SAGE: A Unified Evaluation Framework for Data Augmentation and Few-Shot Learning on Small and Imbalanced Tabular Datasets

Yuhao Yan[1*], Linlu Chen[2], Houyan Zhang[1], Chong Chen[1], Leran Liang[1], Meng Yang[1]
[1]School of Medical Informatics Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, China
[2]School of Public Health and Management, Guangzhou University of Chinese Medicine, Guangzhou 510006, China
E-mail: vogthepburn08904@outlook.com
[*]Corresponding author

Keywords: few-shot learning, class imbalance, algorithm evaluation framework, model selection, sage framework

Received: October 22, 2025

In critical domains such as healthcare and finance, structured data often suffers from sample scarcity and class imbalance, undermining the traditional machine learning assumption that training data adequately reflects the true distribution. To address this challenge, this study proposes SAGE (Small-sample Adaptive Generalization Evaluation), a unified framework for systematically comparing data-driven augmentation methods with model-driven few-shot learning (FSL) approaches. The framework integrates a standardized data conditioning pipeline, a comprehensive spectrum of 12 models (including 6 classical classifiers and 6 FSL architectures), and multi-dimensional evaluation metrics. Experimental validation was conducted on three diverse datasets: UCI Heart Disease (297 samples), Hepatitis (155 samples), and Glass Identification (214 samples), covering medical and forensic domains. Results demonstrate the complementary strengths of both paradigms. For data-driven methods, CatBoost augmented with Large Language Models (LLMs) achieved a Macro-F1 of 0.4219 on the heart disease dataset, significantly outperforming traditional oversampling methods like SMOTE ($p<0.001$). However, for extreme scarcity, model-driven approaches proved superior; Siamese Networks achieved the highest Macro-F1 of 0.5959 on heart disease and maintained robustness across datasets, specifically attaining an F1-score of 0.64 on the rarest class (Class 4). Furthermore, SHAP analysis confirmed that the best-performing models successfully captured clinically relevant features, such as albumin levels in hepatitis prediction. The SAGE framework thus provides empirical evidence to guide paradigm selection: FSL for extreme scarcity and LLM-based augmentation for enhancing ensemble classifiers.

Povzetek: Študija predstavlja okvir SAGE za analizo majhnih in neuravnoteženih strukturiranih podatkov. Rezultati kažejo, da je FSL primernejši pri ekstremnem pomanjkanju vzorcev, medtem ko LLM-povečanje izboljša klasične modele.

1 Introduction

In high-stakes industrial process monitoring, adaptive control systems, financial risk, and medical diagnosis, data scarcity and class imbalance are pervasive. Critical minority classes—such as diseased medical samples [1] or financial fraud [2]—are often underrepresented, violating standard algorithmic assumptions and leading to biased generalization [3]. This challenge persists even in large-scale datasets [4-5], necessitating robust models.

Two paradigms address this [6]: data-centric and model-centric. Data-centric approaches rebalance datasets using synthetic generation, ranging from classical SMOTE [7] and ADASYN [8-9] to CTGAN [10] and LLMs [11-12], though they risk introducing noise and lowering interpretability [13]. Model-centric approaches, led by Few-Shot Learning (FSL) [3], utilize architectures like MAML [14] or Siamese and Matching Networks [15] to learn from sparse data. FSL preserves data integrity but can falter under distribution divergence [16]. Crucially, systematic cross-paradigm evaluation remains limited [17], as most research is fragmented into single-paradigm comparisons, hindering evidence-based decision-making.

To bridge this gap, we propose SAGE (Small-sample Adaptive Generalization Evaluation), a unified framework for rigorous comparative analysis of imbalanced tabular data. SAGE integrates a standardized conditioning pipeline (with robust imputation), diverse algorithms (SMOTE, LLMs, MAML, Reptile [18], Siamese Networks), and a diagnostic layer for reproducible benchmarking. Validation across Hepatitis, Glass Identification, and UCI heart disease [19] datasets reveals complementary paradigm strengths: FSL models (e.g., Siamese Networks) offer superior stability, while LLM augmentation significantly boosts conventional classifiers. SHAP-based interpretability further confirms these gains are grounded in meaningful feature attributions.

To guide our investigation, we pose the following three Research Questions (RQs):

RQ1: Can advanced data augmentation techniques, specifically LLMs, effectively enable traditional classifiers to outperform few-shot learning models on small, imbalanced tabular data?

RQ2: Do model-centric few-shot learning approaches offer superior robustness and stability across minority classes compared to data-centric approaches?

RQ3: How do different model families interact with augmentation strategies, and does model complexity correlate with the risk of overfitting to synthetic noise?

2 Related works

2.1 Data-centric approaches for imbalanced data

Imbalance is traditionally addressed via resampling. SMOTE [7] uses linear interpolation, and ADASYN [8] adaptively shifts boundaries, though both risk noise in high-dimensional spaces [9]. CTGAN [10] models column-specific distributions but faces mode collapse on small data. LLMs [11-12] offer semantic generative potential, yet medical tabular applications remain underexplored.

2.2 Model-centric few-shot learning

FSL utilizes architectural bias. Meta-learning like MAML [13] optimizes task adaptation, while metric-learning (e.g., Siamese [14], Matching Networks [15]) clusters similar samples in embedding spaces. However, FSL lacks systematic benchmarking against data-centric methods on imbalanced tabular data.

2.3 Comparison with SAGE

Previous research often isolates individual paradigms on single datasets without stability testing. SAGE bridges this gap by unifying both under a standardized evaluation framework (Table 1).

Table 1: Comparison of SAGE with Existing Frameworks.

Feature	Traditional Augmentation [7, 8]	GAN-based Methods [10]	FSL Studies [14, 15]	SAGE (Ours)
Core Technique	Interpolation (SMOTE)	Generative (GAN)	Meta/Metric Learning	Unified (LLM + FSL)
Paradigm	Data-centric only	Data-centric only	Model-centric only	Dual-Path (Data & Model)
Dataset Scope	Single Domain	Single Domain	Image/Text mostly	Multi-Domain (Medical & Ind.)
Evaluation	Accuracy/F1 (Single run)	Accuracy/F1	N-way K-shot Accuracy	Robustness (Mean± Std)
Imputation	Often ignored	Internal handling	Often ignored	Robust Median/Mode Imputation

3 Methodology of SAGE framework

3.1 Framework overview

SAGE addresses the lack of standardized evaluation protocols in small-sample, imbalanced modelling—a fragmentation that undermines result comparability and practitioner guidance. By unifying data preparation, model implementation, and performance evaluation, SAGE ensures observed performance reflects intrinsic model capabilities (Figure 1).

The framework consists of three layers:

(1) Data Conditioning Layer: This layer standardizes preprocessing across four areas: (i) a unified feature engineering pipeline; (ii) stratified sampling to preserve rare classes; (iii) leakage-free imputation using training-set statistics (addressing incompleteness as seen in the Hepatitis dataset); and (iv) a standardized interface for augmentation (e.g., SMOTE, CTGAN, LLM). These are applied exclusively to training sets to prevent data leakage.

(2) Model Spectrum Layer: This layer facilitates fair cross-paradigm comparisons. The data-centric path evaluates classical classifiers on augmented data, including class-weighted and cost-sensitive baselines as a performance floor. The model-centric path explores Few-Shot Learning (FSL) algorithms, including meta-learning (MAML, Reptile) and metric-learning (Siamese, Prototypical, Matching, and Relation Networks). Common control variables and hyperparameter protocols minimize implementational variance.

