

Article

A Multi-Criteria Sample Selection Framework Using Uncertainty, Reliability, Representativeness, and Non-Redundancy for Emergency Department Prediction

Daun Jeong ^{1,2} , SangJun Moon ³ and Jae Yong Yu ^{4,5,6,7,8,*}

- ¹ Division of Critical Care Medicine, Department of Emergency Medicine, Chung-Ang University Gwangmyeong Hospital, Gwangmyeong-si 14353, Gyeonggi-do, Republic of Korea; daun8910@gmail.com
- ² Department of Emergency Medicine, Chung-Ang University College of Medicine, Seoul 06794, Republic of Korea
- ³ Data Science Department, Hana Card Co., Ltd. 66, Eulji-ro, Jung-gu, Seoul 04538, Republic of Korea; msj5589@gmail.com
- ⁴ Research Institute for Data Science and AI (Artificial Intelligence), Hallym University, Chuncheon-si 24252, Gangwon-do, Republic of Korea
- ⁵ Division of Data Science, Hallym University, Chuncheon-si 24252, Gangwon-do, Republic of Korea
- ⁶ Healthcare AI Assurance Lab, Global Frontier Research Center, Hallym University, Chuncheon-si 24252, Gangwon-do, Republic of Korea
- ⁷ Department of Emergency Medicine, Hallym University, Chuncheon-si 24252, Gangwon-do, Republic of Korea
- ⁸ Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, MA 01655, USA
- * Correspondence: icalust@hallym.ac.kr; Tel.: +82-1046171449

Abstract

Background: Selecting informative training samples is a fundamental yet challenging problem in predictive modeling, particularly in heterogeneous clinical data. Although supervised learning is typically formulated as an optimization problem over model parameters, the composition of the training set substantially influences generalization performance. In this study, we propose a multi-criteria score-based sample selection framework for a machine learning setting. **Method:** Four sample-level scores were defined to quantify predictive uncertainty, representativeness, non-redundancy, and reliability. These scores were normalized and combined using three integration schemes: additive weighting, reliability-gated weighting, and rank-based aggregation. For each chief complaint category, a baseline model was trained either on the full training set, on score-selected subsets and on random size-matched subsets. Performance was assessed using the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve, sensitivity, and specificity, with classification thresholds determined by the Youden index. **Results:** Across experiments, integrated score-based subset selection outperformed both full-data training and random subsampling in terms of mean AUROC, while often showing lower variability across chief complaints. **Conclusions:** The results suggest that sample utility in clinical tabular data is intrinsically multi-dimensional and that explicitly modeling this structure can improve predictive discrimination.

Keywords: sample selection; predictive modeling; emergency department; uncertainty; representativeness; redundancy; reliability

MSC: 68T09; 62H30



Academic Editors: Lin Ge, Shengxin Tu, Brani Vidakovic and MinJae Lee

Received: 29 March 2026

Revised: 4 June 2026

Accepted: 22 June 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In standard supervised learning, one usually fits a predictor on the full empirical dataset, implicitly assuming that more observations necessarily lead to better generalization [1,2]. This assumption is not mathematically guaranteed when the data contain redundancy, local duplication, or weakly informative observations [3,4]. Accordingly, predictive learning may be viewed not only as parameter estimation, but also as sample selection over the empirical distribution.

Recent work in data-centric AI has emphasized that predictive performance depends not only on model choice, but also on the composition and informational quality of the training data [5,6]. In parallel, subset selection, coreset construction, and data pruning studies have shown that carefully selected subsets can preserve or even improve generalization relative to indiscriminate use of the full dataset [7,8]. Related ideas have also been explored in active learning, although that setting primarily concerns label acquisition rather than structured subset construction from an already labeled dataset.

This perspective is consistent with previous mathematical studies. Active learning studies when selectively chosen samples improve learning efficiency over indiscriminate sampling [9,10]. Core-set methods show that a geometrically chosen subset can remain competitive with the full dataset [11]. Submodular subset selection formulates training-set design as a set-function optimization problem with controlled loss relative to full-data training [12]. Redundancy-aware learning further shows that suppressing information can improve sample utility beyond relevance-only selection.

However, most existing approaches optimize only one notion of sample utility at a time. Uncertainty-based methods prioritize unstable samples, geometric methods emphasize coverage, and redundancy-aware methods emphasize dispersion. Some hybrid methods combine uncertainty, representativeness, and diversity, but these have been developed primarily for active learning or sample acquisition, where the goal is to decide which instances should receive labels next, rather than to explicitly construct an informative subset from an already labeled heterogeneous dataset [13,14]. Thus, a gap remains between multi-criteria notions of subset utility and practical row-selection frameworks for fixed labeled data.

Formally, let

$$D = \{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^p, y_i \in \{0, 1\}$$

and consider a selection operator

$$S : D \rightarrow S \subset D$$

where each observation receives a utility score s_i , the observations are ranked by s_i , and the predictor is trained only on the top-ranked fraction. The question is whether one can construct s_i so that training on S outperforms both full-data training and random size-matched subsampling.

In this study, we propose an integrated multi-criteria sample-selection framework for category-stratified emergency prediction. For each training observation, we define four normalized sample-utility scores representing predictive uncertainty, local representativeness, non-redundancy, and reliability. These scores are combined through additive, reliability-gated, and rank-based aggregation schemes to generate prioritized training subsets from already labeled clinical data. The key contribution of this framework is that it formulates sample selection as a multi-criterion training-set construction problem, rather than as a single-criterion filtering procedure. We then evaluate whether the resulting structured subsets improve prediction performance relative to full-data training and random size-matched subsampling across chief complaint categories.

The remainder of this paper is structured as follows. Section 2 describes the study design, data construction, sample-selection framework, and evaluation strategy. Section 3 reports the main predictive performance results and robustness analyses. Section 4 discusses the implications and limitations of the findings, and Section 5 presents the conclusion.

2. Materials and Methods

2.1. Study Design and Analytical Overview

We conducted a prediction study using emergency department records stratified by chief complaint category. For each category, a binary prediction task was defined using structured triage variables and demographic features. Within each train–test split, we computed observation-level scores that characterize uncertainty, representativeness, redundancy, and reliability. Based on these scores, we generated reduced training subsets and compared their predictive performance against full-data training and random size-matched subsampling. The overall workflow is summarized in Figure 1 as follows:

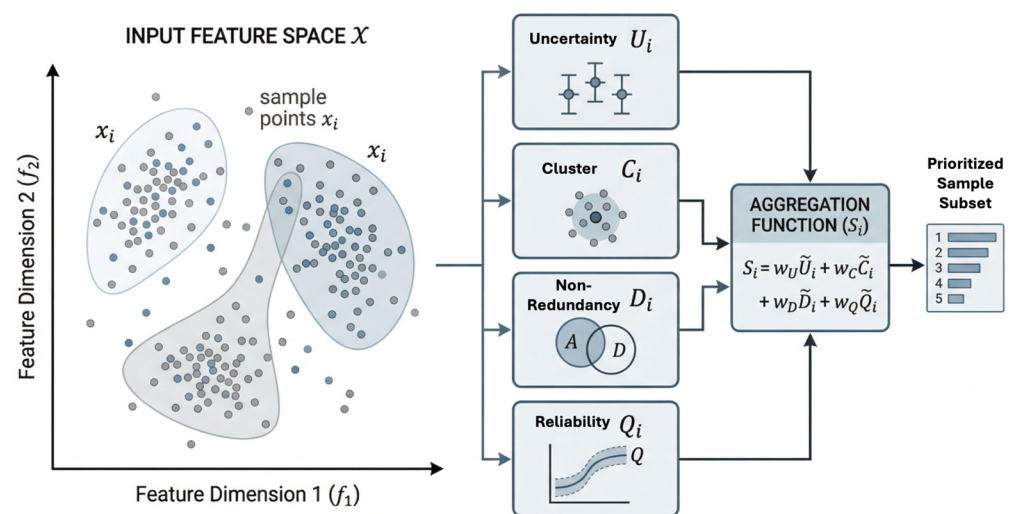


Figure 1. Schematic overview of the proposed multi-criteria row-selection framework. Each training sample is evaluated in the input feature space using four component scores: uncertainty, cluster representativeness, non-redundancy, and reliability. These normalized scores are then combined through an aggregation function to produce an integrated priority score, which is used to construct a prioritized training subset.

