Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v7i2.517

Research Article

Survey-Aware Self-Training Under Label Scarcity for Low-Hemoglobin Screening in NHANES

Irawati [1,*]; Lilis Nur Hayati [2]
[1] Universitas Muslim Indonesia, Makassar 90231, Indonesia, irawati.irawati@umi.ac.id
[2] Universitas Muslim Indonesia, Makassar 90231, Indonesia, lilis.nurhayati@umi.ac.id

Correspondence should be addressed to Irawati; irawati.irawati@umi.ac.id

Received 16 February 2026; Accepted 12 June 2026; Published 31 July 2026

© Authors 2026 CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:
Introduction: Semi-supervised learning may reduce labeled-outcome requirements, but confident pseudo-labels can propagate error. We assessed survey-aware self-training for screening smoking-adjusted low hemoglobin using NHANES August 2021–August 2023. **Method:** The measured cohort included 5,731 adults, of whom 655 met the outcome definition. Predictors comprised demographic, anthropometric, and blood-pressure measures. Outcomes were deliberately withheld from measured respondents after outcome-stratified splitting and mask construction. Five masks provided each fitted model with 20% of training labels and 20% of validation labels; shared tuning used labels from all five masks. Naturally missing-hemoglobin respondents were reserved for descriptive diagnostics. The primary contrast was survey-weighted test average precision (AP) for self-training minus matched supervised boosting, with a 2,000-replicate design-and-mask bootstrap interval. **Results and Discussion:** The weighted test prevalence was 0.0990. Mean AP was 0.1555 for supervision and 0.1539 for self-training (difference, -0.0016; 95% CI, -0.0076 to 0.0024). Mean Brier scores were 0.08969 and 0.08986 (difference, 0.00017; 95% CI, -0.00105 to 0.00188). Four masks admitted no pseudo-labels and produced identical predictions; the fifth admitted 258 and reduced AP by 0.0080. **Conclusion:** The comparison was inconclusive. These findings concern the evaluated boosting and admission rules, and the per-model budgets do not represent the total outcome information used in study design and tuning.

Keywords: Semi-Supervised Learning; Low Hemoglobin; National Health and Nutrition Examination Survey; Complex Survey Data; Probability Calibration

1. Introduction

Machine-learning studies of anemia often begin with laboratory information close to the endpoint. Palika *et al.* predicted inflammation-adjusted iron-deficiency anemia from complete-blood-count indices in women aged 15–50 years using overlapping National Health and Nutrition Examination Survey (NHANES) cycles and external Indian data [1]. That work demonstrates the value of careful outcome construction and external validation, but it addresses an etiologic, laboratory-rich target rather than label efficiency for non-laboratory screening. In the United States, official estimates for August 2021–August 2023 show that anemia remains common and unevenly distributed across adult age and sex groups [2]. The 2024 World Health Organization (WHO) guideline defines anemia from hemoglobin concentration with sex-specific adult cutoffs and categorical adjustments for cigarette smoking [3]. Commentary accompanying the revision emphasizes that a population threshold is not itself a diagnosis of cause [4], while pooled healthy-reference analyses support the cutoffs through age 65 and identify less direct evidence at older ages [5]. These distinctions make low hemoglobin a useful, externally defined prediction endpoint, provided that it is not mislabeled as iron deficiency or clinical diagnosis.

Existing prediction studies leave a specific methodological gap. Supervised classifiers using complete-blood-count variables can distinguish iron-deficiency anemia, but their strongest predictors are measured in the same laboratory panel as hemoglobin and therefore do not represent laboratory-free screening [6]. Anthropometric markers have shown

screening value in Malaysian young adults [7], suggesting that non-laboratory measurements can carry signal, although that population and fully supervised design differ from a general-adult United States survey. A recent NHANES preprint developed a supervised, laboratory-heavy anemia model for older adults with temporal validation [8]; its non-peer-reviewed status, restricted population, and unweighted evaluation do not answer whether unlabeled records help under a fixed annotation budget.

NHANES provides a demanding test bed for this question. The August 2021–August 2023 program combines household interviews with standardized mobile-examination-center measurements in a stratified, multistage probability sample [9]. The released cycle includes masked strata and primary sampling units, and its analytic guidance stresses weight selection and nonresponse assessment [10]. Hemoglobin was measured from venous blood under documented quality-control procedures [11], and analyses involving the blood component require the phlebotomy examination weight rather than an interview-only weight [12]. A model that ignores these design features may summarize the observed sample while misrepresenting population performance.

Semi-supervised learning (SSL) offers a plausible response to outcome-label scarcity but no guaranteed advantage. A two-stage semi-supervised method has been studied for chronic-disease risk prediction in NHANES [13], establishing precedent for unlabeled survey records but not for smoking-adjusted low hemoglobin or design-aware performance estimation. More generally, self-training is a wrapper method that converts confident predictions into additional training targets and depends on assumptions about class structure and labeled–unlabeled comparability [14]. Tabular representation learning shows that unlabeled rows can be useful when those assumptions are sufficiently aligned [15]. Realistic SSL evaluation, however, requires a strong supervised comparator, explicit accounting for validation labels, and attention to distribution mismatch [16]. Pseudo-labeling can amplify the learner's own errors through confirmation bias [17], making confidence control and post-selection diagnostics central rather than optional.

Complex survey structure adds a second layer of risk. Cross-validation and loss estimation for probability samples should preserve relevant design structure and incorporate survey weights [18]. In NHANES gradient boosting, using sampling weights can materially change the population-relevant result [19]. Design weights are likewise needed for population sensitivity, specificity, and discrimination estimates [20], including receiver-operating-characteristic methods developed specifically for complex surveys [21], [22]. Strata, clusters, and weights also inform variance estimation [23]. In the admission rule studied here, a stratum cap limits concentration of pseudo-label mass and a class-prior cap limits negative mass relative to admitted positive mass.

