Evidence-Guided Feature Space Transformation for Heart Disease Classification

Adi Suwondo[1*], Kusrini[2], Ema Utami[3], Kumara Ari Yuana[4]

[1,2,3,4]Informatics Department, Faculty of Computer Science, Universitas Amikom Yogyakarta, Sleman, D.I. Yogyakarta, Indonesia
E-mail: [1*]adisuwondo@students.amikom.ac.id, [2]kusrini@amikom.ac.id, [3]ema.u@amikom.ac.id, [4]kumara.a@amikom.ac.id

(Received: 16 Sep 2026, revised: 29 Oct 2026, 13 Nov 2026, accepted: 18 Nov 2026)

Abstract

Feature relevance in clinical machine learning is commonly used for explanation and selection, while its implications for feature-space geometry require further investigation. This study evaluates an evidence-guided feature-space transformation combining class-discriminative distributional evidence and permutation-based model relevance through mean-one feature multipliers. The primary experiment used 303 Cleveland heart disease records with 14 predictors and five repetitions of stratified 10-fold cross-validation. Representation geometry and predictive performance were evaluated using four geometric indices and four classifiers. Fused transformation increased the mean Geometric Separation Ratio from 0.8466 to 1.1107. For Support Vector Classification, ROC-AUC increased from 0.8907 to 0.9066 and PR-AUC from 0.8943 to 0.9123; however, these differences were not statistically significant under dependence-corrected inference. Additional experiments examined conventional scaling, fixed fusion coefficients, and alternative evidence generators. External evaluation on 294 Hungarian records using 11 harmonized predictors showed descriptive ROC-AUC increases for three classifiers, although this required retraining the Cleveland models with a reduced feature set. These findings suggest that evidence-guided transformation modifies representation geometry, although predictive gains remain uncertain and classifier-dependent. Generalizability is limited by the small Cleveland development cohort and the reduced-feature external validation protocol.

Keywords: Explainable Artificial Intelligence, Feature Transformation, Feature Weighting, Heart Disease Classification, Feature-Space Geometry.

I. INTRODUCTION

Explainable artificial intelligence (XAI) has become increasingly important in healthcare because predictive performance alone does not reveal how individual clinical variables contribute to model behavior. Systematic reviews show that healthcare XAI is used to improve transparency, justification, learning from models, and trust, while explanation quality and effectiveness remain active evaluation challenges [1], [2]. Feature-level relevance is therefore often used to interpret predictions or to guide the construction of clinical machine-learning models. Model-agnostic explanation frameworks such as SHAP and LIME provide widely used feature-level attribution mechanisms [3], [4].

Permutation-based importance provides a model-oriented mechanism for quantifying feature relevance through controlled perturbation [5]. More recent cardiovascular studies have moved beyond post-hoc explanation by incorporating explainability into feature-selection pipelines. Elemam et al. used SHAP-based ranking followed by optimized wrapper selection for coronary artery disease diagnosis [6], while Cenitta et al. combined optimization-based feature selection with SHAP and LIME for ischemic heart disease prediction [7]. These approaches primarily convert relevance into a decision about which variables should be retained.

Feature weighting offers a different perspective because variables can remain in the representation while contributing unequally. De Amorim reviewed feature-weighted K-means methods that explicitly model unequal relevance among dimensions [8], and Panday et al. used feature weighting as a basis for unsupervised feature selection [9]. A recent experimental review further showed the breadth of feature-weighted fuzzy clustering and its evaluation using Silhouette, Davies-Bouldin, Friedman, and Holm procedures [10]. Galis et al. extended this idea by using SHAP-derived information for global feature weighting in unsupervised clustering [11]. In heart disease classification, recent work has also evaluated algorithm-sensitive feature weighting using Mutual Information and ReliefF [12].

DOI: 10.34148/teknika.v15i3.1592
TEKNIKA, Volume 15(3), November 2026, pp. 597-608
ISSN 2549-8037, EISSN 2549-8045

CC BY SA

Existing research establishes that feature relevance can support explanation, variable selection, and feature-weighted learning [6]–[13]. However, these applications do not necessarily distinguish changes in representation geometry from changes in predictive performance. Conventional feature scaling and diagonal feature weighting also modify the geometry of the input space; therefore, multiplying normalized relevance scores by the number of predictors should not, by itself, be regarded as a novel transformation algorithm. Unlike conventional standardization, which derives scaling factors from feature dispersion, the present approach derives feature-specific multipliers from class-discriminative and model-dependent relevance. Unlike learned diagonal metric methods, the proposed procedure does not directly optimize a distance-based objective or learn a metric from pairwise similarity constraints. Its contribution is therefore evaluated as an evidence-construction and representation-analysis framework rather than as a new metric-learning algorithm. Accordingly, the present study investigates an evidence-guided weighting framework in which distributional and model-derived relevance are converted into mean-one feature multipliers and evaluated jointly through geometric and classifier-level outcomes. The mean-one constraint provides a common multiplier normalization, allowing the effects of unequal feature-wise weighting to be examined without introducing an additional uniform shrinkage factor. The research gap is thus primarily methodological and evaluative: whether evidence-guided diagonal rescaling produces interpretable geometric changes, whether these changes correspond to predictive gains, and whether the observed behavior persists under alternative weighting configurations and external cohort evaluation. A structured comparison with the most relevant prior studies is presented in Table 1.

Table 1. State-of-the-art Positioning and Research Gap

Study	Paradigm	Evidence	Primary use	Geometry
Altmann et al. [5]	Permutation relevance	Model-derived	Importance / interpretation	No
de Amorim [8]	Feature weighting	Data structure	Weighted clustering	Clustering context
Panday et al. [9]	Weighting / selection	Cluster relevance	Feature selection	Clustering context
Golzari Oskouei et al. [10]	Feature-weighted fuzzy clustering	Data-driven	Clustering	Yes
Galis et al. [11]	XAI feature weighting	SHAP	Weighted clustering	Yes
Elemam et al. [6]	XAI-guided selection	SHAP	Feature selection	No explicit geometry
Prihantono et al. [12]	Heart-disease feature weighting	MI + ReliefF	Weighting / selection	No explicit geometry
Talukder et al. [13]	Explainable heart-disease prediction	XAI	Explanation/detection	No explicit geometry
This study	Evidence-guided transformation	Distributional + model-derived	Mean-one diagonal feature rescaling	Yes + downstream behavior

Conventional feature scaling applies predetermined transformations to adjust feature magnitudes, whereas feature selection reduces the predictor set. Supervised metric learning, in contrast, learns a distance or similarity function from labeled observations, commonly through an explicit optimization objective. Kulis [14] reviewed linear and nonlinear metric-learning approaches, while Weinberger and Saul [15] demonstrated the optimization of a Mahalanobis distance metric for large-margin nearest-neighbor classification. Unlike these approaches, the proposed evidence-guided transformation does not optimize a learned distance function or introduce a new metric-learning algorithm. Instead, it constructs a diagonal feature transformation from training-derived distributional and permutation-importance evidence, with mean-one normalization ensuring that the arithmetic mean of the feature multipliers equals one without preserving individual feature scales or pairwise distances. Accordingly, its contribution is evaluated through representation geometry, classifier-dependent predictive performance, and sensitivity analyses rather than through claims of superiority over supervised metric-learning methods.