The primary evaluation metric was the area under the receiver operating characteristic curve (AUROC). Secondary metrics included the area under the precision–recall curve (AUPRC), sensitivity, and specificity, where the classification threshold was selected by the Youden index on the training set.

2.2. Data Construction and Outcome Definition

Emergency department data provide a natural testbed for this problem because they are large and structurally heterogeneous [15,16]. The MIMIC-IV-ED benchmark, derived from MIMIC-IV, is a publicly available dataset of ED visits that includes patient demographics, triage information, vital signs, laboratory measurements, and clinical outcomes recorded during ED stays. It defines standardized prediction tasks, enabling consistent and reproducible evaluation of predictive models in emergency care settings. The dataset also captures the temporal progression of patient states, allowing models to leverage time-dependent clinical information.

Let $x_i = (x_{i1}, \dots, x_{ip})^\top$ denote the predictor vector for patient i . The predictors consisted of age, sex, systolic blood pressure, diastolic blood pressure, heart rate, res-

piratory rate, body temperature, and oxygen saturation measured at the emergency department triage.

For each patient encounter, we defined a binary outcome

$$y_i = \begin{cases} 1 & \text{indicates death with in } \tau \text{ days} \\ 0 & \text{Otherwise} \end{cases}$$

where $\tau = 30$ in the primary analysis. Thus, the prediction problem is a binary classification task over the dataset $D_c = \{(x_i, y_i)\}_{i=1}^{n_c}$, for each chief complaint category c .

To reduce obvious data corruption, we removed observations with missing predictor values and applied broad physiological bounds.

Each original chief complaint was searched for the presence of predefined complaint-related strings, and binary indicators were assigned accordingly [17]. Individual complaint indicators were then further aggregated into 10 broader organ system- or syndrome-oriented categories to facilitate analysis and improve interpretability. The final 10 grouped chief complaint categories were as follows: Cardiovascular, Respiratory, Neurologic, Gastrointestinal, Genitourinary & Obstetric/Gynecology, Musculoskeletal, Trauma, Infectious & Dermatological, Psychiatric, and Metabolic. Each grouped category was defined as the logical union of one or more binary complaint indicators derived from the keyword dictionary. Full details of the complaint dictionary and keyword-to-category mapping are provided in github [17].

2.3. Train-Test Splitting

For each chief complaint category c , the cleaned dataset D_c was first partitioned into a training set D_c^{train} and a held-out test set D_c^{test} using stratified random sampling with an 80:20 ratio. Within each split, D_c^{train} was further divided into an inner training set and a validation set. Sample-level selection scores were computed exclusively from the inner training set, and candidate subset-selection configurations were selected using validation AUROC, with AUPRC used as a secondary criterion. The final classifier was then refitted using the selected configuration and evaluated only once on D_c^{test} , which was reserved solely for out-of-sample evaluation. All experiments were repeated over 10 independent stratified splits generated with different random seeds.

SMOTE was applied only to the training data to mitigate class imbalance. For each sampling ratio condition of 0.3, 0.5, and 0.7, the minority class was oversampled to achieve a minority-to-majority ratio of 0.3. The test set was left unchanged to prevent information leakage and preserve the original outcome prevalence for evaluation.

2.4. Baseline Prediction Model

As the base learner, we used logistic regression. For a training subset S the model is

$$Pr(Y_i = 1|x_i) = \pi_i = \frac{1}{1 + \exp(-(\beta_0 + x_i^\top \beta))} \quad (1)$$

The parameter vector (β_0, β) was estimated by maximum likelihood on S . Predicted probabilities on the test set were then used to compute AUROC and AUPRC.

For threshold-based metrics, let $\hat{\pi}_i^{train}$ denote the fitted probabilities on the training subset. The threshold t^* was selected by maximizing the Youden index:

$$t^* = \underset{t}{\operatorname{argmax}} \{ \text{Sensitivity}(t) + \text{Specificity}(t) - 1 \}.$$

This threshold was then applied to the test predictions to compute sensitivity and specificity.

2.5. Sample-Level Scoring Framework

For each observation i in the training set, we defined four sample-wise scores. These were designed to quantify complementary aspects of the utility of observation i for downstream predictive modeling.

2.5.1. Uncertainty Score

To measure predictive instability, we generated B bootstrap resamples of the training set. For each bootstrap iteration $b \in \{1, \dots, B\}$, a logistic model was fitted and applied back to all original training observations, yielding a predicted probability $\hat{\pi}_i^{(b)}$. The uncertainty score was defined as the standard deviation of bootstrap predictions [18]:

$$U_i = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\pi}_i^{(b)} - \bar{\pi}_i \right)^2}, \bar{\pi}_i = \frac{1}{B} \sum_{b=1}^B \hat{\pi}_i^{(b)}. \quad (2)$$

A larger U_i indicates that the predicted probability for observation i is more unstable under resampling-based perturbation of the training set.

We also computed an entropy-based uncertainty quantity

$$H_i = -\bar{\pi}_i \log(\bar{\pi}_i + \varepsilon) - (1 - \bar{\pi}_i) \log(1 - \bar{\pi}_i + \varepsilon) \quad (3)$$

2.5.2. Cluster Representativeness Score

Let z_i denote the standardized predictor vector for observation i . We applied k -means clustering of the standardized training data, obtaining cluster centers $\{\mu_1, \dots, \mu_k\}$. If observation i belongs to cluster $c(i)$, its Euclidean distance to the assigned cluster center is

$$d_i^{\text{clust}} = |z_i - \mu_{c(i)}|_2 \quad (4)$$

The cluster representativeness score was defined as $C_i = \frac{1}{1 + d_i^{\text{clust}}}$.

Thus, observations closer to the center of their local cluster receive larger scores and are considered more representative of their surrounding data structure [13].

2.5.3. Non-Redundancy Score

To reduce local redundancy, we computed the nearest-neighbor distance in the standardized predictor space. Let d_i^{NN} denote the Euclidean distance from z_i to its nearest non-identical neighbor [4]. The non-redundancy score was defined as $D_i = d_i^{\text{NN}}$.

A larger D_i implies that the observation lies farther from its nearest local neighbor and therefore contributes less redundant information.

2.5.4. Reliability Score

Reliability was designed to combine physiological plausibility and multivariate centrality. For each predictor, we defined an indicator of physiologically plausible range:

$$I_{ij} = \begin{cases} 1, & \text{if predictor } j \text{ lies in a clinically plausible range} \\ 0, & \text{Otherwise} \end{cases}$$

Physiological plausibility was assessed using predefined clinically acceptable ranges for each vital sign and demographic variable. Specifically, age was considered plausible if it ranged from 18 to 120 years, systolic blood pressure (SBP) from 40 to 300 mmHg, diastolic blood pressure (DBP) from 20 to 200 mmHg, heart rate (HR) from 20 to 250 beats/min, respiratory rate (RR) from 5 to 80 breaths/min, body temperature (BT) from 30 to 43 °C, and oxygen saturation (SpO2) from 50 to 100%. These bounds were intentionally permissive,

allowing the reliability score to further distinguish plausible from implausible observations within the retained sample.

The plausibility count for observation i was then $P_i = \sum_{j=1}^m I_{ij}$, where m is the number of plausibility-checked variables. To quantify multivariate centrality, we computed the Mahalanobis distance $M_i = (x_i - \bar{x})^\top \Sigma^{-1} (x_i - \bar{x})$, where $\bar{x}$ and Σ are the sample mean vector and covariance matrix of the numerical predictors in the training set. Since a smaller Mahalanobis distance indicates greater centrality, we transformed it into a normalized centrality score:

$$\tilde{M}_i = \frac{\max_j M_j - M_i}{\max_j M_j - \min_j M_j + \varepsilon} \quad (5)$$

The final reliability score was defined as

$$Q_i = \frac{1}{2} \cdot \frac{P_i}{m} + \frac{1}{2} \tilde{M}_i \quad (6)$$

Hence, Q_i is large when the observation is both physiologically plausible and centrally located in the multivariate predictor distribution.