Discrimination alone is insufficient. Calibration intercepts, slopes, and curves reveal whether predicted risks agree with observed frequencies [24]. Average precision is informative for imbalanced outcomes because its no-skill level follows prevalence [25], while Brier score and log loss are proper scoring rules for probabilistic predictions [26]. Decision-curve analysis provides a complementary description of net benefit across action thresholds [27]. Current reporting and appraisal guidance further requires transparent separation of development and evaluation, complete accounting of participants and predictors, and explicit limits on intended use [28], [29].

Regional prediction studies illustrate logistic regression for liver disease [30], tree tuning for diabetes [31], leakage control in insurance claims [32], and class balancing for dementia classification [33]. Other work examines SSL across labeled fractions [34], boosting with synthetic balancing for cirrhosis [35], decision trees on imbalanced skin images [36], and non-laboratory obesity prediction [37]. These studies motivate careful comparator and validation choices, rather than serving as direct comparisons with our outcome.

We asked whether prior-constrained, survey-aware confidence self-training improves population-weighted screening discrimination over an identical supervised learner when each fitted model uses a 20% subset of training and validation outcomes. The primary comparison uses smoking-adjusted WHO thresholds and a respondent-disjoint test set. Secondary analyses examine other per-model label budgets, a confidence-only variant, an interview-only panel, an unadjusted endpoint, and a fully labeled supervised condition. The contribution is the combined evaluation design and its empirical result.

2. Methods

Study design and estimand

We conducted a retrospective prediction-methods study using public, deidentified NHANES August 2021–August 2023 data. The respondent was the unit of analysis. The confirmatory estimand was the paired difference in survey-weighted test AP between survey-aware prior-constrained self-training and a matched supervised histogram-gradient-boosting classifier at a 20% outcome-label budget, averaged across five deterministic label masks. Both methods used the same eligible respondents, predictors, preprocessing recipe, label masks, validation-label budget, base-learner search, calibration rule, and test set. The contrast therefore isolates the incremental contribution of admitted pseudo-labels under the evaluated procedure.

The estimand concerns label efficiency within one survey cycle. It is not a causal effect, temporal forecast, external-transport estimate, etiologic anemia diagnosis, or clinical-utility claim. The interpretation was fixed before model fitting: benefit required a positive point estimate whose 95% interval excluded zero; an interval crossing zero was inconclusive; and a negative interval excluding zero indicated harm.

Data sources, eligibility, and outcome

Five official public-use components were linked one-to-one by respondent sequence number (SEQN): demographics (DEMO_L), body measures (BMX_L), oscillometric blood pressure (BPXO_L), complete blood count (CBC_L), and smoking–cigarette use (SMQ_L). CBC_L supplies hemoglobin (LBXHGB) and the phlebotomy weight (WTPH2YR); DEMO_L supplies the interview weight (WTINT2YR), masked stratum (SDMVSTRA), and PSU (SDMVPSU). Eligible respondents were aged at least 20 years and were not coded pregnant in the released pregnancy-status variable. That field is publicly reported only for women aged 20–44, so eligibility does not prove nonpregnancy outside that range. The primary modeling cohort additionally required measured hemoglobin and a positive phlebotomy weight. The outcome was derived only after eligibility was established.

Current cigarette smoking adjusted measured hemoglobin according to the 2024 WHO categorical rule. We subtracted 0.3 g/dL for current smokers reporting fewer than 10 cigarettes per day or an unknown quantity, 0.5 g/dL for 10–19 cigarettes per day, and 0.6 g/dL for at least 20 cigarettes per day. No subtraction was applied to respondents not reporting current smoking. Low hemoglobin was defined as an adjusted concentration below 13.0 g/dL for men and below the WHO nonpregnant-female cutoff of 12.0 g/dL for eligible women. Smoking fields were outcome-construction variables only and were prohibited as predictors. The endpoint does not identify iron deficiency, etiology, persistence, symptoms, or treatment need.

Predictor panels and leakage control

The primary exam-enhanced panel contained age, sex, race and Hispanic origin, birthplace, educational attainment, marital status, family income-to-poverty ratio, household size, body mass index, waist circumference, upper-arm circumference, and respondent-level means of up to three systolic-pressure, diastolic-pressure, and pulse readings. The secondary interview-only panel retained the demographic and socioeconomic predictors but omitted body and blood-pressure measures.

Respondent identifiers, survey weights, strata, primary sampling units, pregnancy, interview or examination status, source-component presence flags, all smoking variables, hemoglobin, every other complete-blood-count variable, laboratory measures, and values derived from the outcome were excluded from prediction. Implausible examination values outside prespecified ranges were set missing; numeric missing values were median-imputed and robust-scaled. Refused and unknown demographic responses were set missing and categorical missingness was represented explicitly before one-hot encoding. Preprocessing was fitted on training-partition predictor rows, including rows with withheld labels, and reused unchanged for paired methods. Supplementary Table S1 specifies source variables, derivations, and missing-value handling.

Respondent split and label budgets

The 5,731 measured-outcome respondents were deterministically assigned 60% to training, 20% to validation, and 20% to test using seed 20260916. Allocation used all measured outcomes to stratify within outcome-by-survey-stratum-by-PSU cells where their size supported all partitions and within outcome-by-stratum cells otherwise.

Respondents were disjoint across roles. Both released PSUs remained represented in each partition of every stratum; their appearance in more than one partition precludes a claim of unseen-cluster transport.

Five deterministic mask seeds, 20260916 through 20260920, were generated within training and validation using all 4,585 development outcomes to form outcome-by-stratum-by-PSU cells. Labels were then deliberately withheld from individual fits in nested 10%, 20%, and 40% prefixes. Each 20% mask supplied 688 training labels (78 events) and 229 validation labels (28 events); calibration and threshold selection used labels within that validation subset. Across masks, shared tuning directly used 2,301 distinct training labels (265 events) and 779 distinct validation labels (94 events), or 3,080 unique development outcomes. Thus the nominal budget is per fitted model, not the total outcome information required by the retrospective design. The 10% and 40% analyses reused the configuration selected using the 20% masks.