This study addresses three research questions: RQ1, how does evidence-guided diagonal feature rescaling affect supervised class geometry; RQ2, do changes in geometric separation correspond to consistent predictive improvements across classifier families; and RQ3, how sensitive are the geometric and predictive outcomes to distributional evidence, model-derived evidence, fusion coefficients, and evidence-generator choices? The methodological contribution of this study is not the introduction of a new diagonal scaling operator, but the explicit integration of class-discriminative distributional evidence and permutation-derived model relevance into a reproducible feature-weighting procedure. The framework combines two normalized evidence sources through a fold-specific variance-based fusion heuristic and converts the resulting relevance profile into mean-one feature multipliers. Its empirical contribution is a paired evaluation of how this evidence construction changes supervised feature-space geometry and downstream classification, complemented by conventional scaling and ANOVA feature-selection baselines, fusion-coefficient and evidence-generator ablations, and evaluation on an independently collected cohort. These analyses distinguish measurable representation changes from predictive improvements and establish the limits of the proposed framework rather than assuming universal superiority. Rather than claiming a fundamentally new

CC BY SA

diagonal scaling operation, the study evaluates the conditions under which evidence-guided rescaling provides measurable benefits and identifies the limitations of geometry-based conclusions.

II. RESEARCH METHODOLOGY

A. Dataset and Experimental Protocol

The primary experiment used the Cleveland heart disease cohort comprising 303 observations and a binary disease-presence target [16]. The working representation consisted of 14 predictors: eight original attributes (sex, cp, fbs, restecg, exang, slope, ca, and thal) and six contextualized attributes (age_h, trestbps_cat, chol_ratio, thalach_ratio, hrmax_pred, and oldpeak_h). The six contextualized predictors were generated using deterministic patient-level rules. Specifically, age_h retained the recorded age; trestbps_cat categorized resting blood pressure into three levels (0 for values ≤120, 1 for values >120 and ≤140, and 2 for values >140); chol_ratio was calculated as total cholesterol divided by 200; hrmax_pred was computed as $208 - 0.7 \times$ age; thalach_ratio was defined as observed maximum heart rate divided by hrmax_pred; and oldpeak_h retained the recorded ST-depression value. These transformations did not estimate quantiles, thresholds, or other parameters from the full cohort. Missing-value imputation and subsequent scaling were performed separately within each training partition. These contextualized predictors were treated as upstream representations rather than novel components of the proposed transformation. Their implementation was audited against the raw Cleveland records and confirmed to use only individual patient measurements and prespecified constants, without estimating cohort-level statistics. Missing-value imputation, scaling, and evidence estimation were performed within each training partition, and the fitted preprocessing parameters and feature multipliers were applied unchanged to the corresponding held-out partition.

Internal evaluation employed five repetitions of stratified 10-fold cross-validation, yielding 50 paired train–test partitions. All feature representations and downstream classifiers were evaluated on identical partitions. The same fold-specific preprocessing and evidence-construction procedures were applied consistently across the four representations, ensuring that observed differences could be attributed to the representation configurations rather than different data partitions.

The repeated folds were used to estimate predictive variability and enable paired descriptive comparisons. Because observations and training sets overlap across repetitions, the 50 fold-level results were not assumed to be statistically independent. Dependence-aware inference was therefore conducted separately, as described in Section II.D.

Table 2 summarizes the experimental configuration. Additional experiments examined StandardScaler, fixed evidence-fusion coefficients, and Logistic Regression versus Random Forest evidence generators. External validation was performed using a separate Hungarian cohort under a harmonized feature protocol described in Section II.E.

Table 2. Experimental Configuration

Item	Configuration
Dataset	Cleveland heart disease
Instances / predictors	303 / 14
Validation	Repeated Stratified 10-fold x 5 repeats
Evidence reference estimator	Logistic Regression
Permutation importance	Recall scoring; 30 repeats
Representations	Original, Distributional-SP, XAI-SP, Fused-SP
Downstream classifiers	LR, SVC, MLP, Gaussian NB
Geometry metrics	GSR, Silhouette, Davies-Bouldin, Calinski-Harabasz
Predictive metrics	Accuracy, Recall, Precision, F1, ROC-AUC, PR-AUC
Statistics	Primary: corrected resampled t-test; exploratory: Friedman, paired Wilcoxon signed-rank, Holm correction, rank-biserial effect size, and Spearman correlation.

Table 3. Classifier Configurations Used in the Internal Experiments.

Classifier	Configuration
LR	max_iter=2000, class_weight='balanced', solver='lbfgs'
SVC	kernel='rbf', C=5, probability=False, class_weight='balanced', gamma='scale'
MLP	hidden_layer_sizes=(64,32), activation='relu', solver='adam', max_iter=1000, early_stopping=True
Gaussian NB	GaussianNB() with default parameters

The four downstream classifiers were evaluated using fixed configurations, as summarized in Table 3. Logistic Regression employed the lbfgs solver, L2 regularization, C = 1.0, balanced class weights, a convergence tolerance of 10^{-4}, and a maximum of 2,000 iterations. Support Vector Classification used an RBF kernel, C = 5, gamma = 'scale', balanced class weights, a tolerance of 10^{-3}, and probability=False. The Multilayer Perceptron comprised two hidden layers of 64 and 32 neurons, ReLU activation, the Adam optimizer, an initial learning rate of 0.001, an L2 penalty parameter of 0.0001, early stopping with a validation fraction of 0.1, a tolerance of 10^{-4}, 10 iterations without improvement, and a maximum of 1,000 iterations. Gaussian Naive Bayes used its default configuration, including var_smoothing = 10^{-9}. All experiments were conducted in Python 3.11.5 on Windows 10 using NumPy 1.24.4, Pandas 2.2.3, SciPy 1.14.1, and scikit-learn 1.5.2. A fixed random

DOI: 10.34148/teknika.v15i3.1592
TEKNIKA, Volume 15(3), November 2026, pp. 597-608
ISSN 2549-8037, EISSN 2549-8045

seed of 42 was used for reproducibility. The primary evaluation employed five repetitions of stratified 10-fold cross-validation, yielding 50 paired evaluation partitions. Permutation importance was computed using 30 repetitions per feature. Classifier hyperparameters were fixed for the reported comparisons, without nested hyperparameter optimization.

As an additional conventional feature-selection baseline, SelectKBest with the ANOVA F-score was evaluated using seven of the 14 predictors (k = 7). The number of selected predictors was fixed before evaluation, without optimizing k on held-out data. For each of the 50 repeated stratified cross-validation partitions, median imputation and Min–Max scaling were fitted on the training subset, followed by ANOVA feature ranking using training labels only. The fitted selector was then applied unchanged to the corresponding held-out subset. The four downstream classifiers retained the same configurations and evaluation partitions as the primary experiment. This baseline enables comparison between retaining all predictors with evidence-derived continuous weighting and selecting a fixed subset through a conventional univariate criterion.