2.6. Score Normalization and Subset Selection

Because the four scores are defined on different scales, each score was normalized to the unit interval [3]:

$$\tilde{U}_i = \frac{U_i - \min_j U_j}{\max_j U_j - \min_j U_j + \varepsilon} \quad (7)$$

ε is a fixed small constant added to the denominator to prevent division by zero when the empirical range of the score is zero or extremely small. Similarly for $\tilde{C}_i$, $\tilde{D}_i$, and $\tilde{Q}_i$. Given a scalar selection score S_i , the selected subset was defined as the top ρ fraction of the ranked training observations: $S_\rho = \{i \in D_c^{\text{train}} : S_i \geq q_{1-\rho}(S)\}$, where $q_{1-\rho}(S)$ is the $(1 - \rho)$ -quantile of the score distribution. We first considered four single-score subsets using $S_i \in \{\tilde{U}_i, \tilde{C}_i, \tilde{D}_i, \tilde{Q}_i\}$. The selected subset consisted of the 30% of training observations ranked by the relevant selection score. For sensitivity analysis, additional experiments were conducted using sampling ratios of 0.3, 0.5, and 0.7. Subset construction was based on a rank-based quantile threshold rather than an absolute score cutoff, ensuring a comparable retention ratio across methods and complaint categories.

2.7. Integrated Scoring Schemes

To combine the four normalized scores into a unified selection criterion, we considered three integration schemes.

2.7.1. Additive Weighted Score

The additive integrated score was defined as

$$S_i^{\text{add}} = w_U \tilde{U}_i + w_C \tilde{C}_i + w_D \tilde{D}_i + w_Q \tilde{Q}_i \quad (8)$$

where $w_U, w_C, w_D, w_Q \geq 0$, $w_U + w_C + w_D + w_Q = 1$.

2.7.2. Reliability-Gated Score

To emphasize the moderating role of reliability, we defined a gated score,

$$S_i^{gate} = \left(w_U \tilde{U}_i + w_C \tilde{C}_i + w_D \tilde{D}_i \right) \tilde{Q}_i^g + w_Q \tilde{Q}_i \quad (9)$$

where $g > 0$ is a gate exponent. This formulation gives greater weight to uncertainty, representativeness, and non-redundancy when reliability is high, while suppressing their contribution when reliability is low.

2.7.3. Rank-Based Aggregation

Let $R_U(i)$, $R_C(i)$, $R_D(i)$, and $R_Q(i)$ denote reversed normalized ranks of the four component scores, each scaled to $[0, 1]$, with larger values indicating better rank positions. The rank-based integrated score was defined as

$$S_i^{rank} = w_U R_U(i) + w_C R_C(i) + w_D R_D(i) + w_Q R_Q(i) \quad (10)$$

This scheme reduces sensitivity to score magnitude and instead emphasizes relative ordering.

2.8. Weight Sensitivity Analysis and Best Integrated Selection

For each integrated scheme, we evaluated a grid of admissible weights (w_U, w_C, w_D, w_Q) , subject to the simplex constraint, $w_U + w_C + w_D + w_Q = 1$. For the gated scheme, multiple gate exponents, g were also examined, which is, $\gamma \in \{0.5, 1, 2\}$. For each candidate parameter combination, a subset was selected, a logistic model was fitted, and predictive performance was evaluated on the corresponding test set [9]. Within each split, we identified the best additive, gated, rank-based and overall integrated configuration across all three schemes, according to the highest AUROC, with AUPRC used as a secondary criterion.

2.9. Robustness Analysis for Non-Linear Baseline

To assess whether the observed effect of structured row selection depended on the choice of a linear base learner, we additionally repeated the comparative experiments using a random forest classifier implemented with the ranger package as a representative non-linear ensemble model. For this robustness analysis, the random forest was trained using 100 trees. The same train–test splitting protocol, subset-construction procedure, and evaluation criteria were applied as in the primary analysis.

2.10. Statistical Interpretation

The proposed framework does not claim that smaller training sets are universally preferable. Rather, it tests the hypothesis that when informative samples are identified through a mathematically constructed utility score, a reduced subset may approximate a more efficient empirical distribution for classifier fitting. The study therefore evaluates whether structured sample selection can improve out-of-sample discrimination relative to both unrestricted full-data fitting and unstructured random subsampling. To evaluate the differences in AUROC between each selection method and full-data training for logistic regression and random forest separately, linear mixed-effects models, adjusting for sampling ratio, category were utilized. p -values were adjusted using the Benjamini–Hochberg procedure.

2.11. Reproducibility

For reproducibility, the following implementation details were used. Predictive uncertainty was estimated using 30 bootstrap iterations. Diversity was quantified using k-means clustering with 20 clusters. Local non-redundancy was assessed using nearest-neighbor distances, computed with $k = 2$, from which the first nearest-neighbor distance

was used as the non-redundancy score. Distributional representativeness was evaluated using Mahalanobis distance, computed from the covariance structure of the training data. All component scores were normalized to a common 0–1 scale before integration to ensure comparability across criteria. For gated integration, gate-power values of 0.5, 1, and 2 were evaluated. For weighted integration, the full weight grid consisted of all combinations of uncertainty, clustering, diversity, and quality weights ranging from 0.1 to 0.7 in increments of 0.1, constrained to sum to 1.0.

3. Results

3.1. Data Structure and Category-Wise Class Imbalance

Table 1 summarizes the baseline characteristics of the study cohort after preprocessing and variable harmonization. The final analytic dataset comprised 425,043 patients, including 9451 mortality cases, corresponding to an overall event rate of 2.22%. Patients in the mortality group were markedly older than those in the non-mortality group (median age, 76.0 vs. 53.0 years), and they also showed systematic differences across multiple vital signs, including higher HR, lower SpO₂, and lower SBP and DBP. These patterns suggest that the outcome was associated with meaningful shifts in the predictor distribution. At the same time, the low overall event rate indicates substantial class imbalance, implying that model performance should be evaluated using discrimination metrics that are robust to imbalance, particularly AUROC and AUPRC, rather than accuracy alone. The demographic characteristics according to other chief complaint categories are presented in Supplementary Table S1.

Table 1. Baseline characteristics of the study cohort according to 30-day mortality status after preprocessing and variable harmonization.

	Non-Mortality (<i>n</i> = 415,592)	Mortality (<i>n</i> = 9451)	<i>p</i> -Value
Age, Median [IQR]	53.0 [34.0; 68.0]	76.0 [64.0; 86.0]	<0.001
Gender, <i>n</i> (%)			<0.001
Female	225,442 (54.2%)	4437 (46.9%)	
Male	190,150 (45.8%)	5014 (53.1%)	
Vital sign, Median [IQR]			
body temperature (BT)	36.7 [36.4; 37.0]	36.7 [36.3; 36.9]	<0.001
Heart rate (HR)	84.0 [72.0; 96.0]	89.0 [76.0; 104.0]	<0.001
Respiratory rate (RR)	18.0 [16.0; 18.0]	18.0 [16.0; 20.0]	<0.001
Oxygen saturation (SpO ₂)	99.0 [97.0; 100.0]	98.0 [95.0; 99.0]	<0.001
Systolic blood pressure (SBP)	133.0 [120.0; 148.0]	123.0 [107.0; 141.0]	<0.001
Diastolic blood pressure (DBP)	77.0 [68.0; 87.0]	69.0 [58.0; 80.0]	<0.001

Values are presented as median [IQR] for continuous variables and *n* (%) for categorical variables. *p*-values were calculated using Mann–Whitney U test for continuous variable and chi-square test for categorical variables to compare the non-mortality and mortality groups. SpO₂, peripheral oxygen saturation; IQR, interquartile range.

