. Third, although improvements were observed for Logistic Regression and SVM, the primary hypothesis was not supported by statistically

significant results. The absence of significance likely reflects limited statistical power due to the small test set (n = 22), meaning that observed performance differences cannot be reliably distinguished from chance.

5.2. Interpretation and Biological Significance

The prominence of interaction features is consistent with the current understanding of SLE as a network-level disease driven by coordinated dysregulation across immune pathways, including interferon signaling, B-cell activation, and cytokine networks. By explicitly incorporating gene–gene interactions, the Taylor expansion captures relationships that are not fully represented by linear models based on individual gene expression. In this study, these features are implemented as polynomial expansions that explicitly encode nonlinear and interaction effects in gene expression.

While the overall gains in accuracy are modest, improvements in sensitivity, particularly the correction of previously misclassified SLE cases, may be clinically relevant, but require validation in larger cohorts. However, because these improvements were not statistically significant, they should be interpreted with caution, as they may reflect sampling variability rather than consistent underlying effects. In diagnostic settings, false negatives carry greater risk than false positives, and even small improvements in detecting true disease cases may be valuable.

From a methodological perspective, the Taylor-series approach offers a structured way to introduce nonlinear relationships into machine learning models. Unlike kernel or tree-based methods that capture such effects implicitly, this approach makes interactions explicit and interpretable, while L1 regularization ensures that only a sparse and informative subset of features is retained.

5.3. Limitations

A key limitation of this study is the substantial risk of overfitting, driven by the combination of small sample size and high-dimensional feature space. Although feature selection reduced the Taylor-series derived feature set from 20,300 to 53 predictors, the initial feature construction process and the high flexibility of the models increase the likelihood that patterns specific to the training data were captured rather than generalizable signal. This risk is further amplified by the extreme feature-to-sample ratio (49,207 baseline features and 50 training samples), which is well known to produce unstable model estimates in genomic analyses.

Evidence of overfitting is reflected in the near-perfect performance observed across models, including 100% cross-validation accuracy and perfect test-set classification for both Random Forest and SVM. While these results may indicate strong signal, such high performance is uncommon in small biomedical datasets and raises concern that the models may be fitting noise or dataset-specific artifacts. The lack of statistically significant differences despite noticeable accuracy gains further suggests that the observed improvements may not be robust.

As a result, model performance should be interpreted cautiously. The apparent gains from Taylor-series derived feature engineering may partially reflect overfitting rather than true generalization improvements. Validation on larger, independent datasets is therefore essential to determine whether these findings represent reproducible biological signal or are specific to this cohort.

5.4. Methodological Strengths

Several aspects of the experimental design strengthen the validity and reproducibility of these findings. First, strict separation of training and test data was maintained by performing data splitting prior to preprocessing and feature selection, preventing information leakage and ensuring unbiased evaluation. Second, the use of multiple learning algorithms (Logistic Regression, Random Forest, and SVM) demonstrates that the observed improvements are not model-specific, with consistent gains observed in two of the three models.

Feature selection was performed using LASSO with cross-validated regularization, enabling automated identification of a sparse and informative subset of predictors while avoiding subjective feature choice. In addition, reproducibility was supported through fixed random seeds, explicit parameter settings, and clearly defined computational procedures. Finally, class imbalance was addressed through appropriate weighting, ensuring that model performance reflected meaningful detection of SLE cases rather than bias toward the majority class.

5.5. Implications for Clinical Practice and Research

These findings should be considered exploratory, as the study was not sufficiently powered to statistically confirm the observed performance improvements. While exploratory, these findings suggest several potential research directions for improving molecular characterization of SLE. First, gene expression–based screening may help identify patterns associated with early disease risk during the preclinical phase, when standard diagnostic criteria are not yet met, enabling earlier intervention. Second, molecular signatures derived from gene–gene interactions may support future approaches to differential diagnosis by distinguishing SLE from overlapping autoimmune conditions, potentially reducing diagnostic delays.

In addition, this approach may provide a basis for disease stratification in future studies by capturing pathway-level variation, helping to classify patients into biologically distinct subtypes associated with different clinical manifestations. It may offer preliminary insight into treatment-response patterns by identifying interaction patterns linked to therapeutic response, thereby supporting more targeted and efficient treatment decisions.

Beyond lupus, the Taylor-series derived feature engineering framework provides a generalizable approach for modeling gene interactions in other diseases characterized by complex regulatory networks, including autoimmune disorders, cancer, and neurodegenerative diseases.

More broadly, these findings illustrate that meaningful improvements in biomedical machine learning can arise from how data are represented, rather than solely from more complex models or larger datasets. By incorporating mathematically grounded feature transformations, this approach highlights the value of combining classical methods with modern genomic analysis to extract biologically relevant signals.