After the outcome-stratified split was constructed, test outcomes were not used for preprocessing, fitting, tuning, calibration, threshold selection, or development diagnostics. Test outcomes were joined to fixed test probability projections for final evaluation. No post-test model refitting, retuning, recalibration, threshold reselection, or probability regeneration was performed.

Learners and survey-aware self-training

A no-skill reference predicted the survey-weighted prevalence among budget-permitted training labels. Weighted logistic regression was included as a secondary baseline. The matched comparator was a histogram-gradient-boosting classifier fitted with class-balanced phlebotomy weights. The grid varied maximum leaf nodes (7 or 15), minimum leaf size (20 or 40), L2 regularization (1 or 5), confidence gate (0.90 or 0.95), and pseudo-label multiplier (0.25 or 0.50). Mean survey-weighted validation AP across the five 20% masks selected the shared configuration: 15 leaves, minimum leaf size 40, L2 5, learning rate 0.05, 250 iterations without early stopping, gate 0.95, and multiplier 0.50. Ties used a fixed ordering.

Survey-aware self-training began from the matched supervised model. At each of at most three iterations, its raw predictions were calibrated using the permitted validation labels. A hidden training row was eligible as positive at calibrated probability at least 0.95 or as negative at probability at most 0.05. Positives were considered before negatives; within each class candidates were ranked by distance from 0.5, with SEQN breaking ties. The iteration stopped without admitting pending labels if either class had fewer than five candidates before or after caps.

For cap calculations, pseudo-label mass equals the row's phlebotomy weight multiplied by 0.50 and its confidence, the larger of the calibrated probabilities for the two classes. Cumulative pseudo mass within a stratum could not exceed the original labeled phlebotomy-weight mass in that stratum. Cumulative negative pseudo mass could not exceed cumulative positive pseudo mass multiplied by (1 - pL)/pL, where pL is the phlebotomy-weighted positive prevalence among the original training labels. For fitting, each original weight was multiplied by its class-balancing factor, calculated as the original labeled count divided by twice that class's total phlebotomy weight. Each pseudo-label's fitting weight was WTPH2YR multiplied by its pseudo-class factor, 0.50, and confidence. Supplementary Table S2 gives the per-mask admission and stopping record.

The matched supervised condition used the selected base-learner configuration without pseudo-labels. Naive confidence-only self-training removed the stratum and class-prior admission caps but otherwise shared the selected base learner and confidence rule. A 100%-label supervised model provided an upper-reference condition within the same family; it was not a separately optimized competitor.

Calibration, thresholds, and performance measures

Predicted probabilities were calibrated on budget-permitted validation labels after model fitting and, for admission, at each self-training iteration. Weighted isotonic calibration was used with at least 100 validation observations and at least 10 observations of each class; otherwise weighted sigmoid calibration was used. For thresholded metrics, the probability cutoff maximizing weighted Youden's J on the same permitted validation subset was fixed separately for each seed and method.

The primary endpoint was survey-weighted AP because low hemoglobin was imbalanced and the precision–recall baseline depends on prevalence. Secondary population-weighted measures included AUROC, Brier score, log loss, ECE, calibration intercept and slope, sensitivity, specificity, balanced accuracy, and F1. For each mask and method, calibration intercept and slope were estimated jointly by unpenalized weighted logistic regression of test outcomes on

the logit of predicted probability, bounded to 0.000001–0.999999. The intercept is not calibration-in-the-large with the slope fixed at one. Ideal joint calibration has intercept zero and slope one. Decision-curve net benefit was estimated at thresholds from 0.05 to 0.30.

Uncertainty, subgroups, and sensitivity analyses

The primary 95% interval used 2,000 paired replicates generated with seed 20261916. In each test stratum, one of its two PSUs was sampled and its respondents' phlebotomy weights were doubled; the other PSU received zero weight. The same PSU choice applied to both methods and all masks in a replicate. Five mask seeds were sampled with replacement, and their paired contrasts were averaged. The 2.5th and 97.5th percentiles formed the AP interval; the Brier contrast was treated analogously. This is a two-PSU Rao–Wu-style weight-resampling adaptation [38], [39] combined with mask resampling. It is conditional on the fixed split, masks, fitted models, calibrated predictions, and selected configuration; it does not include refitting or all model-selection and external-transport uncertainty.

Prespecified descriptive subgroups were sex, age 20–59 versus at least 60 years, and race/Hispanic-origin groups with at least 40 test respondents and 10 events. Smaller cells were suppressed, and no interaction claim was permitted. Secondary sensitivity analyses evaluated 10% and 40% labels, the interview-only panel, the unadjusted-hemoglobin endpoint, naive self-training, and the fully labeled supervised reference. Naturally missing-Hb respondents were reserved for comparing predictor distributions and model-implied pseudo-positive prevalence with the measured cohort after all model settings had been prespecified. No outcome-performance claim was possible for those records.

Software, reproducibility, and ethics

The analysis used Python 3.12 with NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.16.2, joblib 1.5.2, Matplotlib 3.10.6, and seaborn 0.13.2. All randomness was seeded, and cohort construction, partitions, masks, preprocessing, model selection, calibration, predictions, metrics, tables, and figures were generated programmatically. A deterministic replay reproduced the cohort, split, masks, selected configuration, and outcome-free test projections; tables and figures were rebuilt from hash-verified recorded results without recomputing test metrics.

NHANES public-use files are deidentified and publicly released for secondary analysis. No new participant contact or data collection occurred. We make no institutional-review determination beyond those source facts. The models were evaluated only as retrospective methodological screening tools and must not be used as diagnostic devices, clinical triage systems, or substitutes for hemoglobin measurement.

3. Results

Study population and primary comparison

Among 11,933 linked respondents, 7,768 adults met the released public-use age and pregnancy-status rules. Of these, 5,731 had measured hemoglobin and a positive phlebotomy weight, whereas 2,037 had naturally missing hemoglobin. The measured cohort contained 655 smoking-adjusted low-hemoglobin cases, corresponding to 11.43% unweighted and 10.10% survey-weighted prevalence. The deterministic split assigned 3,439 respondents to training, 1,146 to validation, and 1,146 to test. The test set contained 126 events and had survey-weighted prevalence of 9.90%.