B. Evidence Construction

Within each training fold, distributional evidence was quantified from the absolute standardized difference between the two outcome classes [17]. This quantity captures distributional class discrimination and should not be interpreted as expert-elicited guideline evidence. For feature j, the standardized effect magnitude and its normalized evidence score were defined in Equation 1:

$$d_j = \begin{cases} \dfrac{|\mu_{j,1} - \mu_{j,0}|}{S_{j,pool}}, & s_{j,pool} > 10^{-12} \\ 0, & s_{j,pool} \le 10^{-12} \end{cases} \quad (1)$$

$$s_{j,pool} = \sqrt{\frac{s_{j,0}^2 + s_{j,1}^2}{2}}$$

For feature *j*, the distributional effect magnitude was calculated as the absolute difference between the positive- and negative-class means divided by the root mean of their sample variances, as defined in Equation 1. Here, $\mu_{j,1}$ and $\mu_{j,0}$ denote the class-conditional means, while $s_{j,1}^2$ and $s_{j,0}^2$ denote the corresponding sample variances. The denominator $s_{j,pool}$ is the square root of the arithmetic mean of the two class-specific variances and is not weighted by class sample sizes. To avoid numerical instability, d_j is set to zero when $s_{j,pool} \le 10^{-12}$. All quantities are estimated exclusively from the training partition.

$$C_j = \begin{cases} \dfrac{d_j}{\sum_{k=1}^{p} d_k}, & \sum_{k=1}^{p} d_k > 10^{-12} \\ \dfrac{1}{p}, & \sum_{k=1}^{p} d_k \le 10^{-12} \end{cases} \quad (2)$$

Equation 2 normalizes the distributional effect magnitudes obtained from Equation 1 into a nonnegative evidence vector. Here, C_j denotes the normalized distributional evidence for feature *j*, d_j is its standardized class-discriminative magnitude, and *p* represents the number of predictors. When the sum of all effect magnitudes exceeds 10^{-12}, each magnitude is divided by their total. Otherwise, a uniform weight of *1/p* is assigned to every predictor. Consequently, the distributional evidence scores sum to one and remain well-defined even when no feature exhibits a sufficiently large standardized class difference.

Model-derived evidence was estimated using permutation importance from a fixed Logistic Regression reference estimator trained only on the current training fold. Recall was used as the scoring function and each feature was permuted 30 times. Negative mean importance was clipped to zero before normalization:

$$\mathrm{q_j} = \max(PI_j, 0) \quad (3)$$

$$S_j = \begin{cases} \dfrac{q_j}{\sum_{k=1}^{p} q_k}, & \sum_{k=1}^{p} q_k > 10^{-12} \\ \dfrac{1}{p}, & \sum_{k=1}^{p} q_k \le 10^{-12} \end{cases} \quad (4)$$

In Equation 3, PI_j denotes the mean recall-based permutation importance of feature *j*, estimated from 30 permutations within the training partition, and q_j represents its nonnegative value after clipping negative importance scores to zero. Equation 4 normalizes these values to obtain the model-derived evidence vector *S*, where S_j denotes the normalized evidence assigned to feature *j*, and *p* is the number of predictors. When the sum of the nonnegative importance values exceeds 10^{-12}, each q_j is divided by their total. Otherwise, a uniform weight of *1/p* is assigned to every feature to avoid division by zero. Consequently, all normalized model-derived evidence scores are nonnegative and sum to one. All importance estimates and normalization parameters are derived exclusively from the training partition and are not fitted using the corresponding test observations.

A fixed reference estimator was deliberately used to decouple evidence generation from downstream classifier evaluation, ensuring that all classifiers received the same transformed representation. Distributional and model-derived evidence were combined through the upstream weighting mechanism defined in Equations 5 and 6:

$$W_j = \alpha C_j + (1 - \alpha) S_j \quad (5)$$

$$\alpha = \begin{cases} \dfrac{Var(S)}{Var(S)+Var(C)}, & Var(S) + Var(C) > 10^{-12} \\ 0.5, & otherwise \end{cases} \quad (6)$$

Equations 5 and 6 define the adaptive fusion of distributional and model-derived evidence. In Equation 5, W_j denotes the fused evidence assigned to feature *j*, C_j and S_j represent the corresponding normalized distributional and model-derived evidence scores, and α determines their relative contributions. Because both evidence vectors are nonnegative and sum to one, their convex combination also produces a nonnegative, unit-sum evidence vector. In Equation 6, *Var(C)* and *Var(S)* denote the population variances calculated across the *p* feature-level scores within each normalized evidence

vector. The coefficient follows an inverse-variance-inspired heuristic: a more unevenly concentrated evidence vector receives a smaller fusion contribution, whereas a less concentrated vector receives a larger contribution. This rule is intended to moderate differences in evidence concentration rather than provide a statistically optimal estimate of feature relevance. If the sum of the two variances does not exceed 10^{-12}, equal weighting $\alpha = 0.5$ is applied to maintain numerical stability. The coefficient is recalculated independently within each training partition, without using the corresponding test observations or labels. Its empirical suitability is examined through fixed-coefficient sensitivity experiments rather than assumed a priori.

The reference estimator was fixed to ensure that the four downstream classifiers received the same transformed feature representation. Because Logistic Regression was also evaluated as a downstream classifier, a supplementary experiment replaced the reference estimator with Random Forest to assess sensitivity to evidence-generator choice.

C. Mean-one Feature-space Transformation

Direct multiplication by a normalized evidence vector would introduce global magnitude shrinkage because the mean normalized weight equals 1/p. To isolate unequal feature-wise rescaling from this global effect, each evidence vector was converted into a mean-preserving multiplier. For fused evidence:

$$m_j = pW_j \tag{7}$$

$$X'_{ij} = m_j X_{ij} \tag{8}$$

$$\frac{1}{p}\sum_{j=1}^{p} m_j = 1 \tag{9}$$

Equations 7–9 formalize the conversion of normalized evidence into mean-one feature multipliers. In Equation 7, m_j denotes the multiplier assigned to feature (j), (p) is the total number of predictors, and W_j is the normalized fused evidence obtained from Equation 5. Equation 8 applies this multiplier to the preprocessed value X_{ij} of feature j in observation i, yielding the transformed value X'_{ij}. Because the fused evidence vector satisfies $\sum_{j=1}^{p} W_j = 1$, substituting $m_j = pW_j$ yields $\frac{1}{p}\sum_{j=1}^{p} m_j = \sum_{j=1}^{p} W_j = 1$, as stated in Equation 9. The factor p therefore removes the uniform shrinkage in the average multiplier that would occur if unit-sum evidence weights were applied directly. It also preserves relative evidence ratios, since $\frac{m_j}{m_k} = \frac{W_j}{W_k}$ whenever $W_k > 0$. This normalization does not preserve individual feature scales, variances, pairwise distances, or the original geometry. Mathematically, the operation is a diagonal feature rescaling rather than a fundamentally new transformation operator. The methodological contribution lies in the evidence-construction and fusion procedure, its conversion into comparable mean-one multipliers, and the joint evaluation of representation geometry and downstream predictive behavior.

Distributional-SP and XAI-SP were obtained by applying the same mean-one multiplier construction to the distributional evidence vector C and the model-derived evidence vector S, respectively. Fused-SP used the combined evidence vector W, whereas Original retained the preprocessed predictors without evidence-derived weighting. No additional normalization was performed after applying the feature-specific multipliers, allowing changes in representation geometry to be evaluated without a subsequent scaling operation. Figure 1 illustrates the overall evidence-guided feature-space transformation workflow.