Table 2, Figures S1 and S2 show that the prediction problem was highly heterogeneous across chief complaint categories. In the categories shown, the number of patients ranged from 14,566 in Gynecology to 78,926 in Gastrointestinal, while outcome prevalence ranged from 0.4% in Psychiatric to 4.0% in Respiratory presentations. Baseline model performance also varied across categories, with AUROC values ranging from 0.780 to 0.888 and AUPRC values ranging from 0.027 to 0.141. These findings indicate substantial category-level heterogeneity in sample size, class imbalance, and baseline predictive separability, suggesting that the learning problem differs meaningfully across complaint groups. Therefore, the effect of sample selection should be interpreted not as a uniform adjustment applied to a single population, but as a category-specific redistribution of the training sample space.

Table 2. Category-specific sample size, event prevalence, and baseline predictive performance across chief complaint groups in logistic regression.

Chief Complaint	Number of Patients	Prevalence	AUROC (SD)	AUPRC (SD)	Sensitivity	Specificity
Gastrointestinal	78,926	0.016	0.840 (0.007)	0.091 (0.011)	0.776	0.737
Musculoskeletal	73,211	0.008	0.872 (0.014)	0.074 (0.020)	0.804	0.760
Trauma	45,128	0.016	0.850 (0.012)	0.097 (0.018)	0.788	0.735
Metabolic	43,586	0.03	0.803 (0.011)	0.110 (0.013)	0.752	0.694
Cardiovascular	40,583	0.014	0.845 (0.018)	0.092 (0.021)	0.758	0.771
Neurologic	33,995	0.021	0.855 (0.018)	0.129 (0.020)	0.769	0.763
Respiratory	30,882	0.04	0.804 (0.014)	0.141 (0.016)	0.747	0.700
Psychiatric	26,921	0.004	0.780 (0.058)	0.027 (0.020)	0.660	0.701
Infectious & Dermatological	24,164	0.013	0.849 (0.021)	0.088 (0.021)	0.749	0.766
Genitourinary & Obstetric/ Gynecology	14,566	0.005	0.888 (0.059)	0.062 (0.037)	0.785	0.839

Prevalence denotes the proportion of mortality cases within each chief complaint category. AUROC and AUPRC are reported as mean (SD) across repeated experiments. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; SD, standard deviation.

As an additional robustness analysis, we repeated the comparative row-selection experiments using a random forest classifier. The main qualitative findings were preserved under this non-linear model setting. In particular, the best overall integrated strategy achieved a mean AUROC of 0.812, compared with 0.790 for full-data training and 0.735 for random size-matched subsampling.

Supplementary Table S4 showed that, compared with full-data training, integrated selection methods achieved significantly higher AUROC in both logistic regression and random forest models. The largest gains were observed for the best overall integrated approach, with AUROC increases of 0.0104 for logistic regression and 0.0268 for random forest, both with $p < 0.001$. Detailed category-specific results of the supplementary random forest analysis are provided in Supplementary Tables S2 and S3 and Figures S3–S5.

3.2. Overall Comparison of Row-Selection Operators

Table 3 summarizes the overall predictive performance of the row-selection methods across repeated experiments. The integrated row-selection strategies demonstrated the strongest discrimination performance among the evaluated approaches. Gated integration achieved the highest mean AUROC of 0.849, followed closely by rank integration and additive integration, both of which achieved a mean AUROC of 0.848. By comparison, the full-data baseline achieved a mean AUROC of 0.840. For AUPRC, rank integration and additive integration yielded the highest mean value of 0.094, whereas gated integration, uncertainty-only selection, and the full-data baseline each achieved a mean AUPRC of 0.093. These results indicate that multi-criterion integration, particularly the gated approach, provided the greatest improvement in AUROC-based discrimination, whereas gains in precision–recall performance were modest and largely comparable across the best-performing methods.

Selective retention of training instances based on structured scoring can perform comparably to, and in some cases better than, training on the full dataset. In particular, the superior AUROC of the integrated methods relative to the full-data baseline implies that appropriate subset construction may improve the geometric separability of the empirical training distribution. This supports the view that row selection can act not merely as a reduction in sample size, but as a transformation of the effective learning distribution.

Table 3. Comparison of aggregate discrimination performance across row-selection operators in logistic regression.

Method	AUROC (SD)	AUPRC (SD)	Sensitivity	Specificity
Rank integration	0.848 (0.039)	0.094 (0.039)	0.763	0.761
Gated integration	0.849 (0.040)	0.093 (0.036)	0.769	0.757
Additive integration	0.848 (0.040)	0.094 (0.037)	0.773	0.755
Full data	0.840 (0.038)	0.093 (0.037)	0.803	0.723
Uncertainty only	0.838 (0.039)	0.093 (0.037)	0.599	0.853
Non-redundancy only	0.833 (0.046)	0.091 (0.039)	0.690	0.793
Cluster only	0.830 (0.045)	0.086 (0.036)	0.828	0.673
Reliability only	0.829 (0.046)	0.084 (0.036)	0.836	0.663
Random subsampling	0.820 (0.049)	0.087 (0.040)	0.748	0.732

Performance measures are reported as mean (SD) across experiments. Rank integration, Gated integration, and Additive integration denote the best-performing integrated row-selection strategies under the corresponding aggregation schemes. Full data indicates model training on the entire available training set without row selection. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; SD, standard deviation.

The single-criterion methods showed more variable behavior than the integrated approaches. Uncertainty-only selection achieved a mean AUPRC of 0.093, comparable to the full-data baseline and gated integration, and showed the highest specificity of 0.853, but this was accompanied by substantially reduced sensitivity of 0.599, indicating a conservative operating profile. Non-redundancy-only selection produced relatively balanced threshold-dependent performance, with a sensitivity of 0.658 and specificity of 0.821, although its AUROC of 0.843 remained below that of the best-performing integrated methods. In contrast, cluster-only and reliability-only selection yielded lower AUROC values of 0.803 and 0.799, respectively. Random subsampling showed inferior overall discrimination, with an AUROC of 0.784, confirming that the gains observed with structured row selection cannot be explained by sample-size reduction alone.

A similar contrast was observed in threshold-dependent measures. The full-data model achieved the highest sensitivity of 0.803, but with a lower specificity of 0.723, whereas several row-selection methods, particularly uncertainty-only and non-redundancy-only selection, shifted the operating point toward higher specificity. By comparison, the integrated methods achieved more balanced operating characteristics, with sensitivities ranging from 0.763 to 0.773 and specificities ranging from 0.755 to 0.761. Among these, gated integration attained the highest mean AUROC of 0.849, whereas rank and additive integration showed the highest mean AUPRC of 0.094. Overall, these findings suggest that integrated row selection improved discrimination without inducing the pronounced sensitivity–specificity imbalance observed in some single-criterion approaches.

3.3. Category-Wise Performance Landscape

Figure 2 shows the mean AUROC for each category and method, highlighting substantial heterogeneity across chief complaint groups. Overall, the integrated methods were more often positioned near the top of the category-specific AUROC range, whereas single-criterion methods showed larger dispersion and more frequent performance deterioration. The gap between methods was relatively small in categories with high baseline separability but became more pronounced in more difficult categories. These findings suggest that the effectiveness of row selection was category-dependent and that multi-criteria integration provided a more stable strategy than reliance on any single score alone.

Logistic regression showed broadly consistent category-level AUROC patterns across sampling ratios of 0.3, 0.5, and 0.7, with integrated selection methods generally performing comparably to or better than single-criterion and random subsampling approaches.

Random forest showed greater variability across categories and sampling ratios, but integrated selection methods generally maintained stronger AUROC performance than random subsampling and most single-criterion approaches. Detailed category-specific results are shown in Figures S6 and S7.