Table 1. Study population, outcome frequency, and deterministic respondent split. Counts are unweighted unless stated otherwise; Hb denotes hemoglobin.

Measure	Value	Basis
Eligible adults	7,768	Age at least 20; not coded pregnant
Observed Hb	5,731	Positive phlebotomy weight
Naturally missing Hb	2,037	No confirmatory outcome
Adjusted low-Hb cases	655	Unweighted count
Adjusted low-Hb prevalence	10.10%	Survey weighted
Training partition	3,439	60% split
Validation partition	1,146	20% split
Test partition	1,146	20% split

At the prespecified 20% label budget, mean survey-weighted AP was 0.1555 for matched supervision and 0.1539 for survey-aware self-training. The paired self-training-minus-supervised difference was -0.0016 (95% design-and-mask CI, -0.0076 to 0.0024). Because the point estimate was negative and the interval included zero, the criterion for benefit was not met and the result was inconclusive. Mean Brier score was 0.08969 for supervision and 0.08986 for self-training; the difference was 0.00017 (95% CI, -0.00105 to 0.00188). Secondary primary-budget metrics were similar (Table 2).

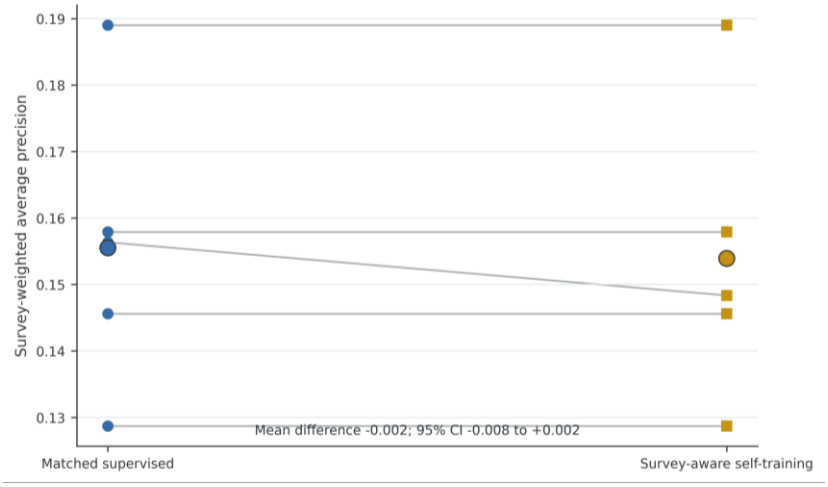


Figure 1. Paired primary comparison at the 20% label budget.

Table 2. Primary test performance at the 20% label budget.

Metric	Supervised	Survey-aware SSL	Difference (95% CI)
Average precision	0.1555	0.1539	-0.0016 (-0.0076, 0.0024)
AUROC	0.6494	0.6470	-0.0024
Brier score	0.08969	0.08986	0.00017 (-0.00105, 0.00188)
Log loss	0.3877	0.3807	-0.0070
Expected calibration error	0.0256	0.0268	0.0012
Calibration intercept	-1.8193	-1.8201	-0.0008
Calibration slope	0.1512	0.1517	0.0004
Sensitivity	0.6085	0.6062	-0.0023
Specificity	0.6147	0.6203	0.0056
Balanced accuracy	0.6116	0.6133	0.0016
F1	0.2424	0.2435	0.0011

The two approaches produced identical test probabilities for four masks. In the remaining mask, AP changed from 0.1563 under matched supervision to 0.1484 under self-training, a difference of -0.0080; its Brier score increased by 0.00085. This single nonzero mask contrast determined the negative mean primary point estimate.

Label budgets and secondary comparators

The direction of the self-training contrast varied across per-model label budgets (Table 3 and Figure 2). At 10% labels, mean AP was 0.1527 for matched supervision and 0.1568 for survey-aware self-training, a descriptive difference of 0.0041. At 40% labels, the values were 0.1857 and 0.1868, a difference of 0.0012. Contrasts were calculated before rounding, so subtracting displayed means can differ by one in the last displayed digit. Naive self-training yielded AP 0.1576 at the primary budget. Weighted logistic regression achieved AP 0.1799 at 20% and 0.1989

at 40%, descriptively exceeding the restricted-label boosting models. The fully labeled supervised reference achieved AP 0.2030, AUROC 0.7324, and Brier score 0.08505.

Table 3. Prespecified sensitivity and comparator performance

Endpoint / panel	Labels	Method	AP	AUROC	Brier
Adjusted / exam	10%	Supervised	0.1527	0.6288	0.09210
Adjusted / exam	10%	Survey-aware SSL	0.1568	0.6399	0.09270
Adjusted / exam	20%	Naive SSL	0.1576	0.6510	0.08883
Adjusted / exam	20%	Weighted logistic	0.1799	0.6751	0.08762
Adjusted / exam	40%	Supervised	0.1857	0.6937	0.08671
Adjusted / exam	40%	Survey-aware SSL	0.1868	0.6945	0.08658
Adjusted / exam	40%	Weighted logistic	0.1989	0.6893	0.08567
Adjusted / exam	100%	Supervised reference	0.2030	0.7324	0.08505
Adjusted / interview	20%	Supervised	0.1606	0.6510	0.09184
Adjusted / interview	20%	Survey-aware SSL	0.1538	0.6448	0.09067
Unadjusted / exam	20%	Supervised	0.1620	0.7004	0.08031
Unadjusted / exam	20%	Survey-aware SSL	0.1663	0.7019	0.07952