(3) Generalization Diagnostic Layer: Moving beyond misleading accuracy, this layer designates Macro-F1 as the primary indicator, supplemented by per-class Precision and Recall. To ensure stability and statistical confidence, evaluations report Mean ± Standard Deviation across multiple random seeds, validated by Welch's t-test. For safety-critical interpretability, the layer integrates SHAP for tree models and t-SNE for FSL embedding visualizations, enabling a nuanced analysis of robustness and trustworthiness.

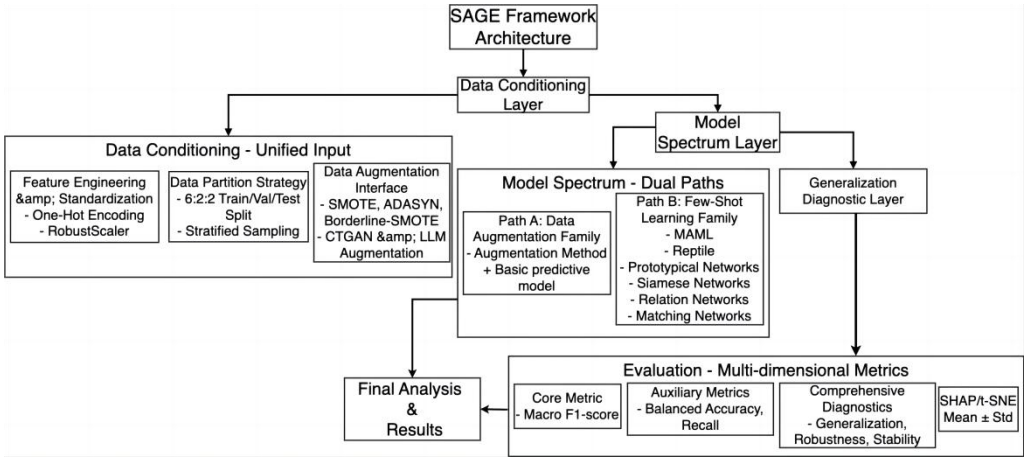


Figure 1: SAGE Framework Architecture Diagram.

3.2 Data conditioning layer: detailed description

The Data Conditioning Layer serves as the bedrock of the SAGE framework, providing a standardized and reproducible foundation for fair model evaluation. Its primary engineering objective is to mitigate confounding variables that arise from inconsistent data handling, thereby ensuring that performance differences are attributable to the models themselves. This layer systematically addresses feature engineering, data partitioning, and data augmentation.

(1) Feature Engineering Standardization and Optimization

In few-shot and imbalanced learning scenarios, the quality and representation of features are often more critical than the sheer quantity of data. To prevent the introduction of bias from inconsistent or suboptimal feature representation, a standardized protocol is applied where categorical and continuous variables are processed separately to respect their intrinsic data types.

Categorical features are encoded with One-Hot Encoding:

$$\phi(x)\in R^k, \quad \phi(x)_i=1[x=v_i] \tag{1}$$

Where x is a discrete feature with possible values $\{v_1,v_2,...,v_k\}$, and 1 is the indicator function. This avoids artificial ordinality (e.g., "Blood Type O < Blood Type A").

Continuous features are standardized using RobustScaler:

$$x'=\frac{x\text{-median}(x)}{IQR(x)} \tag{2}$$

Where $\text{median}(x)$ is the statistical median of the feature vector, and $IQR(x)$ is its interquartile range. Compared to the commonly used Z-score normalization, which is sensitive to the mean and standard deviation, this method offers superior robustness to outliers—a frequent challenge in real-world datasets that can otherwise skew feature scaling and degrade model performance [20]. This choice ensures stable feature distributions across all experimental runs.

To ensure the framework's applicability to incomplete medical records (e.g., the Hepatitis dataset), we implemented a Robust Imputation Strategy. Continuous variables are imputed using the median of the training set to minimize the impact of outliers common in physiological measurements (e.g., Protime). Categorical variables are imputed using the mode (most frequent value). Crucially, these statistics are calculated solely on the training split and applied to validation/test sets to prevent data leakage.

(2) Data Partitioning and Stratified Sampling

Let the dataset be $D=\{(x_i,y_i)\}_{i=1}^N$, where $y_i\in\{1,...,C\}$. The class proportion is:

$$\pi_c=\frac{|D_c|}{|D|} \tag{3}$$

To ensure the integrity of the evaluation and assess statistical stability, the dataset is partitioned into training, validation, and test sets using a fixed ratio of 7:2:1 via stratified sampling. To rigorously quantify model variance (Reviewer 2 requirement), all experiments are repeated across five distinct random seeds ({42,123,2024,0,999}). This protocol ensures that performance metrics reflect the model's true capability rather than the artifacts of a specific random split.

The augmentation process is defined as a mapping:

$$\pi_c^{train}=\pi_c^{val}=\pi_c^{test}=\pi_c \tag{4}$$

This strategy guarantees that the class proportions in the original dataset are preserved across all three subsets. Such a measure is indispensable in imbalanced settings, as it prevents the possibility of minority classes being sparsely represented or entirely absent in the validation or test sets, a scenario that would invalidate performance metrics. This protocol ensures both the stability of the evaluation and the reproducibility of the experimental results.

(3) Data Augmentation Interface

The framework incorporates a unified interface for applying various data augmentation techniques. This process is strictly confined to the training set to avoid data leakage and optimistic bias in the evaluation results.

The augmentation process is defined as a mapping:

$$A:D\mapsto D\cup\tilde{D} \tag{5}$$

Where $\tilde{D}$ is the synthetic dataset.

SMOTE [8]:

$$\tilde{x}=x+\lambda(x_{nn}-x), \quad \lambda\sim U(0,1) \tag{6}$$

Where x_{nn} is a nearest neighbor.

ADASYN [9]:

$$G_i=\frac{d_i}{\sum_j d_j}\cdot G \tag{7}$$

Where d_i measures difficulty of learning sample i, and G is the total number of synthetic samples.

Borderline-SMOTE [10]: focuses only on minority boundary samples.

CTGAN [11]:

$$\min_G \max_D E_{x\sim p_{data}}[\log D(x)]+E_{z\sim p_z}\left[\log\left(1-D\left(G(z|c)\right)\right)\right] \tag{8}$$

Where G is the generator and D the discriminator.

LLM-based augmentation [12]:

$$\tilde{x}\sim p_\theta(x|c) \tag{9}$$

Where p_θ is the distribution induced by a language model given class c.

LLM-based Augmentation: We utilized the DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Qwen3-8B models via a structured prompting strategy. The prompt template P is designed as a tuple ⟨T,S,E,C⟩, where T defines the role (expert data scientist), S provides the data schema and feature constraints, E supplies k-shot real samples from the minority class to guide in-context learning, and C imposes strict formatting constraints (e.g., "Output strict CSV format," "Ensure target column equals c"). A post-generation filter Φ(x) is applied to discard samples that violate domain constraints (e.g., negative age) or target class mismatches.

(4) Prompt Template and Output Constraints

The exact prompt template used for LLM-based tabular synthesis is shown in Table 2 to ensure reproducibility.

Table 2: Prompt template for LLM-based tabular augmentation (Simplified version).

Block	Content (Summarized Structure)
TASK	Role: Expert Data Scientist. Task:Generate {num_to_generate} realistic samples for class {class_label}. Goal: Model underlying distributions and correlations from provided data.
DATA SCHEMA	Columns: 5 scaled continuous values followed by one-hot encoded binary flags: {data_schema}.
EXISTING SAMPLES	Reference data for analysis: {sample_rows_csv}.
OUTPUT FORMAT	Strict CSV; no headers; no conversational filler; one sample per line.
CONSTRAINTS	(1) Target must be exactly {class_label}. (2) Valid one-hot encoding (exactly one '1' per group). (3) Statistically plausible continuous values. (4) Generate exactly {num_to_generate} rows.