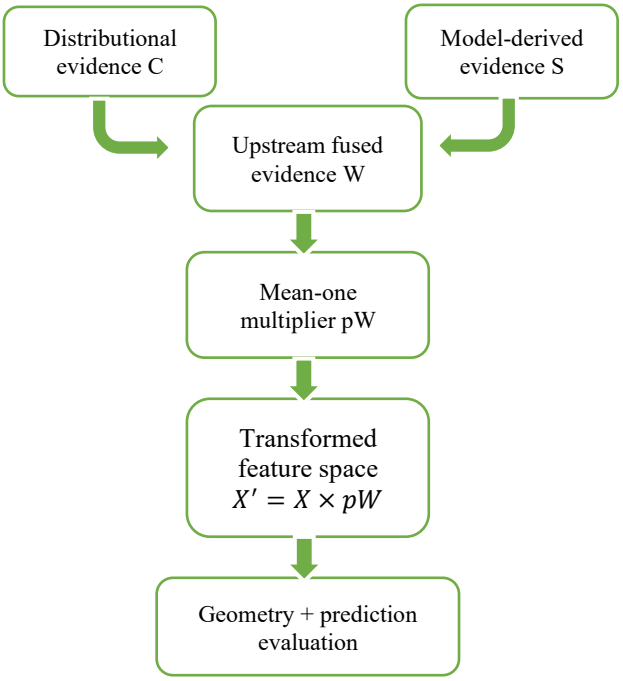


Figure 1. Evidence-guided Mean-one Feature-space Transformation Framework.

D. Geometry, Prediction, and Statistical Analysis

Representation geometry was evaluated using four complementary indices. The Geometric Separation Ratio (GSR) quantified the Euclidean separation between the two class centroids relative to their within-class dispersion. For each class $c \in \{0,1\}$, the centroid $\boldsymbol{\mu}_c$ was computed as the arithmetic mean of its transformed feature vectors, while the within-class dispersion D_c was calculated as the mean Euclidean distance of class-c observations from that centroid. GSR was then defined as the centroid distance divided by the arithmetic mean of the two class dispersions, with a numerical stabilization constant of 10^{-12} added to the denominator.

$$GSR = \frac{\|\mu_1 - \mu_0\|2}{\frac{D_0 + D_1}{2} + 10^{-12}} \tag{10}$$

As defined in Equation 10, larger GSR values indicate greater centroid separation relative to within-class spread. Because feature-specific rescaling changes Euclidean distances, GSR is not invariant under the proposed

DOI: 10.34148/teknika.v15i3.1592
TEKNIKA, Volume 15(3), November 2026, pp. 597-608
ISSN 2549-8037, EISSN 2549-8045

transformation and should be interpreted as a representation-level geometric measure rather than a direct indicator of predictive or clinical utility. Silhouette, Davies-Bouldin, and Calinski-Harabasz indices provided established complementary measures of separation and compactness [18]–[20]. Predictive behavior was evaluated using Logistic Regression (LR), Support Vector Classification (SVC) [21], Multilayer Perceptron (MLP), and Gaussian Naive Bayes (NB).

Predictive performance was evaluated using Accuracy, Recall, Precision, F1-score, ROC-AUC, and PR-AUC on the held-out partition of each cross-validation split. For each classifier–representation combination, fold-level means and sample standard deviations were calculated across the 50 paired partitions. The four geometry indices were likewise summarized using their fold-level means and standard deviations. Geometric and predictive outcomes were analyzed separately to avoid interpreting improved feature-space separation as direct evidence of classification or clinical benefit.

For the primary paired comparisons, an approximate corrected resampled t-test was applied to the fold-level performance differences [22]. With n = 50 paired partitions, the corrected standard error was calculated as

$$SE_{corr} = \sqrt{\left(\frac{1}{n} + \frac{n_{test}}{n_{train}}\right) s_d^2} \quad (11)$$

In Equation 11, n = 50 denotes the number of paired evaluation partitions, and s_d is the sample standard deviation of the paired metric differences. The ratio $\frac{n_{test}}{n_{train}}$ is approximately 1/9 for stratified 10-fold cross-validation, because each test partition contains approximately one-tenth of the observations and the corresponding training partition contains approximately nine-tenths. For each prespecified paired comparison, the test statistic was calculated as $t = \frac{\bar{d}}{SE_{corr}}$, where $\bar{d}$ denotes the mean paired difference. Approximate two-sided p-values and 95% confidence intervals were obtained using the Student t distribution with n − 1 = 49 degrees of freedom. The confidence interval was calculated as $\bar{d} \pm t_{0.975,49} SE_{corr}$.

Holm adjustment was applied within prespecified families of multiple comparisons to control the family-wise error rate [23]. Friedman and paired Wilcoxon signed-rank tests were retained as exploratory analyses of representation differences, consistent with established nonparametric comparison procedures [24]. Dependence arising from overlapping cross-validation partitions was addressed separately using the corrected resampled standard error. Because this correction is approximate, the resulting confidence intervals and significance tests were interpreted cautiously. The corrected resampled t-test was designated as the primary inferential procedure for prespecified paired comparisons of predictive performance. Holm-adjusted p-values and rank-biserial effect sizes were reported for the Wilcoxon comparisons, but their results were not treated as independent confirmatory evidence because the repeated cross-validation partitions overlap. Spearman rank correlation was additionally used to examine the fold-level association between changes in GSR and Recall [25]. The corresponding correlation coefficients and nominal p-values were interpreted as exploratory because fold-level dependence can affect statistical uncertainty. Conclusions were therefore based primarily on dependence-corrected comparisons, effect magnitudes, confidence intervals, and the consistency of observed results.

E. External Cohort Validation

External validation was conducted using the Hungarian Heart Disease cohort (n = 294), with the Cleveland cohort (n = 303) used exclusively for model development. The primary external evaluation employed 11 harmonized predictors: sex, cp, fbs, restecg, exang, age_h, trestbps_cat, chol_ratio, thalach_ratio, hrmax_pred, and oldpeak_h. The variables slope, ca, and thal were excluded from this evaluation because of their substantial missingness in the Hungarian cohort. The 11-feature representation was constructed consistently across the two cohorts. All preprocessing parameters, evidence weights, and classifier parameters requiring estimation were fitted using Cleveland observations only and subsequently applied to Hungarian observations without using Hungarian outcome labels for fitting or model selection. The classifiers were retrained on Cleveland using the reduced 11-feature representation; therefore, the external experiment evaluates transfer under feature harmonization rather than direct replication of the original 14-feature models.

Missing values were handled using median imputation, with the imputer fitted exclusively on the Cleveland development cohort. The fitted imputer was then applied unchanged to the Hungarian cohort. Subsequently, Min–Max scaling parameters were estimated from the imputed Cleveland predictors and applied to both cohorts without refitting on Hungarian observations. Evidence construction and classifier training were performed only after these preprocessing operations. Contextualized predictors were generated using the same prespecified deterministic rules in both cohorts.

Uncertainty in external performance differences was assessed using 300 paired patient-level bootstrap resamples of the Hungarian evaluation cohort. Original and Fused-SP predictions were compared on identical resampled patient indices, while the Cleveland-trained models remained fixed. Percentile-based 95% confidence intervals were calculated from the bootstrap distributions of paired metric differences. These intervals therefore quantify conditional evaluation-sample uncertainty rather than uncertainty from model retraining.