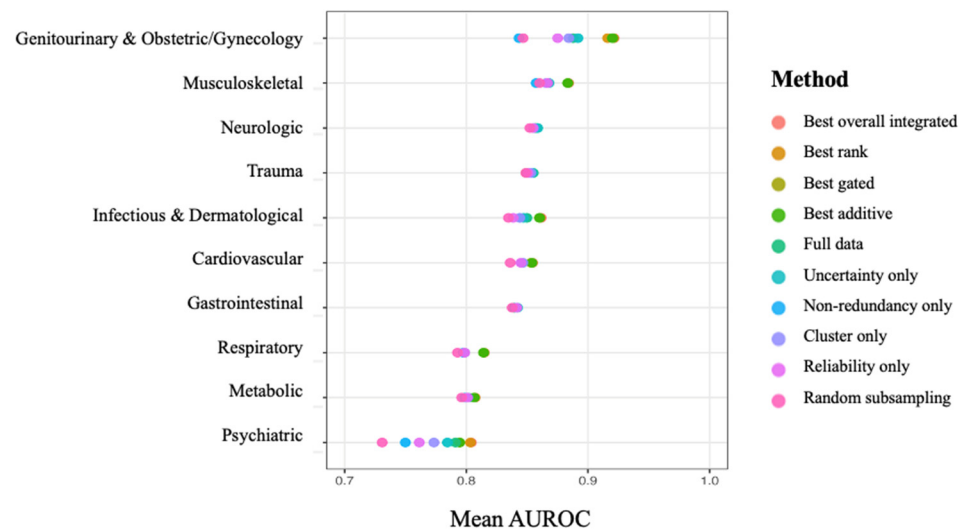


Figure 2. Mean AUROC by chief complaint category and row-selection method across repeated experiments in logistic regression. Points represent the mean AUROC for each method within each chief complaint category, aggregated across repeated experiments. Integrated methods include rank-based, gated, and additive multi-criteria row-selection strategies, whereas single-criterion methods are based on one score only. Random subsampling is shown as a reference method.

For logistic regression with SMOTE, category-level AUROC patterns were generally stable across sampling ratios of 0.3, 0.5, and 0.7, with integrated methods consistently clustering among the higher-performing approaches. For random forest with SMOTE, integrated methods also showed relatively strong performance across sampling ratios, whereas several single-criterion approaches and random subsampling showed greater variability across categories. The detailed results of the SMOTE-based analyses and sampling-ratio sensitivity analyses for each selection method are presented in Table S5, Figures S8 and S9.

3.4. Heterogeneity of the Gain Function

Figure 3 shows the category-specific AUROC change in the best-performing integrated row-selection method compared with full-data training. The effect varied across chief complaint categories, with the largest improvement observed in the Genitourinary & Obstetric/Gynecology category, where AUROC increased by approximately 0.038. More modest positive changes were observed in several other categories, whereas Trauma showed a negative change. In addition, several confidence intervals overlapped zero, suggesting that the effect of integrated row selection was heterogeneous and not uniformly positive across categories. These findings indicate that the benefit of integrated row selection depended on the statistical structure of each category-specific learning problem rather than acting as a uniform improvement across all strata.

From a methodological perspective, these results suggest that integrated row selection acts as a category-dependent filtering operator on the empirical training distribution. In categories with a greater proportion of redundant, weakly informative, or noisy observations, the benefit of selective retention appears to be larger. By contrast, in categories where the original training distribution was already relatively well aligned with the prediction task, the improvement was more limited. Thus, the advantage of the integrated strategy should

be interpreted not as a constant upward shift in model performance, but as an adaptive gain shaped by the local geometry of the category-specific sample space.

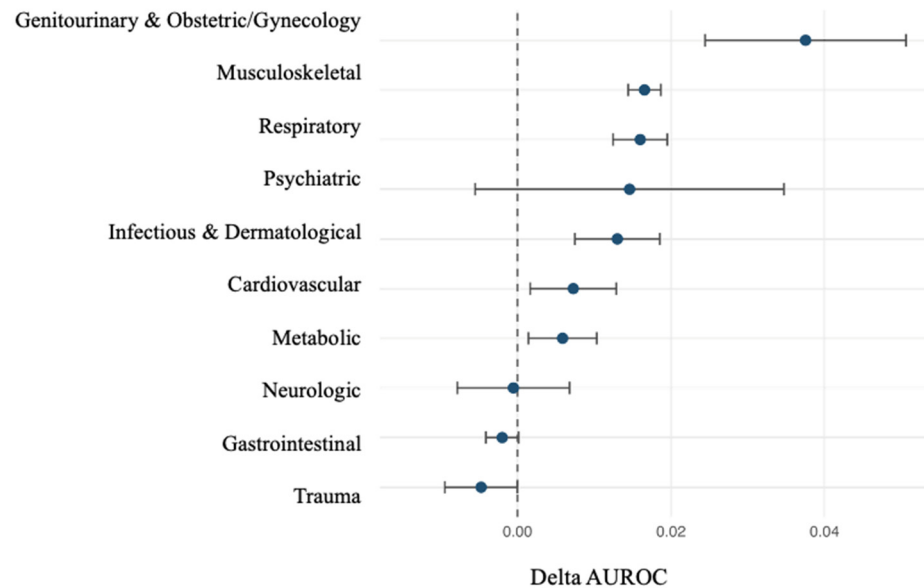


Figure 3. Category-specific AUROC improvement of the best overall integrated method relative to full-data training in logistic regression. Note: Bars represent the difference in mean AUROC between the best overall integrated method and the full-data baseline within each chief complaint category. Positive values indicate improved discrimination under integrated row selection.

3.5. Weight Sensitivity of Integrated Scoring Rules

Figure 4 examines the sensitivity of predictive performance to the weight parameters of the integrated score. Across all three aggregation families, additive, gated, and rank-based, the clearest monotone tendencies were observed for the uncertainty and non-redundancy weights. Increasing the weight assigned to uncertainty (w_U) generally decreased mean AUROC, whereas increasing the non-redundancy weight (w_D) generally increased mean AUROC. The reliability weight (w_Q) showed only a modest positive trend, while the representativeness weight (w_C) had a relatively weak and inconsistent effect across aggregation families.

These trends indicate that, within the explored weight space, the performance surface was not isotropic. Instead, the empirical response favored larger contributions from non-redundancy-related criteria, whereas excessive emphasis on uncertainty tended to reduce discrimination. The effects of reliability and cluster-based representativeness were comparatively modest, suggesting that these components had less dominant influence on AUROC than the uncertainty and non-redundancy weights.

Importantly, the three aggregation families showed broadly similar directional behavior, particularly for the uncertainty and non-redundancy weights. This suggests that the main performance pattern was driven more by the relative weighting of the component criteria than by a single algebraic form of integration. In other words, additive, gated, and rank-based aggregation differed in formal construction, but they recovered broadly comparable response patterns over the explored weight space.

3.6. Effect of Gate Power in Gated Integration

Figure 5 further analyzes the gated formulation by varying the gate power parameter. Across all four weights, the AUROC curves corresponding to gate powers 0.5, 1, and 2 were closely aligned. The direction of the weight effects remained largely unchanged: increasing w_U was associated with reduced AUROC, increasing w_C and w_D was associated

with improved AUROC, and w_Q exerted only a minor effect, with a slight negative trend at higher values.