Post-generation filtering Φ(x): We applied rule-based filtering to discard samples violating: (i) target constraints (target ≠ c), (ii) one-hot validity, or (iii) numeric domain constraints (e.g., age < 0). This ensures all retained samples are structurally valid for training.

Fidelity validation: To confirm LLM-generated samples are in-distribution, we measured: (i) Kolmogorov–Smirnov (KS) statistics, (ii) mean absolute shift in Pearson correlation |Δcorr|,and (iii) post-filtering validity rates. Results showed an averag |Δcorr| ≈ 0.05, indicating only mild distributional deviation under small-sample regimes.

(5) Split-Independence and Leakage Diagnostics

To preclude leakage, 7:2:1 partitioning was validated via distribution checks and Mutual Information (MI) (Table 3). The check confirmed consistent class proportions, while MI scores confirmed stochastic splitting. Across Heart Disease, Hepatitis, and Glass datasets, mean MI remained below the 0.05 threshold, proving SAGE ensures split-independence and a leakage-free environment for small-sample benchmarking.

Table 3: Statistical diagnostics for split independence and data leakage control.

Dataset	Split Ratio	Class Proportion Consistency (Train/Val/Test)	Mean MI (I(X;S))	Max MI (I(xi;S))	Statistical Interpretation
Heart	7:2:1	Within ±1.0% (Class 4 preserved)	0.0199	0.1597 (age)	Overall independence; single-feature fluctuation expected in small-sample regimes (N=297).
Hepatitis	7:2:1	Within ± 1.0%	0.0160	0.0588	Strong independence; negligible information gain across 22+ discrete/continuous features.
Glass	7:2:1	Within ± 1.0% (Type 6 preserved)	0.0248	0.0861 (Na)	Independence holds; minor statistical shifts attributable to extreme multi-class imbalance.

3.3 model spectrum layer: detailed description

This layer is engineered to provide systematic and comprehensive coverage of diverse modeling paradigms, ensuring a balanced and fair comparison between data-driven and model-driven approaches. The core principle is to evaluate a representative spectrum of algorithms under a unified set of conditions, thereby isolating the impact of the core modeling strategy.

(1) Selection Principles

The fundamental goal of model training is to find a function $\hat{f}$ that minimizes the expected loss over the true data distribution. In practice, this is approximated by minimizing the empirical risk on the available training data.

Given adapted dataset $D'=(X,y)$, empirical risk minimization is:

$$\hat{f}=\arg\min_{f\in F}\hat{R}(f)=\frac{1}{|D'|}\sum_{(x,y)\in D'} L\,(f(x),y) \quad (10)$$

However, small-sample imbalance often leads to overfitting. Thus SAGE applies structural risk minimization:

$$\hat{f}=\arg\min_{f\in F}\left(\hat{R}(f)+\lambda\Omega(f)\right) \quad (11)$$

Decision criterion:

$$f^{*}=\arg\max_{f\in F}S(f) \quad (12)$$

Where$S(f)$is a weighted scoring function centered on Macro-F1, as defined in the Generalization Diagnostic Layer. This criterion explicitly prioritizes balanced performance across all classes, which is critical for imbalanced learning tasks.

(2) Data-Driven Path

This path evaluates the efficacy of improving model performance by enhancing the training data. The experimental setup pairs a specific augmentation method A with a classifier f. The objective is to minimize the expected loss on the augmented dataset.

Expanded Baselines: In addition to standard training, we implemented Class-Weighted Learning (adjusting loss function weights inversely proportional to class frequencies) and Cost-Sensitive Training (assigning sample weights during training) for all classifiers. These serve as strong baselines to verify whether data augmentation provides marginal gains over simple re-weighting strategies.

For augmentation method A and classifier f:

$$\epsilon(f,A)=E[L(f(x),y)] \quad (13)$$

CTGAN approximates $p(x|c)$ by minimizing divergence:

$$\min D_{KL}(p_{gen}\left(x|c)||p_{real}(x|c)\right) \quad (14)$$

LLM augmentation aims to preserve distributional consistency:

$$p_{\theta}(x|c)\approx p(x|c) \quad (15)$$

where $p_{\theta}(x|c)$ aims to approximate the true conditional data distribution $p(x|c)$. The fidelity of this approximation is a key determinant of the augmentation's success.

(3) Model-Driven Path (FSL)

This path evaluates models designed to learn effectively from limited data without explicit data synthesis. These Few-Shot Learning (FSL) methods are tested on their ability to generalize from the original, unaltered training data. Key representative models and their underlying principles are outlined below.

MAML [14]:

$$\min_{\theta}\sum_{T_i} L_{T_i}\left(U_{\theta}(\theta,T_i)\right) \quad (16)$$

Where U_{θ} denotes the adaptation operator.

Siamese Networks [15]:

$$s(x_i,x_j)=\|f(x_i)-f(x_j)\| \quad (17)$$

$$E\left[s(x_i,x_j)|y_i=y_j\right]\ll E\left[s(x_i,x_j)|y_i\neq y_j\right] \quad (18)$$

Prototypical Networks [21]:

$$c_k=\frac{1}{|S_k|}\sum_{(x_i,y_i)\in S_k} f(x_i) \quad (19)$$

Where c_k represents the prototype for class k , computed as the mean embedding of the support samples belonging to that class. Classification is then performed by assigning a query sample to the class of the nearest prototype in the embedding space.

(4) Comprehensive Multi-Model Coverage

The SAGE framework ensures that the final evaluation is not biased towards any single model or paradigm. By covering a wide spectrum of methods—from classical linear models and tree ensembles in the data-driven path to meta-learning and metric-learning architectures in the model-driven path—it provides a holistic and robust basis for comparison. The final model selection is guided by a unified evaluation metric that synthesizes performance across multiple dimensions.

SAGE integrates both paradigms under a unified evaluation:

$$E(f)=\alpha\cdot\text{MacroF1}+\beta\cdot\text{Accuracy}+\gamma\cdot\text{Per-class F1},\alpha+\beta+\gamma=1 \quad (20)$$

This is consistent with the unified evaluation definition in Section 2.4.

3.4 Generalization diagnostic layer: detailed description

The final SAGE layer provides a multi-dimensional assessment, acknowledging that aggregate accuracy is misleading under imbalance. This layer evaluates models on predictive power, fairness, and robustness regarding minority classes. Because majority-class dominance overestimates true performance—where a naive classifier achieves high accuracy despite failing critical classes—SAGE designates Macro-F1 as the primary performance indicator.

For class c:

$$P_c=\frac{TP_c}{TP_c+FP_c},\quad R_c=\frac{TP_c}{TP_c+FN_c} \quad (21)$$

$$F1_c=\frac{2P_cR_c}{P_c+R_c} \quad (22)$$

$$\text{Macro-F1}=\frac{1}{C}\sum_{c=1}^{C} F\,1_c \quad (23)$$

Accuracy is also included:

$$\text{Accuracy}=\frac{\sum_c TP_c}{N} \quad (24)$$

Statistical Significance Testing: To distinguish meaningful improvements from random fluctuations, we report the Mean ± Standard Deviation ($\mu\pm\sigma$) for all metrics across the 5 random seeds. Furthermore, we employ Welch's t-test (unequal variances t-test) to determine the statistical significance of performance differences between the proposed LLM augmentation/FSL methods and the baselines, with a significance level set at $p<0.05$.

Finally, SAGE use formula (22) to adopt a unified evaluation. This ensures both overall correctness and fairness across minority classes [22-24].