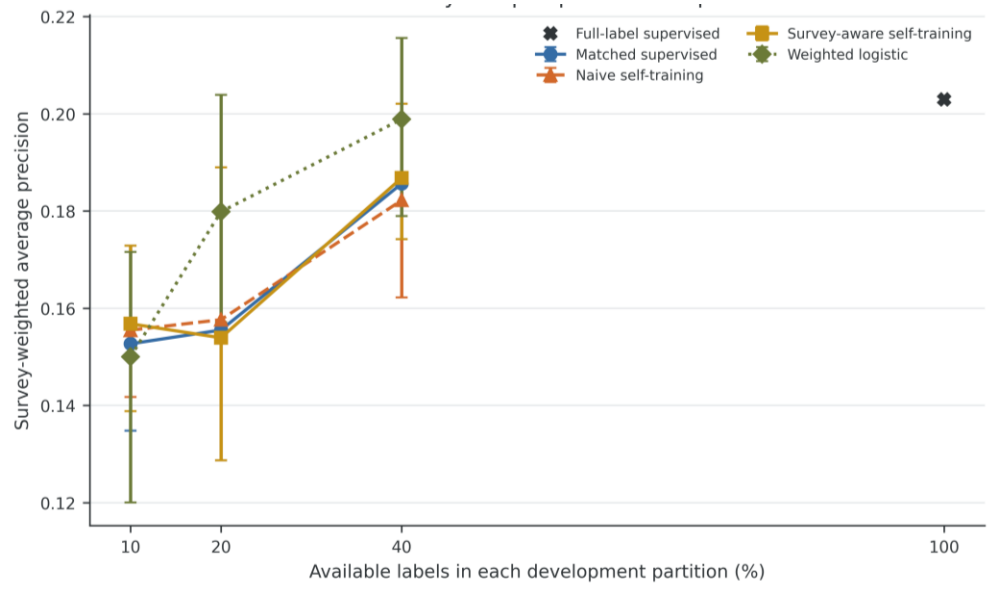


Figure 2. Label budgets and prespecified comparators.

With interview-only predictors at 20% labels, AP was 0.1606 for matched supervision and 0.1538 for self-training. For the unadjusted-hemoglobin endpoint, the direction reversed descriptively: AP was 0.1620 and 0.1663, respectively. These secondary findings did not show a consistent advantage for pseudo-labeling.

Calibration and decision curves

At the primary budget, mean per-mask calibration intercepts were -1.8193 for supervision and -1.8201 for self-training, and mean slopes were 0.1512 and 0.1517. Slopes far below one indicate predicted risks that are too extreme under this calibration model; the similarity of the methods does not imply adequate absolute calibration. Many isotonic predictions reached the numerical probability bounds, so the slopes may depend on those bounds. ECE was 0.0256 and 0.0268. **Figure 3** instead bins the probability averaged across five models for each respondent; it is not a plot of the mean per-mask slope. These ensemble mean probabilities were 0.1140 and 0.1151 against an observed weighted test rate of 0.0990. Upper probability bins contained only 4–19 respondents and were correspondingly unstable.

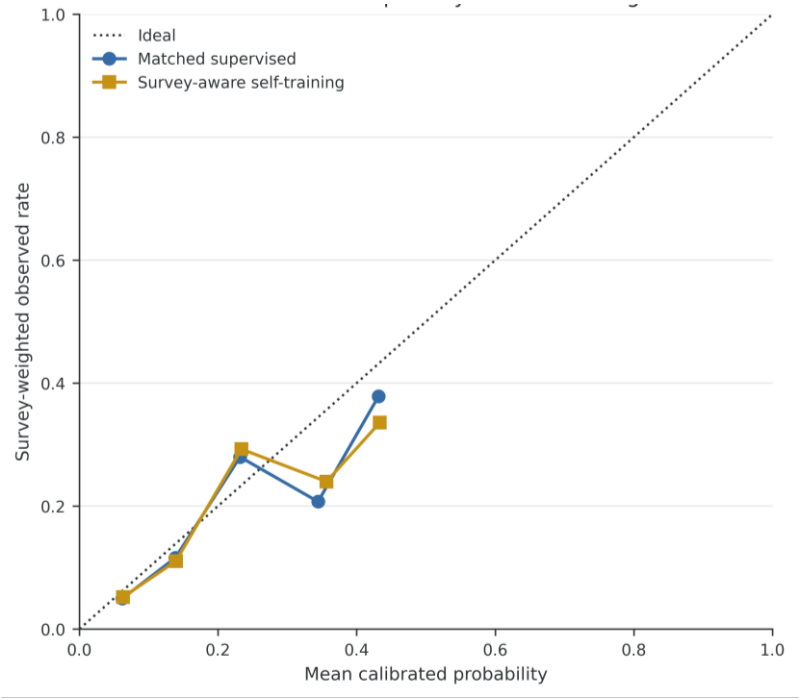


Figure 3. Calibration at the primary 20% label budget.

Decision curves nearly overlapped; the largest absolute self-training-minus-supervised net-benefit difference was 0.0012. Both models had positive estimated net benefit relative to treating none and exceeded treating all across thresholds from 0.05 through 0.17, but neither method showed a consistent advantage over the other. No uncertainty interval was estimated for net benefit, so this descriptive analysis does not establish clinical utility (**Figure 4**).