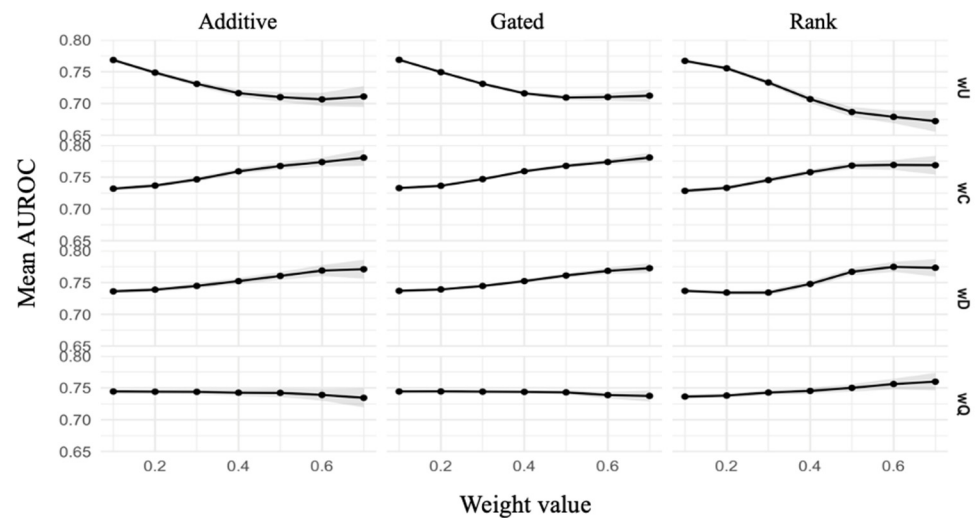


Figure 4. Mean AUROC in logistic regression as a function of criterion-specific weight values under additive, gated, and rank-based score integration. For each row, the weight of one criterion (w_U , w_C , w_D , or w_Q) was varied from 0.1 to 0.7 while the other settings were kept fixed. The plots illustrate the sensitivity of predictive performance to the relative contribution of each criterion. Shaded regions represent variability across repeated evaluations. Note: AUROC, area under the receiver operating characteristic curve. w_U , w_C , w_D , and w_Q indicate the criterion-specific weights in the multi-criteria sample selection framework.

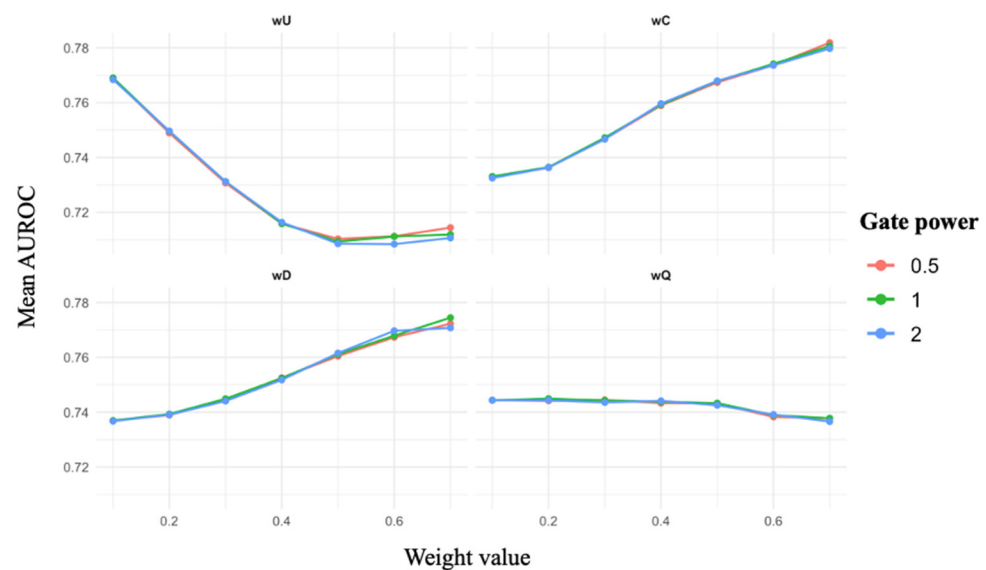


Figure 5. Mean AUROC in logistic regression according to criterion-specific weight values under the gated integration scheme with different gate power settings. The four panels correspond to the weights assigned to uncertainty (w_U), cluster representativeness (w_C), non-redundancy (w_D), and reliability (w_Q). Red, green, and blue lines indicate gate power values of 0.5, 1, and 2, respectively. Note: AUROC, area under the receiver operating characteristic curve. Gate power is a tuning parameter in the gated integration scheme that determines the strength of multiplicative modulation. Higher values indicate stronger gating effects.

This result shows that the gated method is relatively insensitive to moderate changes in the nonlinear gate exponent. Therefore, the empirical performance of the gated operator appears to be driven primarily by the allocation of weights rather than by the precise degree

of multiplicative modulation. Mathematically, this suggests that the gated scoring rule is locally stable with respect to γ over the examined range. Hence, the advantage of the integrated framework cannot be attributed to a narrow tuning effect in the gate exponent, but instead reflects a broader structural property of multi-criterion scoring.

Overall, the results demonstrate three main findings. First, integrated row-selection operators consistently outperformed both full-data training and most marginal selection rules. Second, the gain from integrated selection was heterogeneous across categories, indicating that the method acts as a category-dependent transformation rather than a uniform improvement term. Third, the sensitivity analysis showed that the empirical performance surface is stable across integration forms and relatively robust to the gate exponent, while being systematically shaped by the relative importance of uncertainty, representativeness, diversity, and reliability. Together, these findings support the interpretation of integrated row selection as a mathematically meaningful and robust mechanism for improving predictive discrimination through structured subsampling.

4. Discussion

This study investigated whether predictive model training can be improved by selecting a structured subset of observations rather than indiscriminately using the entire available training set. The results suggest that, in several chief complaint categories, a reduced subset chosen by mathematically constructed sample-level scores can outperform both full-data training and random size-matched subsampling [19,20]. From a modeling perspective, this finding implies that the effective quality of the empirical training distribution may be more important than its raw cardinality [2,21].

The pronounced improvement in AUROC within the Genitourinary & Obstetric/Gynecology and Psychiatric categories suggests that the proposed framework may be particularly useful in domains with substantial clinical heterogeneity and diagnostic uncertainty. Statistically, these categories were characterized by very low AUPRC values (Psychiatric, 0.035; Genitourinary & Obstetric/Gynecology, 0.057) and extreme class imbalance (prevalence: Psychiatric, 0.004; Genitourinary & Obstetric/Gynecology, 0.005), indicating that the prediction task was especially challenging in these strata. In such settings, structured sample selection may help prioritize more informative observations and improve discrimination relative to indiscriminate full-data training.

Mean aggregate AUROC gain masked meaningful subgroup heterogeneity. In several complaint-specific strata, particularly Genitourinary & Obstetric/Gynecology and Psychiatric presentations, the AUROC gain was larger than the overall average. These categories also had extremely low event prevalence and low baseline AUPRC, indicating difficult and clinically vulnerable prediction settings. In such settings, even modest improvements in discrimination may be clinically relevant if they improve the ranking of high-risk patients and support earlier recognition of deterioration. Available signal may be sparse relative to the amount of redundancy or weakly informative variation in the training data. Structured sample selection may therefore provide a larger relative benefit by preferentially retaining observations that are uncertain yet informative, while reducing the influence of redundant or low-value samples. By contrast, in categories where the original training distribution is already more aligned with the prediction task, the incremental benefit of selective retention may be smaller.

The central conceptual contribution of this work is to recast data selection as a formal component of the learning problem. In standard supervised learning, model development is usually framed as parameter estimation conditional on a fixed training set. By contrast, our framework treats the training set itself as an object to be optimized indirectly through score-based ranking [3]. This adds a distinct layer to the learning process: before minimiz-

ing classification loss over model parameters, one may first improve the geometry and informational content of the sample used for fitting. In this sense, the proposed strategy can be interpreted as a data-level regularization mechanism.

A key finding was that the integrated scoring schemes generally outperformed the single-score approaches. This is intuitively and mathematically plausible. Observations in clinical tabular data are rarely optimal with respect to all desirable properties simultaneously. Some samples may be centrally located but highly redundant; others may be uncertain and informative, yet physiologically implausible or structurally noisy. A multi-criteria score is therefore more flexible than any isolated heuristic. The empirical advantage of the integrated formulations supports the idea that sample utility is intrinsically multidimensional.

The reliability-gated and rank-based schemes are especially notable in this regard. The gated scheme introduces a nonlinear interaction between reliability and the remaining criteria, allowing reliability to regulate the influence of uncertainty, representativeness, and non-redundancy. This formulation resembles a multiplicative weighting operator and may be advantageous when low-quality observations should not be promoted even if they appear informative under another metric. The rank-based scheme, in contrast, reduces dependence on score magnitude and instead preserves relative ordering. Its competitive performance suggests that robust subset construction may depend more on correct prioritization than on exact numerical calibration of the component scores.