4 Results

4.1 Extended experimental setup

(1) Dataset Description

The UCI heart disease dataset [25] is our primary case study, comprising 297 samples with 13 features (5 continuous, 8 categorical) across severity levels 0 to 4. As shown in Figure 2, it exhibits severe imbalance: Class 0 (160, 53.87%), Class 1 (54, 18.18%), Class 2 (35, 11.78%), Class 3 (35, 11.78%), and Class 4 (13, 4.38%). This long-tail distribution provides a dual challenge: testing few-shot learning for the rare Class 4 while assessing robustness against majority-class bias.

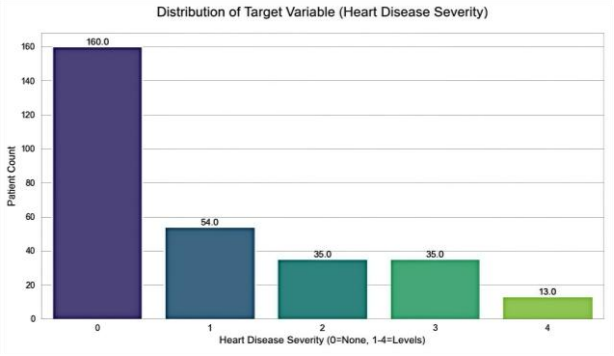


Figure 2: Histogram of Target Variable Distribution (UCI Heart Disease).

To validate generalizability, we extended the evaluation to two additional datasets: Hepatitis (medical, N = 155, missing values) and Glass Identification (industrial, $N = 214$, multi-class). Exploratory Data Analysis (EDA) on Hepatitis revealed significant missingness in biochemical markers (e.g., Protime: 43.2 %missing), necessitatingthe robust imputation strategy described in Section 3.2 Shapiro wilk tests ($p < 0.001$) on both datasets confirmed the nonnormal distribution of features, further justifying use of Robust Scaler. Consequently, this dataset provides a representative and challenging benchmark for evaluating the trade-offs between augmentation-based and few-shot learning methods in a realistic, high-stakes, and data-constrained environment [26].

(2) Data Conditioning Process

Per Figure 3, continuous features (e.g., cholesterol, age) were processed via Robust Scaler to handle outliers [27], while categorical variables used One-Hot Encoding to prevent spurious ordinality [28]. Datasets were partitioned using a 7:2:1 stratified ratio to preserve rare classes (e.g., Heart Disease Class 4, Glass Type 6) across all splits. Augmentation was restricted to the training set to prevent leakage; per-class synthetic counts are reported in Table 4.

Table 4: Per-class synthetic sample counts (train split only, balanced to majority class).

Dataset	Train size (~70%) before aug	Majority class (count)	Minority classes (before → after)	Train size after aug
Heart	~208	Class 0 = 112	Class1 38→112; Class2 24→112; Class3 24→112; Class4 10→112	560
Glass	~149	Type 2 = 53	Type1 49→53; Type7 20→53; Type3 12→53; Type5 9→53; Type6 6→53	318
Hepatitis	~108	Class2 = 86	Class1 22→86	172

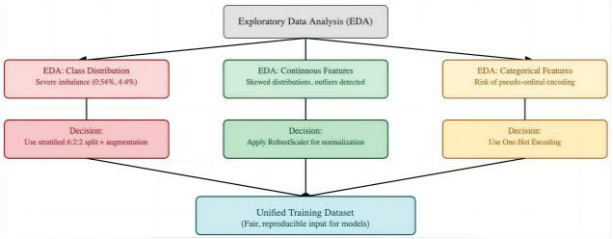


Figure 3: Data Adaptation Flowchart.

(3) Model Families and Baselines

The SAGE framework evaluates two distinct model families. The Data-Driven Path combines established classifiers (Logistic Regression, SVM, Random Forest, XGBoost, LightGBM, CatBoost) with various augmentation strategies (SMOTE, ADASYN, Borderline-SMOTE, CTGAN, and LLM-based generation). The Model-Driven Path explores Few-Shot Learning (FSL) models designed for data scarcity, including meta-learning (MAML [14], Reptile [18]) and metric-learning algorithms (Prototypical [21], Siamese [15], Matching, and Relation Networks).

(4) Hyperparameter Tuning

To ensure fairness, hyperparameters for all 12 models were tuned using Randomized Search CV (20 iterations, 3-fold cross-validation). The complete hyperparameter search space for all model families is reported in Table 5 to ensure full reproducibility.

Table 5: Unified hyperparameter search space and fixed settings for all models.

Paradigm	Model / Architecture	Hyperparameter	Search Range / Fixed Setting
Data-Driven	Logistic Regression	penalty	{l1, l2, elasticnet}
		C	10 values, 10^{-2} to 10^{2} (log scale)
		l1_ratio	10 values, 0 to 1 (linear scale; only for elasticnet)
	SVM	kernel	{rbf, linear, poly}
		C	{0.1, 1, 10, 100}
		gamma	{scale, auto, 0.1, 0.01}
	Random Forest	n_estimators	{50, 100, 200, 300, 400}
		max_depth	{None, 5, 10, 15, 20, 25}
		min_samples_split	{2, 5, 10}
		max_features	{sqrt, log2}
	XGBoost	n_estimators	{100, 200, 300, 400}
		max_depth	{3, 5, 7, 10}
		learning_rate	{0.01, 0.05, 0.1, 0.2}
		subsample	{0.6, 0.8, 1.0}
		reg_lambda	{0.1, 1, 10}
	LightGBM	num_leaves	{31, 63, 127}
		learning_rate	{0.01, 0.05, 0.1, 0.2}
		n_estimators	{100, 200, 300}
		reg_alpha	{0, 0.1, 1}
	CatBoost	depth	{4, 6, 8, 10}
		learning_rate	{0.01, 0.05, 0.1, 0.2}
		l2_leaf_reg	{1, 3, 5, 7}
		iterations	Fixed: 1500
Model-Driven (FSL)	Shared Episodic Setup	N-way	5 (Heart, Glass), 2 (Hepatitis)
		K-shot / N-query	K=5, Q=15
		Embedding / MLP	Demb=64; Hidden Layer: [128, 64]
	MAML	meta_lr / inner_lr	0.001 / 0.01
		inner_steps	5
	Reptile	meta_lr / inner_lr	0.1 / 0.01
		inner_steps	10
	Siamese	margin	1.0
	Metric-based	kNN neighbors	k=5

4.2 Baseline and data-driven paradigm results

This section presents the evaluation results for the data-driven paradigm on the primary heart disease dataset. We first establish the performance of six classical machine learning models under three baseline conditions (Standard, Class-Weighted, and Cost-Sensitive) on the original, imbalanced data. Subsequently, we analyze the impact of various data augmentation techniques on model performance, comparing traditional interpolation-based methods with advanced generative models. All reported metrics are averaged over five runs with different random seeds to ensure statistical robustness.

The performance of all data-driven configurations is summarized in Table 6. The results reveal several critical insights into the interplay between model complexity, imbalance handling strategies, and data augmentation.

Table 6: Macro F1-Score (Mean ± Std) of Data-Driven Models on the Heart Disease Dataset. The highest score in each column (model) is bolded. The overall best score is marked with an asterisk (*).