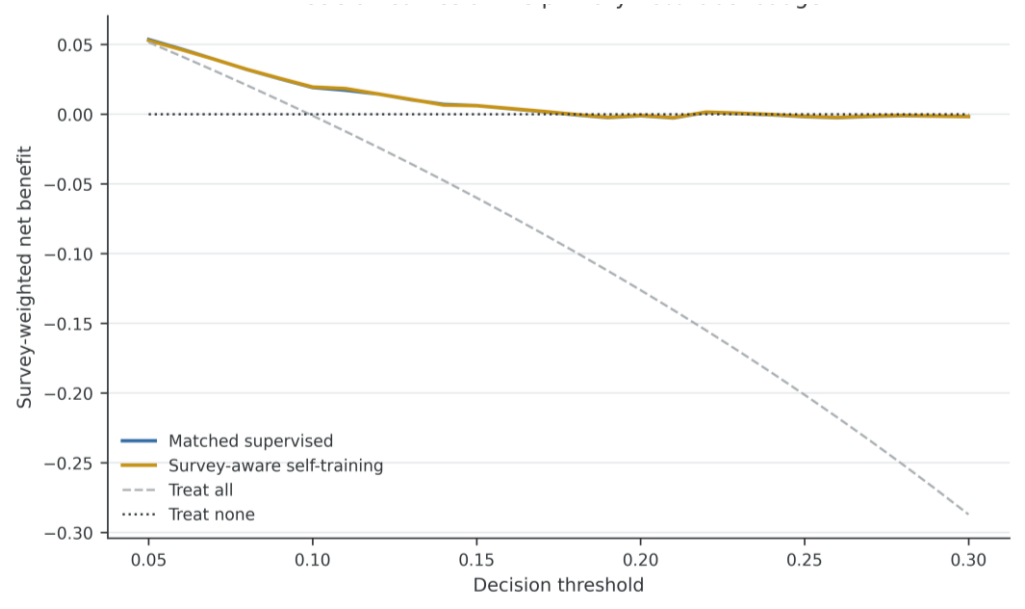


Figure 4. Decision curves at the primary 20% label budget.

Pseudo-label, subgroup, and missing-outcome diagnostics

Survey-aware self-training admitted pseudo-labels in only one of five primary-budget masks. The four inactive masks stopped at their first iteration because no positive candidates reached the gate. The active mask admitted 30

positive and 228 negative pseudo-labels, then stopped at its second iteration because only four positive candidates remained eligible. Their survey-weighted agreement with withheld outcomes was 0.8895, whereas mean model confidence was 0.9999. Naive admission in the same mask accepted 822 records with weighted agreement of 0.9374. The complete per-mask record appears in Supplementary **Table 2**.

Descriptive subgroup AP differences were small and directionally inconsistent, ranging from -0.0064 among adults aged at least 60 years to 0.0087 among non-Hispanic Black respondents. Mexican American and non-Hispanic Asian estimates were suppressed because they had four and nine test events. These analyses were not designed to support interaction or differential-performance claims.

Among 2,037 adults with naturally missing hemoglobin, mean predicted probability was 0.1142, compared with 0.1079 in the measured cohort; threshold-positive fractions were 0.2941 and 0.2459. These estimates used the five calibrated, interview-only survey-aware models at the 20% per-model budget, averaging probabilities per respondent and applying the mean validation-selected threshold of 0.1415. Both cohorts were summarized using interview weights (WTINT2YR). Income-to-poverty missingness was 30.08% versus 13.45%. These quantities describe predictor and prediction shift, not observed prevalence or performance in the naturally missing group.

4. Discussion

In this prespecified, survey-aware comparison, self-training did not improve the primary population-weighted discrimination measure over an otherwise identical supervised learner at a 20% label budget. The AP point estimate slightly favored matched supervision, and the interval included both modest harm and a very small benefit. The Brier-score interval also crossed zero. Accordingly, the result is inconclusive under the fixed decision rule: it is not evidence of superiority, but it is also not an equivalence or noninferiority finding.

The mask-level results clarify how the aggregate contrast arose. The conservative procedure admitted no pseudo-labels in four masks, leaving predictions unchanged from matched supervision. In the fifth, admitted pseudo-labels had very high assigned confidence but lower agreement with withheld reference outcomes, and test AP declined. Stratum and class-prior constraints can limit admitted pseudo-label mass but do not guarantee label correctness. This pattern is consistent with the dependence of self-training on structural assumptions [14] and its susceptibility to confirmation bias [17].

Secondary results did not reveal a stable label-efficiency advantage. Survey-aware self-training had slightly higher mean AP at 10% and 40% labels but lower AP at the confirmatory 20% budget; the direction also changed when the predictor panel or endpoint definition changed. Such variability is consistent with evidence that semi-supervised performance depends on the dataset, learner, and labeled fraction [16], [34]. Because these analyses were descriptive and lacked confirmatory intervals, their point estimates should not rescue or overturn the primary result.

The strong descriptive performance of weighted logistic regression at the 20% and 40% budgets is also informative. For this low-dimensional tabular feature set, a simpler weighted model was competitive with or better than the selected boosting procedure. The fully labeled supervised reference had the highest AP descriptively among the reported models. These observations do not establish a universal algorithm ranking, but they argue for retaining strong simple baselines in label-scarcity studies.

The finding does not conflict with reports that unlabeled tabular data can be useful [15] or with prior semi-supervised applications to NHANES chronic-disease prediction [13]. Those studies evaluated different populations, representations, outcomes, and validation designs. Performance from anemia models using complete-blood-count variables [1], [6] or broader laboratory profiles [8] is not directly comparable with the non-laboratory screen evaluated here. Excluding endpoint-proximal laboratory variables reduces attainable discrimination but avoids circular prediction.

Calibration and decision-curve findings did not identify a practical advantage for self-training. Similar Brier scores and ECE do not remove the concern raised by slopes far below one; those slopes may also be sensitive to the imposed probability bounds. The decision curves almost completely overlapped. Positive estimated net benefit over part of the hypothetical threshold range does not establish clinical utility without a defined action pathway and evidence about downstream benefits and harms [24], [26], [27].

Finally, survey-aware analysis affects the scope of inference. Weighting, strata, and primary sampling units can materially change population-relevant performance estimates [18]–[23]. The naturally missing-hemoglobin group differed in predictor distributions and model-implied risk, reinforcing that phlebotomy nonresponse should not be treated as exchangeable unlabeled data. Descriptive subgroup contrasts were small, inconsistent, and based on limited event counts. External temporal and geographic validation, substantially more label-mask repetitions, and newly developed pseudo-label selection strategies would be needed before stronger generalization or implementation claims could be considered.

5. Limitations

Several limitations qualify these findings. First, label withholding was simulated in a measured cohort. Outcome information from all measured respondents informed the stratified split, and all development outcomes informed mask construction; shared tuning directly used labels from the union of five 20% masks. Consequently, the per-model budgets are not prospective label-acquisition costs, and the 10% analysis also inherits settings selected at 20%. The interval does not establish equivalence or noninferiority.