5.6. Future Directions and Broader Impact

With further validation in larger, independent cohorts, Taylor‑enhanced models may contribute to more sensitive molecular screening, pending validation in larger, independent cohorts for SLE, improving early detection and differential diagnosis in patients with ambiguous symptoms. Future work should focus on testing this approach across external datasets, mapping interaction features to biological pathways, and developing reduced gene panels suitable for clinical platforms such as qRT‑PCR or targeted RNA‑seq. Extending the framework to longitudinal data, treatment-response prediction, and other autoimmune diseases may further demonstrate its broader applicability.

Beyond these immediate extensions, this work highlights a more general principle: meaningful improvements in biomedical machine learning can arise not only from larger datasets or more complex models, but also from rethinking how existing data are represented. By introducing mathematically grounded transformations, such as Taylor-series expansion, it becomes possible to capture nonlinear and interaction-based structure that better reflects underlying biological systems.

This approach emphasizes interpretability as well as performance. Explicitly modeling gene–gene interactions enables clearer connections to known biological pathways, moving beyond “black box” predictions toward mechanistically meaningful insights. Such transparency is particularly important for clinical adoption, where interpretability and biological plausibility are essential.

Finally, the study underscores the importance of rigorous and reproducible methodology in biomedical machine learning. Careful data handling, use of multiple algorithms, and transparent reporting support more reliable findings and contribute to addressing broader concerns regarding reproducibility in computational research. Together, these elements suggest that mathematically informed feature engineering can play a valuable role in advancing both the performance and interpretability of genomic models.

CONCLUSION

This study provides proof-of-concept evidence that Taylor-series–derived features, particularly gene–gene interaction terms, can enhance machine-learning classification of systemic lupus erythematosus (SLE) from gene-expression data. Across three model families, Taylor-enhanced models achieved an average test accuracy of 98.5% versus 93.9% for baselines, and L1 (Lasso) regularization preferentially retained interaction features (~40% of the 53 selected terms). Although McNemar's tests were not significant on a small held-out set (n=22), the consistent direction of improvement and the biological plausibility of the retained interactions suggest the presence of a meaningful signal, though confirmation requires larger studies.

The prevalence of interaction terms aligns with SLE as a network disease, in which coordinated pathway activity, not single-gene effects, drives phenotype. This supports the idea that a calculus-grounded representation capturing local slopes, curvature, and pairwise dependencies can surface structure that static features miss, suggesting potential predictive gains alongside improved interpretability.

Nevertheless, important caveats remain. Perfect cross-validation accuracy and a high feature-to-sample ratio highlight the potential for overfitting, and therefore external validation on independent cohorts ($\geq$ 300–500 samples) is essential to establish generalizability. If validated in larger, independent cohorts, Taylor-enhanced models may support future efforts toward earlier diagnosis, molecular subtype stratification, and treatment-response prediction, improving outcomes for people with SLE.

Beyond lupus, this work provides a generalizable framework for embedding dynamic and interaction structure into standard ML pipelines for diseases characterized by dysregulated molecular networks. As precision medicine advances toward molecular-based diagnosis and therapy selection, methods that explicitly encode biological complexity, while remaining transparent and compatible with conventional classifiers, will be increasingly valuable.

Finally, these results directly address the central question: derivative-based Taylor features improved classification relative to raw features, with the largest gain observed in SVM (+9.1 percentage points). While statistical significance was limited by sample size, effect sizes and sample-level reclassifications indicate practically meaningful improvements, particularly fewer missed SLE cases. Taken together, the findings, to some extent, support the alternative hypothesis and motivate multi-cohort validation, pathway-level interpretation, and eventual reduction to a minimal, clinically actionable gene panel. However, the high performance observed in this study may partly reflect overfitting, and requires confirmation in larger, independent datasets.

ACKNOWLEDGMENTS

This research was conducted as an independent project at W. T. Clarke High School, Westbury, New York. I thank my science research teacher, Mitchel Klass, and my family for their unwavering support throughout this project. I acknowledge the Gene Expression Omnibus (GEO) repository maintained by the National Center for Biotechnology Information (NCBI) for providing open-access gene expression data (GSE39088). I am grateful to the original researchers who generated and shared the GSE39088 dataset, making secondary analyses like this study possible. All computational analyses were performed using free, open-source Python libraries developed by the scientific computing community, and all code was run on the Google Colaboratory environment. No external funding was received for this research. All figures and tables were created by the student author.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

Ahlquist, K. D., Sugden, L. A., & Ramachandran, S. (2023). Enabling interpretable machine learning for biological data with reliability scores. *PLoS Computational Biology*, *19*(5), e1011175. https://doi.org/10.1371/journal.pcbi.1011175

Aliferis, C., & Simon, G. (2024). Overfitting, underfitting, and general model overconfidence and underperformance: Pitfalls and best practices in machine learning and AI. In G. J. Simon & C. Aliferis (Eds.), *Artificial intelligence and machine learning in health care and medical sciences* (pp. 477–524). Springer. https://doi.org/10.1007/978-3-031-39355-6_10