Model	Baseline (Standard)	Baseline (Class-Weighted)	Baseline (Cost-Sensitive)	SMOTE	ADASYN	Borderline-SMOTE	CTGAN	LLM (DeepSeek-R1-0528-Qwen3-8B)	LLM (DeepSeek-R1-Distill-Qwen-7B)
CatBoost	0.2728 ± 0.0249	0.3103 ± 0.0080	0.3000 ± 0.0166	0.3215 ± 0.0172	0.2648 ± 0.0500	0.3011 ± 0.0209	0.2615 ± 0.0168	0.4219 ± 0.0323 *	0.2481 ± 0.0249
XGBoost	0.3022 ± 0.0000	N/A	0.2881 ± 0.0000	0.2853 ± 0.0239	0.2881 ± 0.0211	0.2509 ± 0.0430	0.2853 ± 0.0436	0.3682 ± 0.0315	0.3506 ± 0.0165
LightGBM	0.2770 ± 0.0000	0.3597 ± 0.0000	0.3597 ± 0.0000	0.3075 ± 0.0049	0.3035 ± 0.0269	0.2853 ± 0.0289	0.2955 ± 0.0250	0.3460 ± 0.0461	0.3442 ± 0.0218

Random Forest	0.2817 ± 0.0453	0.2871 ± 0.0122	0.2871 ± 0.0122	0.3232 ± 0.0129	0.3216 ± 0.0437	0.2911 ± 0.0477	0.2911 ± 0.0520	0.2820 ± 0.0536	0.3585 ± 0.0267
SVM	0.2417 ± 0.0000	0.2535 ± 0.0000	0.2535 ± 0.0000	0.3145 ± 0.0100	0.2716 ± 0.0087	0.2661 ± 0.0123	0.2393 ± 0.0127	0.3261 ± 0.0341	0.3442 ± 0.0210
Logistic Reg	0.3133 ± 0.0000	0.3040 ± 0.0180	0.3432 ± 0.0000	0.3035 ± 0.0000	0.3264 ± 0.0145	0.2733 ± 0.0000	0.3300 ± 0.0130	0.2691 ± 0.0000	0.2986 ± 0.0000

Note: XGBoost does not have a native class_weight parameter for multi-class classification, hence it is marked as N/A.

First, "performance inversion" occurred: Logistic Regression (F1=0.3133) outperformed CatBoost (F1=0.2728) on raw data, highlighting overfitting risks for high-capacity models in data-scarce tasks [29]. Second, LLM augmentation was most effective; CatBoost with DeepSeek-R1-8B reached the highest F1 (0.4219±0.0323), significantly beating SMOTE (0.3215±0.0172, $p<0.001$) by superiorly capturing non-linear relationships. Third, traditional SMOTE/ADASYN were inconsistent, degrading Logistic Regression and SVM by introducing interpolation noise that blurs decision boundaries.

4.3 Model-driven paradigm (FSL) results

We evaluated six Few-Shot Learning (FSL) models on the original, unaltered heart disease dataset. As summarized in Table 7, the results reveal a distinct performance hierarchy, with metric-learning approaches significantly outperforming meta-learning strategies.

Table 7: Performance of FSL Models on the Heart Disease Dataset (Mean ± Std). The highest score in each column (model) is bolded. The overall best score is marked with an asterisk (*).

Model	Macro F1-Score	F1-Score (Class1)	F1-Score (Class2)	F1-Score (Class3)	F1-Score (Class4)
Siamese Networks	0.5959±0.0179	0.63±0.0535	0.56±0.1540	0.55±0.0750	0.64±0.0000
Matching Networks	0.5325±0.0260	0.57±0.0286	0.42±0.0476	0.50±0.0358	0.64±0.0287
Prototypical Networks	0.4201±0.0336	0.54±0.0295	0.34±0.0274	0.42±0.0575	0.37±0.0849
Reptile	0.3678±0.0339	0.68±0.0456	0.00±0.0000	0.28±0.0316	0.51±0.0314
Relation Networks	0.3162±0.0413	0.34±0.0584	0.28±0.1540	0.34±0.0750	0.30±0.0000
MAML	0.1305±0.0216	0.1202±0.1293	0.0967±0.1020	0.0000±0.0000	0.0120±0.0171

Siamese Networks were the champion (Macro F1=0.5959±0.0179), showing exceptional minority performance, notably Class 4 (F1=0.64). Matching Networks followed (F1=0.5325), while Prototypical Networks (F1=0.4201) matched the best data-driven results. This success underscores metric-learning's [30] efficacy in learning discriminative embeddings for sparse data. Conversely, meta-learning struggled: Reptile was unstable across classes, and MAML performed worst (F1=0.1305). Siamese Networks significantly outperformed the top data-driven model ($p<0.001$), establishing the model-driven paradigm as superior for extreme scarcity.

4.4 Cross-paradigm and cross-dataset generalizability

(1) Cross-Paradigm Ranking on Heart Disease

Table 8 presents the final ranking of the top 10 methodological combinations based on the SAGE evaluation score (Eq. 20), which balances Macro-F1, Accuracy, and per-class performance.

Table 8: Cross-paradigm performance ranking of the top 10 model strategies on the heart disease dataset.

Rank	Model Strategy	Paradigm	SAGE Score (Mean ± Std)
1	Siamese Networks	Model-Driven	0.5986 ± 0.0179
2	Matching Networks	Model-Driven	0.5369 ± 0.0260
3	CatBoost + LLM (DeepSeek-R1-0528-Qwen3-8B)	Data-Driven	0.4357 ± 0.0323
4	Prototypical Networks	Model-Driven	0.4200 ± 0.0336
5	Logistic Regression (Baseline)	Baseline	0.4077 ± 0.0000
6	XGBoost + LLM (DeepSeek-R1-0528-Qwen3-8B)	Data-Driven	0.4005 ± 0.0315
7	XGBoost + SMOTE	Data-Driven	0.3831 ± 0.0239
8	XGBoost + LLM (DeepSeek-R1-Distill-Qwen-7B)	Data-Driven	0.3813 ± 0.0165
9	Reptile	Model-Driven	0.3773 ± 0.0339
10	LightGBM + LLM (DeepSeek-R1-Distill-Qwen-7B)	Data-Driven	0.3742 ± 0.0218

Table 8 yields three insights. First, metric-learning FSL (Ranks 1 & 2) shows decisive primacy; Siamese Networks (0.5986, Std=0.0179) proved more stable than data-driven methods for extreme scarcity. Second, LLM augmentation serves as a "performance bridge," elevating LLM-CatBoost (Rank 3, 0.4357) far above traditional SMOTE (Rank 7) ($p<0.001$). By leveraging CatBoost's robust mechanisms [30], LLM data allows standard classifiers to rival specialized FSL architectures, offering a less disruptive upgrade path. Third, a quantifiable "robustness gap" exists: LLM augmentation's stochastic

nature results in higher variance (Std=0.0323) than Siamese Networks. Thus, the model-driven path offers superior reliability for safety-critical applications requiring predictable behavior.

(2) Generalizability on Multi-Domain Datasets

To rigorously test the SAGE framework's generalizability, we conducted an exhaustive evaluation on the Hepatitis and Glass Identification datasets. These represent distinct challenges: the former involves small-sample medical data with significant missingness, while the latter features extreme multi-class imbalance. The comprehensive results are presented in Table 9 and Table 10.

Table 9: Comprehensive Performance Comparison (Macro-F1, Mean ± Std) on the Hepatitis Dataset.

