Article

A Remote Smart Health Framework for Anemia Risk Stratification via Edge Medical Vision Systems

Sebastián A. Cruz Romero [1,2], Misael J. Mercado Hernández [1,2], Samir Y. Ali Rivera [1], Jorge A. Santiago Fernández [1] and Wilfredo E. Lugo Beauchamp [1]

[1] Department of Computer Science and Engineering, University of Puerto Rico at Mayagüez, Mayagüez, PR 00681, USA; samir.ali@upr.edu (S.Y.A.R.); jorge.santiago32@upr.edu (J.A.S.F.); wilfredo.lugo1@upr.edu (W.E.L.B.)
[2] Capicú Technologies, Mayagüez, PR 00680, USA; scruzomero@capicu.ai (S.A.C.R.), mmercado@capicu.ai (M.J.M.H
* Correspondence: scruzromero@capicu.ai (S.A.C.R.); wilfredo.lugo1@upr.edu (W.E.L.B.)

Abstract

We present an offline-first edge telemedicine platform designed for clinics and outreach programs where internet access, power, and IT support are unreliable. The system runs a local electronic health record (EHR) and clinical "plug-in" screening services on a single embedded device, accessed through a clinician-facing web app over local Wi-Fi. Data are stored locally with role-based access control and record-level encryption, while interoperability is provided as a best-effort, queued synchronization pathway to external systems using HL7 FHIR when connectivity is available. As a representative plug-in, we implement non-invasive anemia screening from fingernail photographs. Images are processed fully on-device: an INT8-quantized YOLOv8n detector extracts nail regions, lightweight color and summary-statistic features are computed per ROI and concatenated, and a supervised regressor estimates hemoglobin. On an NVIDIA Jetson Orin Nano, ROI extraction runs in 22 ms and hemoglobin inference in 34 ms. Across six training strategies (unbalanced, augmented, and KDE-balanced by remark or severity), test RMSE ranges from 2.05–3.13 g/dL; the strongest numeric performance is achieved by severity-balanced SVR (RMSE 2.048 g/dL) and remark-balanced Gradient Boosting (RMSE 2.091 g/dL). Raincloud analyses restricted to true-anemic test cases show that balancing primarily reduces systematic overestimation (which drives false negatives), while augmentation can widen error tails, highlighting the importance of selecting training strategy to match screening objectives rather than optimizing a single aggregate metric.

Keywords: Offline-first; Edge computing; Edge AI; Telemedicine; Anemia screening; Fingernail pallor

1. Introduction

Primary care delivery in low- and middle-income countries (LMICs) frequently occurs in environments where the clinical workflow is constrained more by infrastructure than by clinical intent [1]. In low- and lower-middle-income countries, close to one billion people are estimated to be served by health care facilities that either lack electricity outright (433 million people) or have an unreliable electricity supply (478 million people)[2,3]. Further, unreliable power is a major contributor to medical device failure, with WHO estimating that roughly one-third of device failures worldwide are attributable to low-quality electricity supply. [3] In such settings (e.g., community health centers, field clinics, and outreach events) a typical day includes paper-first registration, episodic triage under time pressure, intermittent device charging, ad hoc local networks (if any), and clinical documentation that must satisfy program and facility reporting requirements alongside point-of-care decision-making [4].

Digital health has expanded rapidly, yet its real-world impact is often diluted by fragmentation and weak interoperability. The WHO Digital Health Atlas (DHA) registry illustrates both scale and duplication: as of February 2023, 944 digital health projects were registered, reaching an estimated 4.74 million health workers and more than 1.56 million facilities. [5] More recent WHO materials report

Table 1. Representative recent (2020–2025) approaches adjacent to infrastructure-constrained clinical workflows, and the gap addressed by this work.

Approach	What it demonstrates (quantified)	Strengths for low-resource care	Limitations relative to our problem statement
Offline-capable clinical EHR for displaced/remote care (Hikma Health) [12]	Peer-reviewed description of an offline-first EHR used in humanitarian/underserved contexts.	Supports offline data capture and continuity of care; pragmatic focus on deployment realities.	Interoperability and clinical-grade integration remain context-specific; does not center embedded, on-premise AI screening as a modular service co-located with the EHR.
Large-scale CHW digital health intervention (OpenSRP-based DHI) [13]	Scaled in 30 months to 44 peri-urban areas in Karachi, serving >150,000 women and children.	Demonstrates that modular, workflow-oriented digital health can scale substantially in LMIC primary care.	Primarily a mobile intervention for CHWs; does not directly address an on-premise "clinic appliance" model with local EHR + local AI microservices + opportunistic standards-based sync.
Standalone fingernail-image Hb screening app (field evaluation) [14]	Clinic sample: average error magnitude 1.88 g/dL with wide limits of agreement; sensitivity 51.2%, specificity 41.6% (reported for a clinic-based sample). Retraining improved accuracy in a subset.	Quantifies the opportunity and the risk of camera-based, low-cost Hb estimation in the field.	Highlights that performance is context- and population-dependent; as a *tool*, it does not solve workflow integration, longitudinal records, security, or interoperability under constrained infrastructure.