The comparison against random size-matched subsampling strengthens the interpretation of the results. Because the random baseline used subsets of identical size, the observed performance gains cannot be explained simply by smaller-sample fitting or accidental regularization through data reduction. Rather, the improvement appears to stem from the selective inclusion of observations with favorable structural properties. This distinction is important because it indicates that the method is not merely discarding data, but reorganizing the learning signal.

The relationship between the integrated methods and the full-data model is particularly interesting. In classical statistical thinking, more data are usually considered preferable if the observations are independent and identically distributed. However, in finite and noisy real-world datasets, especially in healthcare, this assumption is only approximately satisfied. Training on all samples may propagate redundancy, local density imbalance, measurement artifacts, and structurally weak observations into the estimation process [22,23]. Selected subsets could yield better generalization than indiscriminate full-data fitting, particularly when redundancy and low-value observations are present.

Nevertheless, the results should be interpreted with appropriate caution. First, the benefit was not perfectly uniform across all performance metrics. In some subgroups, AUROC improved while AUPRC showed smaller gains or even mild decreases. This pattern suggests that the proposed method primarily enhances ranking discrimination, whereas improvements in positive predictive concentration under class imbalance may be less consistent. Consequently, the framework should not be viewed as a universal improvement mechanism for every metric simultaneously.

Second, the present analysis identifies the best integrated configuration by evaluating multiple weighting schemes and aggregation rules. From a strict predictive modeling standpoint, future work should separate tuning and final evaluation more explicitly, for example, by using nested resampling, inner-loop tuning on the training data, [24] or an external validation dataset. Such refinements would provide a more conservative estimate of the generalizability of the integrated scoring framework. The proposed framework introduces additional overhead due to bootstrap scoring, clustering, nearest-neighbor computation, and weight search. This added complexity may limit immediate scalability in

very large datasets or resource-constrained settings. Future work should therefore evaluate runtime, memory usage, and computational scalability more formally.

Third, the base learner in this study was logistic regression. This choice was intentional because it offers a transparent and mathematically tractable reference model, thereby allowing the effect of sample selection itself to be more directly observed. However, the proposed framework is not restricted to generalized linear models [25]. In principle, the same sample-selection logic can be coupled with tree-based methods, kernel machines, boosting algorithms, or neural architectures. It would be particularly interesting to study whether the gains from informative subset selection are amplified or diminished when the downstream learner is more flexible. The robustness of the random forest analysis suggested that the main qualitative findings were not confined to a linear base learner. Under the non-linear ensemble model, the best overall integrated strategy again outperformed both full-data training and random subsampling in mean AUROC (0.824 vs. 0.791 and 0.701, respectively). This pattern supports the interpretation that the proposed framework improves the effective training distribution in a way that is not merely an artifact of logistic regression. At the same time, these supplementary results should be interpreted cautiously, as additional work is still needed to evaluate the framework across a wider range of model classes, hyperparameter settings, and external datasets.

Fourth, the scoring components themselves admit alternative formulations. Uncertainty may be defined through entropy, margin instability, conformal width, or Bayesian posterior variance [26]. Representativeness may be estimated using graph-based centrality or density-based clustering rather than k -means distance. Non-redundancy may be extended beyond nearest-neighbor distance to submodular diversity objectives. Reliability may incorporate missingness patterns, temporal consistency, or domain-specific plausibility rules. Thus, the present framework should be seen not as a closed formula, but as a general mathematical template for observation-wise utility construction.

From an application perspective, the chief complaint-specific results suggest that the usefulness of structured sample selection may depend on subgroup heterogeneity [27,28]. Categories with broader physiological diversity or noisier latent substructure may benefit more from integrated selection than relatively homogeneous categories. This raises an important methodological implication: sample-selection strategies may need to be adapted to subgroup topology rather than assumed to be globally optimal. In emergency medicine, where patient presentations differ substantially by complaint category, such adaptive data curation may be especially valuable.

More broadly, this study contributes to the emerging view that training data should not be treated as a passive input to predictive modeling. Instead, the data can be actively organized, ranked, and filtered according to mathematical criteria that encode informativeness, reliability, and structural diversity [29,30]. This perspective connects predictive modeling most directly with recent work on subset optimization, data collection, and data informativeness [31]. In that sense, the proposed method is not merely a heuristic for preprocessing, but a step toward a more principled theory of sample utility in applied machine learning.

This study was based on a single publicly available emergency department dataset, and, therefore, external generalizability remains limited. Although MIMIC-IV-ED provides a large and reproducible benchmark, it reflects the data-generating process, documentation practices, triage workflows, and patient population of a specific health-system context. The performance and optimal weighting structure of the proposed sample-selection framework may differ in other hospitals, countries, electronic health record systems, or clinical workflows. Therefore, the present findings should be interpreted as methodological evidence supporting the feasibility of multi-criteria row selection, rather than as definitive evidence

of universal clinical transportability. Future studies should evaluate the framework in external multicenter ED datasets and prospectively collected cohorts.

5. Conclusions

This study proposed a mathematically structured framework for sample-level subset selection in emergency department prediction tasks. By assigning each observation a utility score based on uncertainty, cluster representativeness, non-redundancy, and reliability, the framework enabled selective construction of training subsets rather than indiscriminate use of the full dataset. The results showed that multi-criteria integrated selection consistently outperformed random subsampling and, in several categories, also improved upon full-data training. These findings suggest that informative sample selection in heterogeneous clinical data is inherently multi-dimensional and that appropriate subset design can improve the effective training distribution for prediction. Overall, this study highlights training subset design as a meaningful methodological problem in supervised learning and provides a general framework for utility-based sample selection in healthcare prediction modeling.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math14142457/s1>. Table S1: Baseline characteristics of the chief complaint cohort according to mortality; Table S2. Category-specific sample size, event prevalence, and nonlinear (Random Forest) baseline predictive performance across chief complaint groups. Table S3. Comparison of aggregate discrimination performance across row-selection operators in Random Forest. Table S4. Statistical comparison of AUROC between each selection method and full-data training by model type. Table S5. Class Distributions Before and After SMOTE in the Inner Training Set by Chief Complaint Category. Figure S1. Distribution of Box plot for AUROC across by row-selection strategy in logistic regression. Figure S2. Distribution of Box plot for AUROC across by row-selection strategy and chief complaint in logistic regression. Figure S3. Mean AUROC by chief complaint category and row-selection method across repeated experiments in Random Forest. Figure S4. Distribution of Box plot for AUROC across by row-selection strategy in Random Forest. Figure S5. Distribution of Box plot for AUROC across by row-selection strategy and chief complaint in Random Forest. Figure S6. Category-specific AUROC comparison across sampling ratios for logistic regression. Figure S7. Category-specific AUROC comparison across sampling ratios for Random Forest. Figure S8. Category-specific AUROC comparison across sampling ratios for logistic regression with SMOTE. Figure S9. Category-specific AUROC comparison across sampling ratios for Random Forest with SMOTE. Algorithm S1. Overall multi-criteria sample selection framework. Algorithm S2. Construction of sample-level utility scores. Algorithm S3. Integrated Score-Based Subset Selection and Evaluation.

Author Contributions: Conceptualization, J.Y.Y.; methodology, J.Y.Y.; software, J.Y.Y.; validation, D.J. and J.Y.Y.; formal analysis, J.Y.Y.; investigation, J.Y.Y.; resources, J.Y.Y.; data curation, J.Y.Y.; writing—original draft preparation, J.Y.Y.; writing—review and editing, D.J. and S.M.; visualization, J.Y.Y.; supervision, J.Y.Y.; project administration, J.Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (RS-2025-02223418); by the Regional Innovation System and Education (RISE) Glocal University 30 Project through the Gangwon RISE Center, funded by the Ministry of Education (MOE) and Gangwon State (GS), Republic of Korea (2026-RISE-10-009).

Data Availability Statement: The data used in this study are publicly available from the PhysioNet repository as MIMIC-IV, although access is restricted to users who complete the required training and data use agreement.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT 4o for the purposes of English editing and generation of graphical abstract.