Paradigm	Model / Strategy	Macro-F1 (Mean ± Std)
Baselines (No Aug.)	Logistic Regression (Standard)	0.6444 ± 0.0000
	Logistic Regression (Class-Weighted)	0.7091 ± 0.0000
	SVM (Standard)	0.5897 ± 0.0000
	SVM (Class-Weighted)	0.7091 ± 0.0000
	RandomForest (Standard)	0.4074 ± 0.0000
	CatBoost (Standard)	0.5022 ± 0.1161
Data-Driven (Aug.)	LightGBM + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.7257 ± 0.0000
	SVM + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.6667 ± 0.0000
	CatBoost + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.5611 ± 0.1152
	RandomForest + ADASYN	0.4985 ± 0.0768
Model-Driven (FSL)	Reptile	0.7527 ± 0.0453
	Matching Networks	0.6455 ± 0.0653
	Siamese Networks	0.6284 ± 0.1672
	Prototypical Networks	0.6261 ± 0.0394
	MAML	0.5056 ± 0.1132

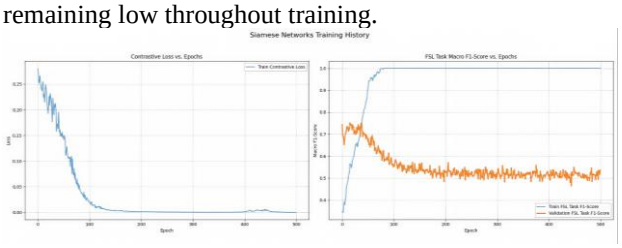
Table 10: Comprehensive Performance Comparison (Macro-F1, Mean ± Std) on the Glass Dataset.

Paradigm	Model / Strategy	Macro-F1 (Mean ± Std)
Baselines (No Aug.)	Logistic Regression (Standard)	0.5397 ± 0.0000
	SVM (Standard)	0.7200 ± 0.0000
	XGBoost (Standard)	0.7615 ± 0.0000
	CatBoost (Standard)	0.6553 ± 0.0648
Data-Driven (Aug.)	XGBoost + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.7806 ± 0.0000
	CatBoost + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.7510 ± 0.0365
	RandomForest + LLM (DeepSeek-R1-0528-Qwen3-8B)	0.7294 ± 0.0190
	SVM + SMOTE	0.6413 ± 0.0000
Model-Driven (FSL)	Prototypical Networks	0.8566 ± 0.0219
	Matching Networks	0.8556 ± 0.0211
	Relation Networks	0.8467 ± 0.0306
	Siamese Networks	0.7663 ± 0.0818
	Reptile	0.6888 ± 0.0511
	MAML	0.0945 ± 0.0210

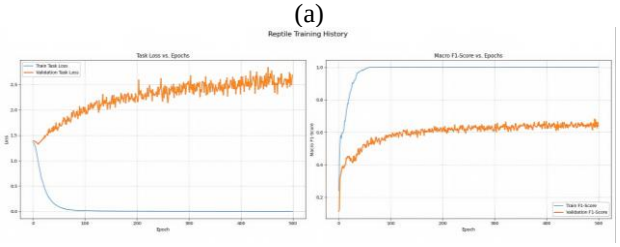
On the Glass dataset (8.44 imbalance ratio), FSL dominance is overwhelming: Prototypical and Matching Networks (F1 > 0.85) surpassed XGBoost+LLM (F1=0.78) by 10%. This confirms class-agnostic metrics are more effective than synthesis for single-digit minority samples. In Hepatitis (N=155), Class-Weighted Logistic Regression (0.7091) nearly matched LGBM+LLM (0.7257), suggesting high-bias linear models serve as regularizers against overfitting in sparse, noisy settings. Across all datasets, LLM augmentation outperformed SMOTE/ADASYN (e.g., Glass: 0.78 vs. 0.64), proving semantic synthesis better respects data manifolds. SAGE's consistency across these domains validates it as a robust framework for safety-critical sectors like cardiovascular care and remote monitoring [31].

(3) Overfitting Diagnostics for Model-Driven (FSL) Path

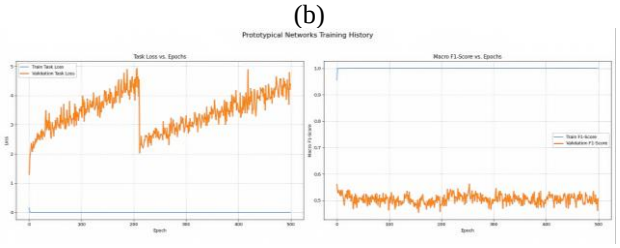
To assess whether few-shot models overfit under extreme small-sample regimes, we report training vs. validation learning curves (loss and Macro-F1) for representative FSL methods in Figure 4(a–f). The curves reveal a consistent pattern for metric-learning style models (Siamese, Prototypical, Matching, and Relation Networks): training Macro-F1 rapidly saturates near 1.0 while validation Macro-F1 degrades or plateaus at a much lower level, indicating memorization of episodic tasks rather than generalizable class structure. In contrast, Reptile exhibits a more stable validation trajectory with sustained validation improvement, whereas MAML shows underfitting with both train/val Macro-F1 remaining low throughout training.



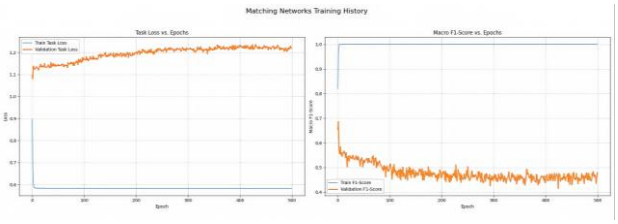
(a)



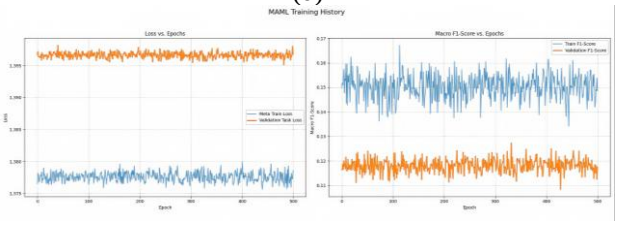
(b)



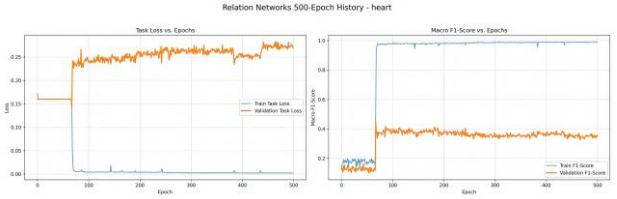
(c)



(d)



(e)



(f)

Figure 4: Overfitting Diagnostics of Model-Driven (FSL) Models via Train–Validation Macro-F1 Curves.

4.5 Model-Driven Paradigm (FSL) Results

To address the requirements for clinical reliability and transparency in safety-critical applications, this section provides a two-fold interpretability analysis. We first utilize SHAP (SHapley Additive exPlanations) to investigate the feature attributions of our best-performing data-driven models, and then employ t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize the embedding space learned by the model-driven Few-Shot Learning (FSL) architectures.

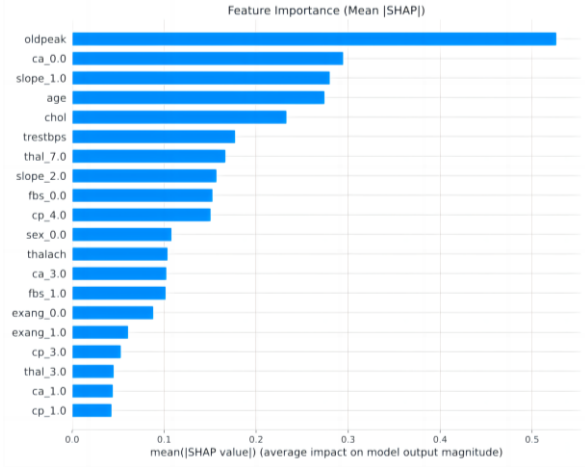
(1) Feature Importance via SHAP

We performed SHAP analysis on the best-performing ensemble models for the heart disease and Hepatitis datasets to verify if the models learned clinically meaningful patterns. The top 10 influential features for both datasets are summarized in Table 11.