Second, the evaluation addressed one self-training rule, one gradient-boosting family, five label masks, and one NHANES cycle. At the primary budget, the survey-aware procedure admitted pseudo-labels in only one mask. The estimand reflects both accepted pseudo-labels and the procedure's frequent abstention. Five masks limit characterization of label-selection variability, and the bootstrap interval does not encompass every source of model-selection or implementation uncertainty.

Third, this was internal, within-cycle respondent validation. It does not establish temporal, geographic, health-system, or unseen-cluster transport. Masked survey primary sampling units appeared across partitions, so the evaluation does not represent deployment to wholly unseen clusters. Naturally missing hemoglobin was associated with predictor and missingness differences and could not be evaluated for outcome performance.

Fourth, the endpoint was threshold-defined low hemoglobin, not iron-deficiency anemia or an etiologic diagnosis. Altitude adjustment was unavailable, and released pregnancy status does not identify pregnancy outside ages 20–44. Evidence for hemoglobin thresholds above age 65 is less direct. The intentional exclusion of hemoglobin, other complete-blood-count indices, smoking variables, and laboratory predictors reduces direct comparability with laboratory-rich anemia models.

Fifth, calibration at low label budgets was based on relatively few positive validation outcomes. The reported calibration slopes use bounded probabilities, and their sensitivity to a different bound was not assessed. Estimation of calibration intercepts and slopes produced optimizer warnings; although the reported coefficients were finite, their numerical stability was not independently verified. Subgroup estimates were descriptive, some race and Hispanic-origin cells were suppressed, and no interaction inference was conducted. Decision curves used hypothetical thresholds and do not demonstrate patient benefit, cost-effectiveness, or readiness for clinical use.

6. Conclusion

For the evaluated admission rule, histogram-gradient-boosting configuration, and single NHANES cycle, self-training did not demonstrate higher survey-weighted AP than matched supervision at the 20% per-model budget. Conservative admission frequently abstained, and high confidence did not ensure accurate pseudo-labels. The inconclusive comparison supports reporting matched supervised baselines and pseudo-label diagnostics when evaluating such methods.

Data Availability

The study used public NHANES August 2021–August 2023 demographic, body-measures, oscillometric blood-pressure, complete-blood-count, and smoking–cigarette-use files distributed by the U.S. National Center for Health Statistics. The cycle overview, component documentation, and access links are available through the official NHANES website. No new participant data were collected.

Code Availability

The supplementary material includes analysis-function extracts, the selected configuration, respondent split and label-mask definitions, stored results, a data dictionary, and diagnostics. An accompanying editorial code record contains the original analysis scripts and protocol. Neither archive is a complete executable end-to-end pipeline or a public repository. The public NHANES source files are available from the National Center for Health Statistics.

References:

[1] R. Palika, T. Dasi, D. J. Parasannanavar, S. K. Banjara, B. Kulkarni, and R. Pullakhandam, "Machine learning-based prediction of inflammation adjusted iron deficiency anaemia using blood cell indices," *The Indian Journal of Medical Research*, vol. 163, no. 4, pp. 469–476, 2026. doi: https://doi.org/10.25259/IJMR_1500_2025.

[2] A. M. Williams, N. Ansai, N. Ahluwalia, and D. T. Nguyen, "Anemia Prevalence: United States, August 2021–August 2023," National Center for Health Statistics (U.S.), *NCHS Data Brief*, no. 519, pp. 1–8, 2024. doi: https://doi.org/10.15620/cdc/168890.

[3] World Health Organization, *Guideline on haemoglobin cutoffs to define anaemia in individuals and populations*. World Health Organization, 2024, ISBN 978-92-4-008854-2. [Online]. Available: https://www.who.int/publications/i/item/9789240088542. [Accessed: Oct. 26, 2026].

[4] S.-R. Pasricha, L. Rogers, F. Branca, and M.-N. Garcia-Casal, "Measuring haemoglobin concentration to define anaemia: WHO guidelines," *The Lancet*, vol. 403, no. 10440, pp. 1963–1966, 2024. doi: https://doi.org/10.1016/S0140-6736(24)00502-6.

[5] S. Braat et al., "Haemoglobin thresholds to define anaemia from age 6 months to 65 years: estimates from international data sources," *The Lancet Haematology*, vol. 11, no. 4, pp. e253–e264, 2024. doi: https://doi.org/10.1016/S2352-3026(24)00030-9.

[6] S. Pullakhandam and S. McRoy, "Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning," *BioMedInformatics*, vol. 4, no. 1, pp. 661–672, 2024. doi: https://doi.org/10.3390/biomedinformatics4010036.

[7] S. Dutta, I. R. Karkada, P. Sengupta, and S. V. Chinni, "Anthropometric Markers With Specific Cut-Offs Can Predict Anemia Occurrence Among Malaysian Young Adults," *Frontiers in Physiology*, vol. 12, Art. no. 731416, 2021. doi: https://doi.org/10.3389/fphys.2021.731416.

[8] Z. Niu, L. Ren, and N. Hu, "Development and temporal validation of a machine learning risk prediction model for anemia in older adults: A NHANES study of an inflammatory–nutritional–renal profile," *Research Square*, preprint, 2026. doi: https://doi.org/10.21203/rs.3.rs-9486902/v1.

[9] A. L. Terry, M. M. Chiappa, J. McAllister, D. A. Woodwell, and J. E. Graber, "Plan and Operations of the National Health and Nutrition Examination Survey, August 2021–August 2023," National Center for Health Statistics (U.S.), *Vital and Health Statistics, Series 1*, no. 66, 2024. doi: https://doi.org/10.15620/cdc/151927.

[10] National Center for Health Statistics, "Brief Overview of Sample Design, Nonresponse Bias Assessment, and Analytic Guidelines for NHANES August 2021–August 2023," *NHANES Analytic Guidelines*, 2024. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewbrief.aspx?Cycle=2021-2023. [Accessed: Sep. 16, 2026].