Conflicts of Interest: Author SangJun Moon is employed by the Data Science Department, Hana Card Co., Seoul, Republic of Korea, and that the company had no involvement in the study. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUPRC	area under the precision–recall curve
AUROC	area under the receiver operating characteristic curve
BT	body temperature;
DBP	diastolic blood pressure;
ED	emergency department;
HR	heart rate;
IQR	interquartile range;
MIMIC	Medical Information Mart for Intensive Care;
RR	respiratory rate;
SBP	systolic blood pressure;
SD	standard deviation;
SpO2	peripheral oxygen saturation.

References

1. Moser, B.B.; Shanbhag, A.S.; Frolov, S.; Raue, F.; Folz, J.; Dengel, A. A coreset selection of coreset selection literature: Introduction and recent advances. *arXiv* **2025**, arXiv:2505.17799.
2. Zha, D.; Bhat, Z.P.; Lai, K.-H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–42. [[CrossRef](#)]
3. Hu, Y.; Hu, P.; Zhao, H. Most influential subset selection: Challenges, promises, and beyond. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 119778–119810. [[CrossRef](#)]
4. Tan, H.; Wu, S.; Huang, W.; Zhao, S.; Qi, X. Data pruning by information maximization. *arXiv* **2025**, arXiv:2506.01701.
5. Gidiagba, O.J.; Tartibu, L.; Okwu, M. Integrating Machine Learning with Multi-Criteria Decision-Making Models for Sustainable Supplier Selection in Dynamic Supply Chains. *Logistics* **2025**, *9*, 152.
6. Matejová, M.; Paralič, J. A Multi-Criteria Decision-Making Approach for the Selection of Explainable AI Methods. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 158.
7. Xu, F.; Wang, S.; Zhang, J.; Sun, C.; Tang, H.; Zhang, L. UNSEEN: Enhancing Dataset Pruning from a Generalization Perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 27224–27232.
8. Yang, S.; Cao, Z.; Guo, S.; Zhang, R.; Luo, P.; Zhang, S.; Nie, L. Mind the boundary: Coreset selection via reconstructing the decision boundary. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
9. Bi, J.; Xu, Y.; Conrad, F.; Wiemer, H.; Ihlenfeldt, S. A comprehensive benchmark of active learning strategies with AutoML for small-sample regression in materials science. *Sci. Rep.* **2025**, *15*, 37167. [[PubMed](#)]
10. Li, D.; Wang, Z.; Chen, Y.; Jiang, R.; Ding, W.; Okumura, M. A survey on deep active learning: Recent advances and new frontiers. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5879–5899. [[CrossRef](#)]
11. Jubran, I.; Maalouf, A.; Feldman, D. Overview of accurate coresets. *WIREs Data Min. Knowl. Discov.* **2021**, *11*, e1429. [[CrossRef](#)]
12. Beck, N.; Pham, T.; Iyer, R. Theoretical analysis of submodular information measures for targeted data subset selection. *arXiv* **2024**, arXiv:2402.13454.
13. Machnio, J.; Nielsen, M.; Ghazi, M.M. When Uncertainty, Coverage, and Representation Matter in Active Learning Frameworks. In Proceedings of the ICLR 2026 (International Conference on Learning Representations), Rio de Janeiro, Brazil, 23–27 April 2026.
14. Cho, Y.; Shin, B.; Kang, C.; Yun, C. Lightweight dataset pruning without full training via example difficulty and prediction uncertainty. *arXiv* **2025**, arXiv:2502.06905.
15. Xie, F.; Zhou, J.; Lee, J.W.; Tan, M.; Li, S.; Rajnithern, L.S.O.; Chee, M.L.; Chakraborty, B.; Wong, A.-K.I.; Dagan, A. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci. Data* **2022**, *9*, 658. [[CrossRef](#)] [[PubMed](#)]

16. Kansal, A.; Chen, E.; Jin, T.; Rajpurkar, P.; Kim, D. Multimodal clinical monitoring in the emergency department (mc-med). *PhysioNet* **2025**. [CrossRef]
17. Cavalab. Github, Mimic-Chief-Complaint. Available online: <https://github.com/cavalab/mimic-iv-ed-epi/blob/main/mimic-chief-complaint.r> (accessed on 29 March 2026).
18. López, L.; Elsharief, S.; Jorf, D.A.; Darwish, F.; Ma, C.; Shamout, F.E. Uncertainty quantification for machine learning in healthcare: A survey. *arXiv* **2025**, arXiv:2505.02874.
19. Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; pp. 8110–8118.
20. Killamsetty, K.; Durga, S.; Ramakrishnan, G.; De, A.; Iyer, R. Grad-match: Gradient matching based data subset selection for efficient deep model training. In Proceedings of the International Conference on Machine Learning, Shenzhen, China, 26 February–1 March 2021; pp. 5464–5474.
21. Borsos, Z.; Mutný, M.; Tagliasacchi, M.; Krause, A. Data summarization via bilevel optimization. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.
22. Rajaraman, S.; Zamzmi, G.; Yang, F.; Liang, Z.; Xue, Z.; Antani, S. Semantically redundant training data removal and deep model classification performance: A study with chest X-rays. *Comput. Med. Imaging Graph.* **2024**, *115*, 102379. [CrossRef] [PubMed]
23. Subasri, V.; Krishnan, A.; Kore, A.; Dhalla, A.; Pandya, D.; Wang, B.; Malkin, D.; Razak, F.; Verma, A.A.; Goldenberg, A. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw. Open* **2025**, *8*, e2513685. [CrossRef] [PubMed]
24. Clift, A.K.; Dodwell, D.; Lord, S.; Petrou, S.; Brady, M.; Collins, G.S.; Hippisley-Cox, J. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: Cohort study. *BMJ* **2023**, *381*, e073800. [CrossRef] [PubMed]
25. Liu, C.; Shi, J.; Wang, F.; Li, D.; Luo, Y.; Yang, B.; Zhao, Y.; Zhang, L.; Yang, D.; Jin, H. Development and validation of an interpretable multi-task model to predict outcomes in patients with rhabdomyolysis: A multicenter retrospective cohort study. *EClinicalMedicine* **2025**, *87*, 103438. [CrossRef] [PubMed]
26. Sreenivasan, A.P.; Vaivade, A.; Noui, Y.; Khoonsari, P.E.; Burman, J.; Spjuth, O.; Kultima, K. Conformal prediction enables disease course prediction and allows individualized diagnostic uncertainty in multiple sclerosis. *npj Digit. Med.* **2025**, *8*, 224. [CrossRef] [PubMed]
27. Churpek, M.M.; Carey, K.A.; Snyder, A.; Winslow, C.J.; Gilbert, E.; Shah, N.S.; Patterson, B.W.; Afshar, M.; Weiss, A.; Amin, D.N. Multicenter development and prospective validation of eCARTv5: A gradient-boosted machine-learning early warning score. *Crit. Care Explor.* **2025**, *7*, e1232. [PubMed]
28. Wyatt, S.; Lunde Markussen, D.; Haizoune, M.; Vestbø, A.S.; Sima, Y.T.; Sandboe, M.I.; Landschulze, M.; Bartsch, H.; Sauer, C.M. Leveraging machine learning to identify subgroups of misclassified patients in the emergency department: Multicenter proof-of-concept study. *J. Med. Internet Res.* **2024**, *26*, e56382. [PubMed]
29. Jain, E.; Nandy, T.; Aggarwal, G.; Tendulkar, A.; Iyer, R.; De, A. Efficient data subset selection to generalize training across models: Transductive and inductive networks. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 4716–4740. [CrossRef]
30. Mahmood, R.; Lucas, J.; Alvarez, J.M.; Fidler, S.; Law, M. Optimizing data collection for machine learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 29915–29928. [CrossRef]
31. Bennouna, O.; Bennouna, A.; Amin, S.; Ozdaglar, A. What data enables optimal decisions? An exact characterization for linear optimization. *arXiv* **2025**, arXiv:2505.21692.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.