III. RESULTS AND DISCUSSION

A. Evidence Characteristics and Feature-Space Geometry

Across the 50 folds, the upstream fusion coefficient had a mean of 0.9300 (SD 0.0375; range 0.8106-0.9760), so fused

CC BY SA

evidence was predominantly aligned with the class-discriminative distributional component while retaining a smaller model-derived adjustment. The model-derived evidence distribution was more concentrated than the distributional evidence. For example, thal received a mean model-derived evidence score of 0.5355, compared with a distributional evidence score of 0.1344 and a fused evidence score of 0.1581. These values indicate differences in feature-relevance concentration rather than independently established clinical importance. The resulting geometric characteristics of the four representations are summarized in Table 4.

Table 4 summarizes the mean and sample standard deviation of four geometric indices across 50 paired cross-validation partitions. The standard deviations characterize the observed variation in representation geometry across evaluation partitions. Because the partitions overlap, these values should not be interpreted as uncertainty estimates obtained from 50 independent samples. The geometric indices are therefore used primarily for descriptive comparison rather than as evidence of clinical effectiveness.

Relative to Original, Fused-SP increased the mean Geometric Separation Ratio (GSR) from 0.8466 to 1.1107 and the Silhouette coefficient from 0.1147 to 0.1881. The Davies–Bouldin index decreased from 2.4550 to 1.9164, whereas the Calinski–Harabasz index increased from 5.0317 to 8.4549. XAI-SP achieved the strongest mean geometric separation, with GSR of 1.3087 and Silhouette of 0.2462, exceeding the corresponding Fused-SP values.

These observations demonstrate that evidence-guided feature rescaling substantially modifies the measured geometry of the Cleveland representation. However, the extremely small p-values obtained from conventional paired fold-level tests should not be interpreted as evidence from 50 independent replications, because repeated cross-validation partitions overlap. Accordingly, the geometric results are primarily interpreted descriptively, with fold-level variability reported to characterize their consistency. The comparative geometric patterns are illustrated in Figure 2.

Table 4. Feature-space Geometry Across Four Representations (Mean ± Sample SD Over 50 Paired Cross-Validation Partitions).

Representation	GSR ↑	Silhouette ↑	DB ↓	CH ↑
Original	0.8466 ± 0.1651	0.1147 ± 0.0450	2.4550 ± 0.4981	5.0317 ± 1.9587
Distributional-SP	1.0816 ± 0.2581	0.1811 ± 0.0686	1.9587 ± 0.4932	8.0621 ± 3.8634
XAI-SP	1.3087 ± 0.5607	0.2462 ± 0.1216	1.7822 ± 0.7337	10.6675 ± 9.0759
Fused-SP	1.1107 ± 0.2797	0.1881 ± 0.0723	1.9164 ± 0.4966	8.4549 ± 4.3477

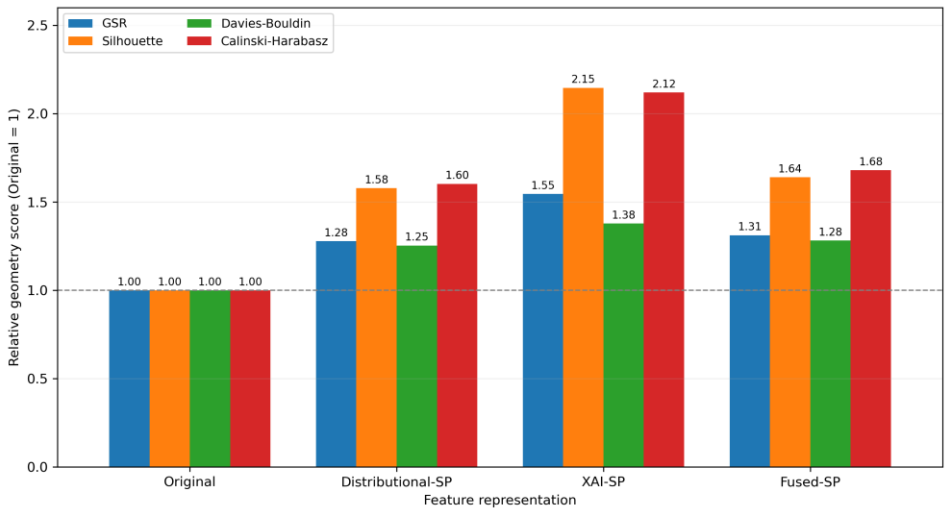


Figure 2. Relative Representation Geometry Across Original, Distributional-SP, XAI-SP, and Fused-SP. All Metrics Were Directionally Normalized Relative To Original So That Values Greater Than One Consistently Indicate a More Favorable Geometric Index.

XAI-SP produced the strongest geometric separation, with GSR 1.3087, Silhouette 0.2462, Davies-Bouldin 1.7822, and Calinski-Harabasz 10.6675. This is consistent with feature-weighting literature showing that unequal relevance can materially alter data geometry [8]–[11]. However, the predictive results below show why geometry should be treated as an intermediate representation outcome rather than a sufficient objective.

B. Classifier-dependent Predictive Effects

The mean predictive performance across the 50 paired evaluation folds is reported in Table 5.

DOI: 10.34148/teknika.v15i3.1592
TEKNIKA, Volume 15(3), November 2026, pp. 597-608
ISSN 2549-8037, EISSN 2549-8045

Table 5. Internal Predictive Performance of Four Classifiers Across Four Feature Representations (Mean ± Sample SD Over 50 Paired Cross-Validation Partitions)

Model	Representation	Accuracy	Recall	F1	ROC-AUC	PR-AUC
LR	Original	.8281±.0742	.7885±.1140	.8062±.0868	.9049±.0577	.9060±.0569
LR	Distributional-SP	.8328±.0716	.7927±.1135	.8113±.0841	.9018±.0584	.9039±.0588
LR	XAI-SP	.7846±.0762	.7612±.1202	.7622±.0893	.8580±.0686	.8404±.0870
LR	Fused-SP	.8328±.0729	.7899±.1145	.8108±.0855	.9014±.0587	.9039±.0588
SVC	Original	.8163±.0687	.7902±.1051	.7970±.0751	.8907±.0646	.8943±.0606
SVC	Distributional-SP	.8222±.0719	.7927±.1116	.8024±.0822	.9045±.0545	.9103±.0491
SVC	XAI-SP	.7866±.0627	.7781±.1062	.7684±.0717	.8506±.0700	.8308±.0825
SVC	Fused-SP	.8176±.0745	.7898±.1168	.7972±.0867	.9066±.0533	.9123±.0482
MLP	Original	.7874±.0664	.7423±.1405	.7570±.0916	.8527±.0631	.8561±.0696
MLP	Distributional-SP	.7814±.0654	.7626±.1513	.7552±.1008	.8559±.0718	.8575±.0717
MLP	XAI-SP	.7589±.0746	.7312±.1006	.7357±.0786	.7779±.0929	.7730±.1079
MLP	Fused-SP	.7820±.0723	.7735±.1239	.7626±.0868	.8527±.0718	.8598±.0679
NB	Original	.8349±.0720	.7957±.1119	.8138±.0843	.8968±.0616	.8939±.0644
NB	Distributional-SP	.8349±.0720	.7957±.1119	.8138±.0843	.8968±.0616	.8939±.0644
NB	XAI-SP	.7799±.0821	.7524±.1266	.7558±.0991	.8579±.0694	.8301±.0918
NB	Fused-SP	.8349±.0720	.7957±.1119	.8138±.0843	.8968±.0616	.8939±.0644

Note: Reported values are fold-level means and sample standard deviations. All classifier–representation combinations were evaluated on identical cross-validation partitions. Inferential comparisons were performed separately using dependence-corrected paired procedures; the standard deviations shown here are descriptive and do not assume independence among repeated partitions.