[11] National Center for Health Statistics, "Complete Blood Count with 5-Part Differential—Whole Blood (CBC_L)," *NHANES Data Documentation*, 2024. [Online]. Available: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/CBC_L.htm. [Accessed: Sep. 16, 2026].

[12] National Center for Health Statistics, "NHANES Tutorials: Weighting Module," *NHANES Tutorials*, n.d. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx. [Accessed: Sep. 16, 2026].

[13] Z. Mao, Y. Fukuma, H. Tsukada, and S. Wada, "Risk prediction of chronic diseases with a two-stage semi-supervised clustering method," *Preventive Medicine Reports*, vol. 32, Art. no. 102129, 2023. doi: https://doi.org/10.1016/j.pmedr.2023.102129.

[14] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020. doi: https://doi.org/10.1007/s10994-019-05855-6.

[15] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 11033–11043, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/7d97667a3e056acab9aaf653807b4a03-Abstract.html. [Accessed: Oct. 26, 2026].

[16] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic Evaluation of Deep Semi-Supervised Learning Algorithms," in *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/c1fea270c48e8079d8ddf7d06d26ab52-Abstract.html. [Accessed: Oct. 26, 2026].

[17] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: https://doi.org/10.1109/IJCNN48605.2020.9207304.

[18] J. Wieczorek, C. Guerin, and T. McMahon, "K-fold cross-validation for complex sample surveys," *Stat*, vol. 11, no. 1, Art. no. e454, 2022. doi: https://doi.org/10.1002/sta4.454.

[19] N. MacNell et al., "Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting," *PLOS ONE*, vol. 18, no. 1, Art. no. e0280387, 2023. doi: https://doi.org/10.1371/journal.pone.0280387.

[20] A. S. Wadekar and J. P. Reiter, "Evaluating Binary Outcome Classifiers Estimated from Survey Data," *Epidemiology*, vol. 35, no. 6, pp. 805–812, 2024. doi: https://doi.org/10.1097/EDE.0000000000001776.

[21] W. Yao, Z. Li, and B. I. Graubard, "Estimation of ROC curve with complex survey data," *Statistics in Medicine*, vol. 34, no. 8, pp. 1293–1303, 2015. doi: https://doi.org/10.1002/sim.6405.

[22] A. Iparragirre, I. Barrio, and I. Arostegui, "Estimation of the ROC curve and the area under it with complex survey data," *Stat*, vol. 12, no. 1, Art. no. e635, 2023. doi: https://doi.org/10.1002/sta4.635.

[23] T. Lumley, "Analysis of Complex Survey Samples," *Journal of Statistical Software*, vol. 9, no. 8, 2004. doi: https://doi.org/10.18637/jss.v009.i08.

[24] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, no. 1, Art. no. 230, 2019. doi: https://doi.org/10.1186/s12916-019-1466-7.

[25] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015. doi: https://doi.org/10.1371/journal.pone.0118432.

[26] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007. doi: https://doi.org/10.1198/016214506000001437.

[27] A. J. Vickers and E. B. Elkin, "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006. doi: https://doi.org/10.1177/0272989X06295361.

[28] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, 2024. doi: https://doi.org/10.1136/bmj-2023-078378.

[29] K. G. M. Moons et al., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, Art. no. e082505, 2025. doi: https://doi.org/10.1136/bmj-2024-082505.

[30] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 1–7, 2024. doi: https://doi.org/10.56705/ijodas.v5i1.121.

[31] D. Anggreani, Hamdani, Nurmisba, and Lukman, "Grid Search Hyperparameter Analysis in Optimizing The Decision Tree Method for Diabetes Prediction," *Indonesian Journal of Data and Science*, vol. 5, no. 3, pp. 190–197, 2024. doi: https://doi.org/10.56705/ijodas.v5i3.190.

[32] D. H. Zulfikar, E. S. J. Atmadji, and W. Wisanti, "Leakage-Aware and Explainable Machine Learning for Healthcare Claim Fraud Detection Using Imbalanced Medical Insurance Data," *International Journal of Artificial Intelligence in Medical Issues*, vol. 4, no. 1, pp. 19–34, 2026. [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/433. [Accessed: Oct. 26, 2026].

[33] R. Puspitasari, T. Amaliah, and H. Darwis, "Comparative Machine Learning Models for Dementia Prediction Using SMOTE," *International Journal of Artificial Intelligence in Medical Issues*, vol. 3, no. 2, pp. 79–90, 2025. doi: https://doi.org/10.56705/ijaimi.v3i2.351.

[34] Purnawansyah, A. P. Wibawa, T. Widiyaningtyas, Haviluddin, H. Darwis, and H. Azis, "An in-depth exploration of supervised and semi-supervised learning on face recognition," *Open Computer Science*, vol. 15, no. 1, Art. no. 20250029, 2025. doi: https://doi.org/10.1515/comp-2025-0029.

[35] A. Latip, H. Azis, H. Himawan, and D. A. Kurnia, "SMOTE Technique Utilization in Cirrhosis Classification: A Comparison of Gradient Boosting and XGBoost," *JICO: International Journal of Informatics and Computing*, vol. 1, no. 1, pp. 34–40, 2025. [Online]. Available: https://iaico.org/index.php/JICO/article/view/5. [Accessed: Oct. 26, 2026].

[36] N. Rismayanti and S. F. Azzahrah, "Classification of Skin Diseases using Decision Tree Algorithm on an Imbalanced Dataset," *International Journal of Artificial Intelligence in Medical Issues*, vol. 2, no. 2, pp. 83–94, 2024. doi: https://doi.org/10.56705/ijaimi.v2i2.155.

[37] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 37–44, 2024. doi: https://doi.org/10.56705/ijodas.v5i1.126.

[38] J. N. K. Rao and C. F. J. Wu, "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 231–241, 1988. doi: https://doi.org/10.1080/01621459.1988.10478591.

[39] J. N. K. Rao, C. F. J. Wu, and K. Yue, "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, vol. 18, no. 2, pp. 209–217, 1992. [Online]. Available: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199200214486. [Accessed: Oct. 26, 2026].