Alnaimat, F., Hamdan, O., Othman, L., Dabbah, T., Marar, M., & Mohammed, R. H. A. (2025). Patient perspectives on reproductive health in systemic lupus erythematosus: Exploring disease manifestations, quality of life, and the role of social support. *International Journal of Women's Health, 17*, 1849–1862. https://doi.org/10.2147/IJWH.S519395

Birmingham, D. J., Irshaid, F., Nagaraja, H. N., Zou, X., Tsao, B. P., Wu, H., Yu, C. Y., Hebert, L. A., & Rovin, B. H. (2010). The complex nature of serum C3 and C4 as biomarkers of lupus renal flare. *Lupus, 19*(11), 1272–1280. https://doi.org/10.1177/0961203310371154

Gu, Z., Ge, B., Wang, Y., Gong, Y., & Qi, M. (2025). Artificial intelligence technologies for enhancing neurofunctionalities: A comprehensive review with applications in Alzheimer's disease research. Frontiers in Aging *Neuroscience, 17*, 1609063. https://doi.org/10.3389/fnagi.2025.1609063

Jiang, J. (2022). *Asymptotic expansions.* In *Large sample techniques for statistics*. Springer. https://doi.org/10.1007/978-3-030-91695-4_4

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kuehn, C., Zschaler, G., & Gross, T. (2015). Early warning signs for saddle-escape transitions in complex networks. *Scientific Reports, 5*, 13190. https://doi.org/10.1038/srep13190

Kumar, M., Yip, L., Wang, F., Marty, S. E., & Fathman, C. G. (2025). Autoimmune disease: Genetic susceptibility, environmental triggers, and immune dysregulation. *Frontiers in Immunology, 16*, 1626082. https://doi.org/10.3389/fimmu.2025.1626082

Lazar, S., & Kahlenberg, J. M. (2023). Systemic lupus erythematosus: New diagnostic and therapeutic approaches. *Annual Review of Medicine, 74*, 339–352. https://doi.org/10.1146/annurev-med-043021-032611

Leventhal, E. L., Daamen, A. R., Grammer, A. C., & Lipsky, P. E. (2023). An interpretable machine learning pipeline based on transcriptomics predicts phenotypes of lupus patients. *iScience, 26*(10), 108042. https://doi.org/10.1016/j.isci.2023.108042

Liu, X., Lu, X., Wang, L., Du, Q., Li, X., Li, Y., Ding, X., Lai, Y., & Chen, X. (2025). Bioinformatics analysis of immune infiltration and cell cycle hub genes in atopic dermatitis. *Biochemistry and Biophysics Reports, 44*, 102253. https://doi.org/10.1016/j.bbrep.2025.102253

Lupus Foundation of America. (2025, April 11). Lupus facts and statistics. https://www.lupus.org/resources/lupus-facts-and-statistics

Maddula, S. (2024, August 8). *Taylor series in AI*. Towards AI. https://towardsai.net/p/artificial-intelligence/taylor-series-in-ai

Mau, J., & Moon, K. (2024). *Incorporating Taylor series and recursive structure in neural networks for time series prediction*. arXiv. https://arxiv.org/abs/2402.06441

Mitchell, J. L. (2024). Understanding the impact of delayed diagnosis and misdiagnosis of systemic lupus erythematosus (SLE). *Journal of Family Medicine and Primary Care, 13*(11), 4819–4823. https://doi.org/10.4103/jfmpc.jfmpc_1177_24

National Institute of Arthritis and Musculoskeletal and Skin Diseases. (2022, October). *Systemic lupus erythematosus (lupus)*. https://www.niams.nih.gov/health-topics/lupus

NHS South Tees Hospitals. (2022, April 27). *Complement C3 and C4*. https://www.southtees.nhs.uk/services/pathology/tests/complement-c3-and-c4/

Parodis, I., Fanouriakis, A., Bortoluzzi, A., Iking-Konert, C., Weinmann-Menke, J., San Román, C., Levy, R. A., Rua-Figueroa, I., Alexander, T., & Amoura, Z. (2025). Practical insights for the clinical implementation of the EULAR recommendations for patients with systemic lupus erythematosus. *RMD Open, 11*(4), e006210. https://doi.org/10.1136/rmdopen-2025-006210

Wang, J. J. (2024). Biostatistical challenges in high-dimensional data analysis: Strategies and innovations. *Computational Molecular Biology, 14*(4), 163–172. https://doi.org/10.5376/cmb.2024.14.0019

Zhou, Y., Wu, Y., Yang, B., Zhu, Q., Long, H., Yuan, L., Cao, W., & Deng, D. (2025). Single-cell and transcriptomic analysis of regulatory mechanisms of key genes in systemic lupus erythematosus. *International Journal of General Medicine, 18*, 3193–3206. https://doi.org/10.2147/IJGM.S522871