The same Fused-SP representation produced different responses across classifier families (Table 5). LR Accuracy increased from 0.8281 to 0.8328, whereas ROC-AUC decreased slightly from 0.9049 to 0.9014. For SVC, ROC-AUC increased from 0.8907 to 0.9066 (Δ = +0.0159), and PR-AUC increased from 0.8943 to 0.9123 (Δ = +0.0180). However, after accounting for dependence between repeated cross-validation partitions, neither difference was statistically significant. The corrected ROC-AUC comparison yielded p = 0.1069 with a 95% confidence interval of −0.0035 to 0.0353; the corresponding PR-AUC comparison yielded p = 0.1355 with a 95% confidence interval of −0.0058 to 0.0419.

MLP Recall increased descriptively from 0.7423 to 0.7735, but this change did not establish statistically significant superiority. Gaussian NB remained unchanged between Original and Fused-SP for the reported predictive metrics. These findings indicate that the predictive consequences of feature-space transformation depend on the downstream classifier and cannot be inferred solely from improved geometric separation. Figure 3 presents the paired mean performance differences between Fused-SP and Original across the 50 repeated cross-validation partitions. For SVC, Fused-SP increased ROC-AUC by 0.0159 and PR-AUC by 0.0180, whereas Recall changed marginally by −0.0004. For MLP, Recall increased by 0.0312, while ROC-AUC remained virtually unchanged. These differences are descriptive and should not be interpreted as statistically significant improvements without reference to the dependence-corrected inferential results.

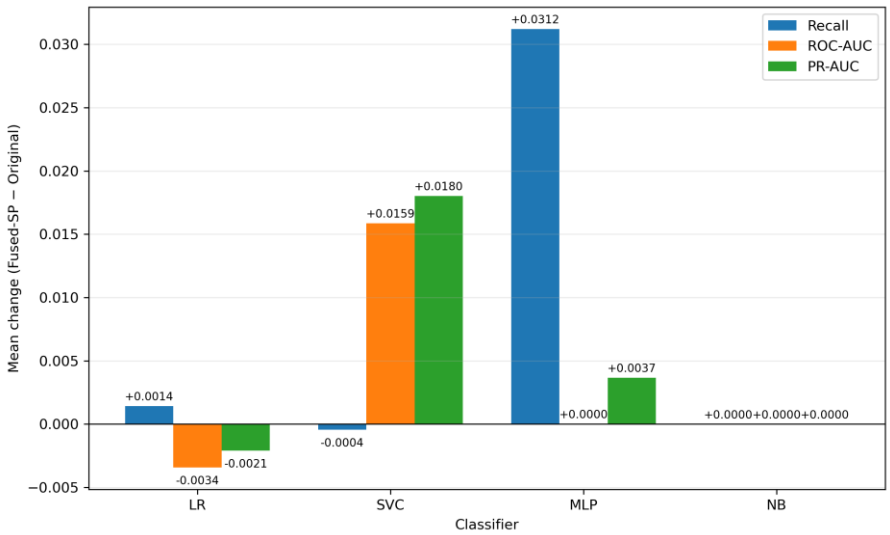


Figure 3. Mean Predictive Changes for Fused-SP Relative to Original Across Classifier Families

Table 6. Internal ROC-AUC Comparison of Conventional Scaling and Fused-SP

Classifier	Original	StandardScaler	Fused-SP
LR	0.9049	0.9018	0.9014
SVC	0.8907	0.8839	0.9066
MLP	0.8527	0.8692	0.8527
NB	0.8968	0.8968	0.8968

To assess whether evidence-guided weighting offered benefits beyond conventional preprocessing, Fused-SP was compared with the unweighted Original representation and StandardScaler under the same internal cross-validation design (Table 6). Fused-SP achieved the highest mean ROC-AUC for SVC (0.9066), compared with 0.8907 for Original and 0.8839 for StandardScaler. Nevertheless, the Fused-SP versus StandardScaler SVC difference did not reach statistical significance under corrected resampled inference ($p = 0.1192$).

The pattern was not uniform across classifiers. StandardScaler achieved the highest mean ROC-AUC for MLP (0.8692), while Original performed slightly better for LR (0.9049). Gaussian NB was unchanged. These findings show that evidence-guided weighting should be considered a classifier-dependent representation strategy rather than a universally superior alternative to conventional scaling.

To further examine the proposed transformation against conventional feature selection, ANOVA SelectKBest ($k = 7$) was evaluated on the same 50 paired cross-validation partitions. Table 7 presents the corresponding ROC-AUC results and dependence-corrected paired comparisons against Fused-SP.

Table 7. ROC-AUC Comparison Between ANOVA SelectKBest and Fused-SP (Mean ± Sample SD Over 50 Paired Partitions)

Model	ANOVA (k=7)	Fused-SP	Δ ROC-AUC	p-value
LR	0.8976 ± 0.0608	0.9014 ± 0.0587	+0.0038	0.5277
SVC	0.8896 ± 0.0705	0.9066 ± 0.0533	+0.0169	0.2099
MLP	0.8598 ± 0.0864	0.8527 ± 0.0718	−0.0070	0.8124
NB	0.8850 ± 0.0691	0.8968 ± 0.0616	+0.0118	0.1740

Fused-SP achieved higher mean ROC-AUC than ANOVA SelectKBest for Logistic Regression (0.9014 versus 0.8976), SVC (0.9066 versus 0.8896), and Gaussian Naive Bayes (0.8968 versus 0.8850). In contrast, ANOVA SelectKBest achieved higher mean ROC-AUC for MLP (0.8598 versus 0.8527). The largest positive difference favoring Fused-SP occurred for SVC ($\Delta = +0.0169$), but the dependence-corrected comparison was not statistically significant ($p = 0.2099$; 95% CI: −0.0099 to 0.0437). None of the four classifier comparisons reached the 0.05 significance threshold. These findings indicate that evidence-guided continuous weighting can yield competitive discrimination relative to univariate feature selection, but do not establish statistically significant superiority.

Table 8. SVC ROC-AUC Under Fixed and Adaptive Fusion Coefficients

α	Mean ROC-AUC
0.00	0.8506
0.25	0.8870
0.50	0.8941
0.75	0.9026
1.00	0.9045
Adaptive	0.9066

The fixed-coefficient ablation demonstrated that SVC performance was sensitive to the relative contribution of distributional and model-derived evidence (Table 8). Mean ROC-AUC increased from 0.8506 at $\alpha = 0$, corresponding to model-derived evidence alone, to 0.9045 at $\alpha = 1$, corresponding to distributional evidence alone. Adaptive fusion achieved a mean ROC-AUC of 0.9066.