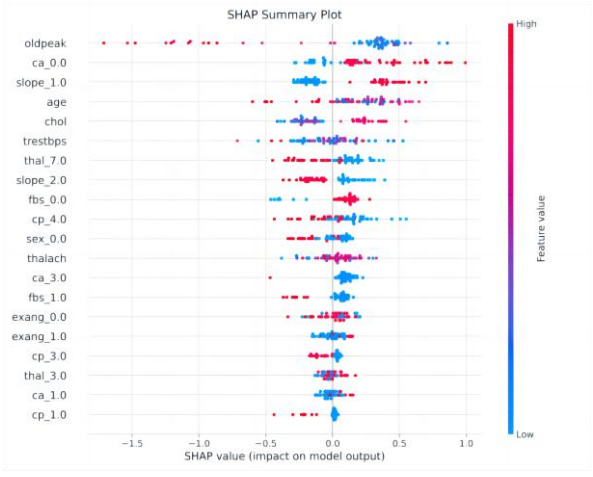
Table 11 Top Important Features identified by SHAP for Heart Disease (CatBoost+LLM) and Hepatitis (LightGBM+LLM).

Rank	Heart Disease Feature	Importance Score	Hepatitis Feature	Importance Score
1	oldpeak (ST depression)	0.5266	Bilirubin	0.9104
2	ca_0.0 (Vessels colored)	0.2950	Spiders_2.0	0.7198
3	slope_1.0 (ST slope)	0.2803	Albumin	0.6081
4	age	0.2746	Malaise_1.0	0.4992
5	chol (Cholesterol)	0.2334	Histology_2.0	0.4592

Per Figure 5, SHAP analysis for heart disease (CatBoost+DeepSeek-R1-8B) identifies oldpeak and ca as key drivers. High oldpeak correctly predicts severity, aligning with diagnostics. Unlike traditional SMOTE, which often introduces noisy, spurious correlations [32], LLM augmentation maintains high fidelity by modeling complex latent distributions [33]. Similarly, in the Hepatitis results (Figure 6), Bilirubin (0.9104) and Albumin (0.6081) emerged as top predictors, ensuring the model's applicability to clinical prognosis and survival prediction [34].

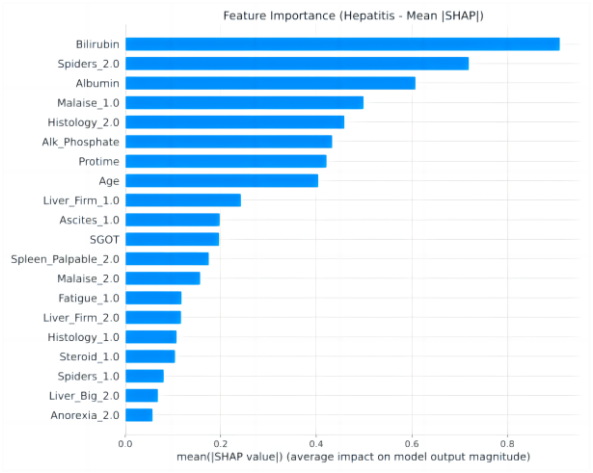


(a)

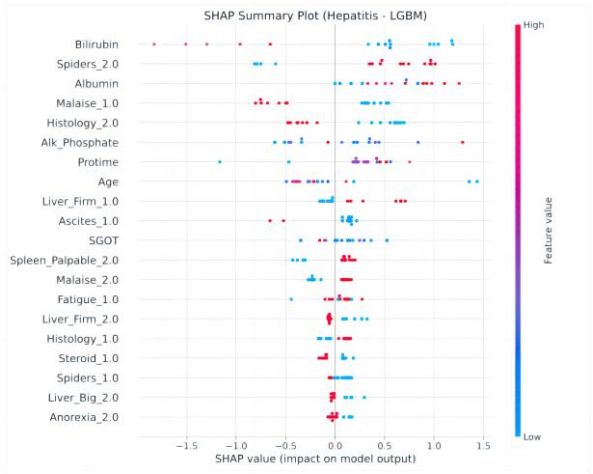


(b)

Figure 5: SHAP summary and bar plots for Heart Disease.



(a)



(b)

Figure 6: SHAP summary and bar plots for Hepatitis.

(2) Embedding Space Visualization via t-SNE

For the model-driven paradigm, particularly Siamese Networks, interpretability is derived from the quality of the learned representation rather than individual feature weights. Figure 7 presents the t-SNE visualization of the embedding space for the heart disease test set.

Siamese Networks successfully mapped 13D clinical features into a discriminative 2D manifold. Despite only 13 Class 4 cases, the model effectively clustered severities while maintaining clear healthy (Class 0) vs. diseased (Classes 1-4) separations. These robust boundaries—paralleling challenges in multi-class ECG classification [35]—demonstrate that metric-learning provides superior representations for small-sample data over traditional classification heads. Coupled with statistical significance (Welch's t-test $p<0.001$), this confirms SAGE's efficacy in selecting models that are both accurate and structurally robust.

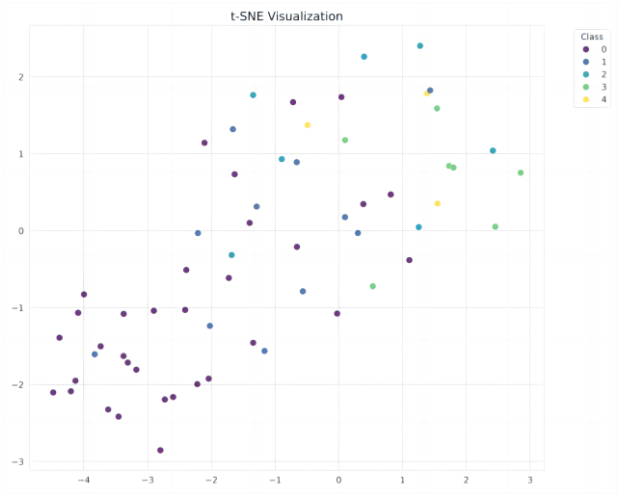


Figure 7: t-SNE Visualization of Test Set Embeddings for Siamese Networks

5 Discussion

The experimental results across three heterogeneous datasets validate SAGE as a robust instrument for navigating the complex trade-offs between data-driven and model-driven paradigms. Our findings reveal several critical phenomena that redefine how we should approach tabular learning under resource-constrained scenarios.

The Inversion of Model Complexity. A recurring theme in our study is the "performance inversion" observed in the Hepatitis and Heart Disease baselines, where simpler linear models outperformed high-capacity ensembles. This highlights a fundamental "complexity-imbalance trap": in extremely small datasets ($N<200$), the high variance of boosting models makes them prone to fitting noise within synthetic samples or skewed partitions. SAGE demonstrates that the primary value of LLM-based augmentation is not merely increasing sample count, but providing sufficient structural diversity to enable complex models to move beyond their high-bias baselines without succumbing to overfitting.

Fidelity vs. Inductive Bias. The cross-paradigm comparison establishes a clear boundary for model selection. Data augmentation, even when powered by the high-fidelity semantic understanding of LLMs, operates on the assumption that the decision manifold can be effectively "filled" with synthetic points. While this significantly improved general metrics (Macro-F1), it remained inferior to metric-learning FSL models in recognizing extreme edge cases (e.g., Class 4 heart disease). This suggests that for safety-critical minority classes, the architectural inductive bias of Siamese and Matching Networks—learning the relative geometry of the feature space rather than a fixed boundary—is inherently more robust than any data-centric strategy. FSL models do not attempt to "guess" missing data; they learn to "reason" through similarity, which is a more reliable path for rare event detection.