Although the adaptive configuration produced the highest observed mean ROC-AUC, its difference from fixed $\alpha = 1$ was small ($\Delta = +0.0021$) and was not statistically significant under corrected resampled inference ($p = 0.4338$). Thus, the adaptive coefficient should not be presented as conclusively superior to distributional-only weighting. Instead, the ablation indicates that the balance between evidence sources influences downstream performance, particularly for SVC.

To examine the sensitivity of the proposed framework to the evidence reference estimator, Logistic Regression and Random Forest generators were compared under the same paired evaluation protocol. The resulting predictive outcomes are summarized in Table 9.

Table 9. Effect of Evidence-Generator Choice on Fused-SP

Classifier	Metric	LR Based	RF Based	RF−LR
SVC	ROC-AUC	0.9066	0.9002	−0.0064
SVC	PR-AUC	0.9123	0.9075	−0.0048
MLP	Recall	0.7735	0.7820	+0.0085
MLP	ROC-AUC	0.8527	0.8621	+0.0093
MLP	PR-AUC	0.8598	0.8670	+0.0071

The original framework used Logistic Regression as a fixed reference estimator for permutation-based evidence generation. Because Logistic Regression was also included among the downstream classifiers, a supplementary experiment replaced the evidence generator with Random Forest while preserving the paired evaluation protocol.

The adaptive fusion coefficient differed substantially between generators, averaging 0.9300 for the Logistic Regression reference and 0.5837 for the Random Forest reference. For SVC, the Random Forest generator reduced mean ROC-AUC from 0.9066 to 0.9002. Conversely, MLP

DOI: 10.34148/teknika.v15i3.1592
TEKNIKA, Volume 15(3), November 2026, pp. 597-608
ISSN 2549-8037, EISSN 2549-8045

mean ROC-AUC increased from 0.8527 to 0.8621, and Recall increased from 0.7735 to 0.7820.

None of the 12 paired generator comparisons remained statistically significant after dependence correction and Holm adjustment. These results indicate that the evidence generator influences the weighting profile and descriptive predictive outcomes, but the current experiment does not establish superiority of either reference estimator. They also address the potential special relationship between Logistic Regression evidence generation and downstream Logistic Regression evaluation.

C. Geometry Does Not Equate to Predictive Utility

The ablation comparison highlighted the distinction between geometric separation and predictive utility. Although feature weighting altered the geometric indices, the corresponding predictive responses varied across classifier families and weighting configurations. Consequently, improved geometric separation cannot be assumed to produce improved classification performance. XAI-SP produced the highest GSR (1.3087), yet LR Recall decreased from 0.7885 to 0.7612 and LR PR-AUC decreased from 0.9060 to 0.8404. Similar predictive degradation occurred for XAI-SP in other classifiers despite its stronger geometry. In contrast, Fused-SP produced a more moderate geometric change while preserving or improving selected predictive metrics. The fold-level associations between changes in geometric separation and Recall are summarized in Table 10.

Table 10. Fold-level Relationship Between Geometry Change and Recall Change

Classifier	Spearman rho (ρ)	p-value	Interpretation
LR	0.015	0.917	Non-significant
SVC	0.175	0.224	Non-significant
MLP	0.271	0.057	Non-significant
NB	—	—	Recall invariant

Fold-level correlation analysis reinforced this distinction. Delta GSR was not significantly associated with delta Recall for LR ($\rho = 0.015$, $p = 0.917$) or SVC ($\rho = 0.175$, $p = 0.224$). MLP showed a positive but non-significant tendency ($\rho = 0.271$, $p = 0.057$), while NB Recall was invariant. Accordingly, the data do not support a causal or universal claim that greater geometric separation produces greater predictive sensitivity.

These ablation results address RQ3 by showing that distributional-only, model-derived-only, and adaptively fused evidence produced different geometric and predictive outcomes. XAI-SP achieved the highest mean geometric separation, whereas Fused-SP provided a more moderate geometric transformation with classifier-dependent predictive effects. The sensitivity analyses did not establish that adaptive fusion was consistently superior to fixed fusion coefficients or alternative evidence generators.

The external evaluation compared Original and Fused-SP using the 11-feature harmonized representation on the Hungarian cohort. The classifier-specific ROC-AUC results are presented in Table 11.

Table 11. External Evaluation on Hungarian Using 11 Harmonized Features

Classifier	Original ROC-AUC	Fused-SP ROC-AUC	Δ
LR	0.8803	0.8843	+0.0041
SVC	0.8659	0.8758	+0.0099
MLP	0.8392	0.8603	+0.0211
NB	0.8511	0.8511	0.0000

External evaluation was conducted on 294 Hungarian observations using models developed exclusively on the 303-record Cleveland cohort. Because of differences in predictor availability, the primary external experiment employed 11 harmonized features, with all models retrained on Cleveland using this reduced representation.

Fused-SP achieved higher descriptive ROC-AUC values than Original for LR (0.8843 versus 0.8803), SVC (0.8758 versus 0.8659), and MLP (0.8603 versus 0.8392), while Gaussian NB remained unchanged at 0.8511. SVC PR-AUC increased from 0.7999 to 0.8198, and MLP PR-AUC increased from 0.7568 to 0.8009.

Paired patient-level bootstrap analyses were additionally performed to quantify uncertainty in the external performance differences. For SVC, the ROC-AUC difference was +0.0099 (95% bootstrap CI: −0.0074 to 0.0269), whereas the PR-AUC difference was +0.0199 (95% bootstrap CI: 0.0025 to 0.0443). For MLP, the Recall difference was +0.1226 (95% bootstrap CI: 0.0573 to 0.1899), and the ROC-AUC difference was +0.0211 (95% bootstrap CI: 0.0061 to 0.0395). These intervals are conditional on the fitted Cleveland models and do not incorporate uncertainty arising from model development or feature harmonization.

These findings provide an initial assessment of transfer to an independent cohort. However, the external experiment used a harmonized feature representation that differs from the original 14-feature internal evaluation. Therefore, the results should not be interpreted as direct external replication of the original models. The Hungarian cohort also exhibited substantial missingness in several original predictors, limiting the reliability of the supplementary 14-feature sensitivity analysis.

D. Implications, Limitations, and Relation to Prior Work

The findings extend prior feature-weighting research by evaluating geometry as an explicit intermediate outcome in supervised clinical classification. Galis et al. demonstrated that SHAP-derived weighting can improve unsupervised clustering [11], while recent cardiovascular XAI studies primarily use relevance for selection or explanation [6], [13]. The present results show that continuous evidence weighting can reorganize supervised clinical feature space without removing variables, but the representation must still be judged by downstream behavior. This is particularly important because maximizing geometric criteria alone would have favored XAI-SP, which was not the best predictive representation.

The mean-one multiplier formulation is also important methodologically. Since normalized weights sum to one, direct weighting would have an average multiplier of 1/p and would globally shrink the input. Multiplication by p retains all pairwise weight ratios while restoring the mean multiplier to one. Meaningful geometry changes remained after this correction, indicating that the observed effects arose from unequal feature-wise rescaling rather than uniform magnitude reduction.

Several limitations constrain the interpretation of these findings. First, the internal experiment was based on 303 Cleveland observations, and repeated cross-validation partitions were statistically dependent. Although corrected resampled inference was applied, its confidence intervals and significance tests remain approximate. Second, the additional Hungarian cohort provided an independent evaluation population, but substantial missingness required an 11-feature harmonized protocol and retraining of the development models. Consequently, the external results cannot be interpreted as direct validation of the original 14-feature classifiers.

Third, the distributional evidence component represents observed class-conditional differences rather than clinical knowledge independently elicited from domain experts. Fourth, the weighting profiles and adaptive coefficients were sensitive to the choice of evidence reference estimator. The diagonal rescaling mechanism also does not explicitly model nonlinear feature interactions. Finally, improvements in geometric separation and discrimination metrics do not establish better calibration, clinical decision utility, or patient outcomes. Larger, more diverse cohorts and prospective clinical evaluation are required before clinical applicability can be established.

An additional methodological limitation concerns the distinction between evidence-guided weighting and established feature transformation techniques. The mean-one multiplier does not constitute a new class of transformation operators; it is a normalized diagonal rescaling whose distinctive aspect lies in the construction and fusion of the evidence vectors. Furthermore, the observed geometric improvements are specific to the chosen distance-based indices and do not imply improved calibration, clinical decision-making, or patient outcomes. The present findings should therefore be interpreted as an evaluation of representation behavior under a defined experimental protocol rather than proof of general clinical utility.

Although conventional scaling and ANOVA-based univariate feature selection were included as supplementary baselines, the experiments did not cover wrapper-based selection, embedded selection, or learned diagonal metric methods. The resulting comparisons therefore do not establish superiority over the broader family of feature-weighting and metric-learning approaches.

IV. CONCLUSION

This study investigated evidence-guided feature-space transformation as an approach to modifying supervised representation geometry without eliminating predictors. The proposed framework combined distributional class-discriminative evidence and model-derived permutation relevance through mean-one feature multipliers, enabling systematic evaluation of geometric and predictive effects.

In the Cleveland cohort, Fused-SP increased mean GSR from 0.8466 to 1.1107 and Silhouette from 0.1147 to 0.1881, while XAI-SP produced the strongest geometric separation. Nevertheless, these geometric changes did not consistently translate into better predictive performance. For SVC, Fused-SP increased mean ROC-AUC from 0.8907 to 0.9066 and PR-AUC from 0.8943 to 0.9123, but neither difference was statistically significant under dependence-corrected inference. Logistic Regression showed marginal changes, MLP exhibited a descriptive Recall increase, and Gaussian Naive Bayes remained unchanged. Sensitivity analyses showed that performance depended on the fusion coefficient and evidence reference estimator, without establishing consistent superiority of adaptive fusion or either evidence generator. External evaluation using the Hungarian cohort and an 11-feature harmonized representation showed descriptive discrimination improvements for selected classifiers; however, this protocol required retraining the Cleveland models and was not directly equivalent to the original 14-feature experiment. Overall, the findings support evaluating evidence-guided feature transformations through both representation geometry and downstream predictive behavior. They do not establish universal predictive superiority or clinical utility. Future work should investigate larger cohorts, harmonized external validation, calibration, and alternative evidence-generation strategies.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this manuscript, the authors used OpenAI ChatGPT to assist with language refinement, manuscript organization, methodological discussion, and code development support. All computational procedures, experimental results, statistical analyses, and scientific interpretations were reviewed and verified by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

[1] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, 'Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011--2022)', *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107161, 2022. doi: 10.1016/j.cmpb.2022.107161.

[2] J. Jung, H. Lee, H. Jung, and H. Kim, 'Essential Properties and Explanation Effectiveness of Explainable

Artificial Intelligence in Healthcare: A Systematic Review', *Heliyon*, vol. 9, no. 5, p. e16110, 2023. doi: 10.1016/j.heliyon.2023.e16110.

[3] S. M. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in neural information processing systems*, 2017, vol. 30, pp. 4765–4774.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why Should I Trust You?" Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD Conference*, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[5] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, 'Permutation Importance: A Corrected Feature Importance Measure', *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010. doi: 10.1093/bioinformatics/btq134.

[6] T. Elemam, H. Refaat, and M. Makhlouf, 'An Explainable AI-Driven Hybrid Feature Selection Approach for Coronary Artery Disease Diagnosis', *Scientific Reports*, vol. 16, p. 10411, 2026. doi: 10.1038/s41598-026-41712-y.

[7] D. Cenitta, N. Arul, R. V Arjunan, K. Chadaga, and J. Andrew, 'An Explainable Artificial Intelligence Framework for Ischemic Heart Disease Prediction Using Enhanced Squirrel Search Feature Selection', *Scientific Reports*, vol. 16, p. 15422, 2026. doi: 10.1038/s41598-026-46823-0.

[8] R. C. de Amorim, 'A Survey on Feature Weighting Based K-Means Algorithms', *Journal of Classification*, vol. 33, pp. 210–242, 2016. doi: 10.1007/s00357-016-9208-4.

[9] D. Panday, R. C. de Amorim, and P. Lane, 'Feature Weighting as a Tool for Unsupervised Feature Selection', *Information Processing Letters*, vol. 129, pp. 44–52, 2018. doi: 10.1016/j.ipl.2017.09.005.

[10] A. Golzari Oskouei *et al.*, 'Feature-Weighted Fuzzy Clustering Methods: An Experimental Review', *Neurocomputing*, vol. 619, p. 129176, 2025. doi: 10.1016/j.neucom.2024.129176.

[11] F. Galis, D. M. Onchis, and C. Istin, 'Refining Filter Global Feature Weighting for Fully Unsupervised Clustering', *Applied Sciences*, vol. 15, no. 16, p. 9072, 2025. doi: 10.3390/app15169072.

[12] D. Prihantono, U. Faqih, and I. Indra, 'Algorithm-Sensitive Feature Weighting Using Mutual Information and ReliefF for Heart Disease Classification', *Edumatic: Jurnal Pendidikan Informatika*, vol. 10, no. 1, pp. 260–269, 2026. doi: 10.29408/edumatic.v10i1.34262.

[13] M. A. Talukder, A. S. Talaat, M. Kazi, and A. Khraisat, 'XAI-HD: an explainable artificial intelligence framework for heart disease detection', *Artificial Intelligence Review*, vol. 58, p. 385, 2025. doi: 10.1007/s10462-025-11385-6.

[14] B. Kulis, 'Metric Learning: A Survey', *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013. doi: 10.1561/2200000019.

[15] K. Q. Weinberger and L. K. Saul, 'Distance Metric Learning for Large Margin Nearest Neighbor Classification', *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[16] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, 'Heart Disease', UCI Machine Learning Repository, 1989. doi: 10.24432/C52P4X.

[17] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[18] P. J. Rousseeuw, 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis', *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[19] D. L. Davies and D. W. Bouldin, 'A Cluster Separation Measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

[20] T. Caliński and J. Harabasz, 'A Dendrite Method for Cluster Analysis', *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. doi: 10.1080/03610927408827101.

[21] C. J. C. Burges, 'A Tutorial on Support Vector Machines for Pattern Recognition', *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998. doi: 10.1023/A:1009715923555.

[22] C. Nadeau and Y. Bengio, 'Inference for the Generalization Error', *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003. doi: 10.1023/A:1024068626366.

[23] S. Holm, 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. doi: 10.2307/4615733.

[24] J. Demšar, 'Statistical Comparisons of Classifiers over Multiple Data Sets', *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.

[25] R. V Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*, 8th ed. Boston, MA, USA: Pearson, 2019.