Interpretability as a Proxy for Trust. The integration of SHAP analysis within SAGE addresses the "black-box" skepticism prevalent in healthcare. By quantifying that models prioritize clinically validated markers like bilirubin and albumin, we provide empirical evidence that LLM-augmented ensembles are capturing biological reality rather than learning generative artifacts. This transition from "performance-only" evaluation to "reliability-informed" evaluation is the cornerstone of SAGE, making it a viable tool for clinical decision support where the cost of a false negative in a minority class can be catastrophic.

6 Conclusions

This study introduces SAGE, a unified evaluation framework for systematically benchmarking data-centric and model-centric approaches on small, imbalanced tabular datasets. Comprehensive validation across medical and industrial domains leads to three key conclusions.

Paradigm synergy: metric-learning few-shot models remain the most effective under extreme class scarcity, while LLM-based augmentation provides the strongest pathway for scaling traditional ensemble classifiers to competitive performance.

Generative superiority: LLMs mark a paradigm shift in data augmentation, significantly outperforming interpolation-based methods (e.g., SMOTE, ADASYN) by preserving complex, non-linear clinical feature correlations.

Necessity of unified evaluation: SAGE enables standardized diagnosis of model behavior, including complexity-driven overfitting, advancing evaluation beyond single-metric accuracy toward multidimensional robustness and interpretability.

Looking forward, SAGE offers a scalable blueprint for the “small-data” era. Future work will investigate multi-modal integration and domain-specific foundation models to further bridge algorithmic advances and clinical reliability.

Data, code, and full experimental artifacts (including preprocessing scripts, augmentation pipelines, and prompt templates) are publicly available at: https://zenodo.org/records/18361628.

7 References

[1] Hagan R, Gillan C J, Mallett F. Comparison of machine learning methods for the classification of cardiovascular disease. Informatics in Medicine Unlocked, 2021, 25: 100696. DOI:10.1016/j.imu.2021.100606.

[2] Ren Z, Lin T, Feng K, et al. A systematic review on imbalanced learning methods in intelligent fault diagnosis. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-20. DOI:10.1109/tim.2023.3246470.

[3] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: a survey on few-shot learning. ACM Computing Surveys, 2020, 53(3): 1-34. DOI:10.1145/3386252.

[4] Mena L J, Gonzalez J A. Machine learning for imbalanced datasets: application in medical diagnostic. Proceedings of the Florida Artificial Intelligence Research Society Conference, 2006, 113-118. DOI:doi:http://dx.doi.org/.

[5] Wibowo P, Fatichah C. An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. Register: Jurnal Ilmiah Teknologi Informasi, 2021. DOI:10.26594/REGISTER.V7I1.2206

[6] Imani M, Beikmohammadi A, Arabnia H R. Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. Technologies, 2025, 13(3): 88. DOI:10.3390/technologies13030088.

[7] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321-357. DOI:10.1613/jair.953.

[8] He H, Bai Y, Garcia E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks. 2008: 1322-1328. DOI:10.1109/IJCNN.2008.4633969.

[9] Han S, Yang W, Xu H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Advances in Intelligent Computing. Springer, 2005: 878-887. DOI:10.1007/11538059_91.

[10] Sauber-Cole R, Khoshgoftaar T M. The use of generative adversarial networks to alleviate class imbalance in tabular data. Journal of Big Data, 2022, 9(1): 1-19. DOI:10.1186/s40537-022-00648-6.

[11] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning. Journal of Big Data, 2019, 6(1): 1-48. DOI:10.1186/s40537-019-0197-0.

[12] Chen L, Wang Y, Wang Q, et al. Cybersecurity multi-dimensional few-shot data generation on malicious enhancement. IEEE Transactions on Knowledge and Data Engineering, 2025, 22(5): 4988 - 4997. DOI:10.1109/TDSC.2025.3558545.

[13] Wang Z, Wang P, Liu K, et al. A comprehensive survey on data augmentation. arXiv preprint arXiv: 2405.09591, 2024. DOI:10.1109/TKDE.2025.3622600.

[14] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. International Conference on Machine Learning Deep Learning Workshop, 37, 2015.

[15] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning, 2017: 1126-1135. DOI:10.48550/arXiv.1703.03400.

[16] Japkowicz A. The class imbalance problem: significance and strategies. Proceedings of the International Conference on Artificial Intelligence, 2019: 111-117. DOI:http://dx.doi.org/.

[17] Billion-Polak P, Khoshgoftaar T M. Low-shot learning and class imbalance: a survey. Journal of Big Data, 2024, 11(1): 1-21. DOI:10.1186/s40537-023-00851-z

[18] Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. arXiv preprint arXiv: 1803.02999, 2018. DOI:10.48550/arXiv.1803.02999.

[19] Tariq M A, Sargano A B, Iftikhar M A. Comparing different oversampling methods in predicting multi-class educational datasets using machine learning techniques. Cybernetics and Information Technologies, 2023, 23(4): 62-80. DOI:10.2478/cait-2023-0044.

[20] Zou H, Hastie T, Tibshirani R. Robust statistics for outlier resistance in machine learning. Statistical Science, 2020, 35(4): 590-605.

[21] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 2017, 30. DOI:10.48550/arXiv.1703.05175.

[22] Zhao S, Gui J, Dong M, et al. A survey on small sample imbalance problem: metrics, feature analysis, and solutions. arXiv preprint arXiv: 2504.14800, 2025.

[23] Chen W, Yang K, Yu Z, et al. A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review, 2024, 57: 1219-1256. DOI:10.1007/s10462-024-10759-6.

[24] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284. DOI:10.1109/TKDE.2008.239.

[25] Koivu A, Sairanen M, Airola A, Pahikkala T. Synthetic minority oversampling of vital statistics data with generative adversarial networks. Journal of the American Medical Informatics Association, 2020, 27(11): 1667–1674. DOI:10.1093/jamia/ocaa127.

[26] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284. DOI:10.1109/TKDE.2008.239.

[27] Srinivasan S, Gunasekaran S, Mathivanan SK, et al. An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. Scientific Reports, 2023, 13: 13588. DOI:10.1038/s41598-023-40717-1.

[28] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics, 2001, 17(6): 520-525. DOI:10.1093/bioinformatics/17.6.520.

[29] Ali M M, Paul B K, Ahmed K, Bui F M, Quinn J M W, Moni M A. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 2021, 136: 104672. DOI:10.1016/j.compbiomed.2021.104672.

[30] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018, 31. DOI:10.48550/arXiv.1706.09516.

[31] Ramezani R, Iranmanesh S, Naeim A. Bench to bedside: AI and remote patient monitoring. Frontiers in Digital Health, 2025, 2: 1584443. DOI:10.3389/fdgth.2025.1584443.

[32] Gholampour S. Impact of nature of medical data on machine and deep learning for imbalanced datasets: clinical validity of SMOTE is questionable. Machine Learning and Knowledge Extraction, 2024, 6(2): 39. DOI:10.3390/make6020039.

[33] Deng Z, Torim A, Yahia S B. Generative AI in intrusion detection systems for Internet of Things: a systematic literature review. IEEE Open Journal of the Communications Society, 2025, 6: 145-163. DOI:10.1109/OJCOMS.2025.3573194.

[34] Newaz A, Ahmed N, Haq F. Survival prediction of heart failure patients using machine learning techniques. Informatics in Medicine Unlocked, 2021, 25: 100245. DOI:10.1016/j.imu.2021.100772.

[35] Nasef D, Basco K J, Singh A, et al. Clinical applicability of machine learning models for binary and multi-class electrocardiogram classification. AI, 2025, 6(3): 59. DOI:10.3390/ai6030059.