1,328 projects across 132 countries, reinforcing the growth of the ecosystem. [6] However, capability is not equivalent to clinical integration. In the Global Digital Health Monitor (GDHM) 2023 synthesis, *Standards and Interoperability* is among the weakest enabling-environment components, with 62% of participating countries remaining at early maturity (Phase 1–2) and only 16% at advanced maturity (Phase 4–5). [7] In practice, this manifests as systems that can "capture data" but cannot reliably behave "as if you never left the clinic"—i.e., they do not preserve a coherent clinical record across sites, do not support local continuity of care when networks fail, and do not make interoperability the default.

Documentation workload further motivates an infrastructure-aware approach. In a multi-country study of primary care data recording systems, a single register entry could require $\sim$ 25 cells of information; recording time was estimated at 2–5 minutes per patient, corresponding to approximately 24–50% of a typical consultation. [8] When documentation practices vary across facilities and outreach teams, "note-taking" becomes a source of variability that directly affects downstream analytics, longitudinal follow-up, and the feasibility of integrating screening outputs into routine care. Systems that treat constraints as an afterthought tend to amplify this variability: users invent workarounds, records become incomplete, and connectivity-dependent synchronization fails to reflect what actually happened on the ground [4].

Accordingly, there is a need for *constraint-driven digital health infrastructure*: platforms that (i) run locally with minimal external dependencies[9], (ii) preserve a clinical workflow under intermittent power and connectivity[10], and (iii) treat interoperability as a first-class design objective rather than a later integration task [11]. Table 1 summarizes representative directions from 2020–2025 that partially address these needs, and highlights the remaining gap that motivates this work.

To ground the platform requirements in a clinically meaningful use case, we focus on anemia screening. Anemia remains highly prevalent globally, affecting an estimated 40% of children aged 6–59

months, 37% of pregnant women, and 30% of women aged 15–49 years. [15] In many outreach settings, suspicion of anemia begins with rapid, low-instrument heuristics (e.g., inspection of pallor, including under the fingernails), but these assessments are subjective and difficult to standardize at scale. [15] Recent field evidence from a fingernail-image smartphone Hb application further underscores both feasibility and limitation: population shift required retraining to improve accuracy, and screening sensitivity/specificity can be poor without careful contextual adaptation. [14] These observations motivate using anemia imaging not as a claim of clinical-grade diagnosis, but as a representative module that stress-tests whether an integrated, offline-first platform can host screening microservices while preserving clinical workflow and record integrity.

In this paper, our objective is to design an offline-first edge clinic platform with secure local records with queued synchronization, and to evaluate end-to-end feasibility using an embedded anemia-screening plug-in as a representative workload (runtime, model error, and error behavior on true-anemic cases under alternative training policies). We present a framework designed for infrastructure-absent environments, where cloud connectivity is treated as an opportunistic enhancement rather than an assumption. The platform consolidates: (i) local clinical data management and role-based access; (ii) encrypted on-device storage; (iii) a clinician-facing web interface accessible over local Wi-Fi; and (iv) asynchronous standards-based synchronization using HL7 FHIR when connectivity becomes available. [16] An embedded medical vision service (fingernail-image anemia risk stratification) is included as a modular demonstration of how the platform can host screening tools without outsourcing computation or data custody to the cloud.

Contributions.

This work makes three systems-engineering contributions: (1) an offline-first, clinic-oriented edge platform architecture that preserves day-to-day screening workflow under intermittent power/connectivity while maintaining secure local records; (2) an interoperability-first synchronization design based on asynchronous HL7 FHIR exchange to support "as-if-in-clinic" continuity when network access returns; and (3) an end-to-end embedded screening microservice deployed on an edge GPU device, used as a representative workload to quantify feasibility (latency and operational overhead) and to illustrate how screening tools can be integrated as modular services rather than standalone apps.

2. Materials and Methods

2.1. System Architecture and Design Objectives

We designed the platform as a layered, offline-first edge system for clinics and outreach settings with intermittent connectivity and limited IT support. The edge device hosts: (i) local EHR storage and clinical services (patient registration, record retrieval, reporting), and (ii) modular screening microservices that perform on-device inference and attach structured outputs to the patient record. A clinician-facing web application connects locally (Wi-Fi) to the edge device. When connectivity is available, the system performs asynchronous, best-effort synchronization through a dedicated interoperability gateway service (Section 2.10). We treat anemia screening as one representative screening microservice within a general module contract: acquire sample(s), run an on-device screening pipeline, and write standardized outputs (numeric estimate + categorical risk label + linked artifacts) into the longitudinal record without modifying the core EHR layer. This separation enables additional point-of-care modules (e.g., dermatology, wound monitoring) to be integrated through the same interface.

2.2. Edge Hardware and Software Stack

All experiments and deployment benchmarking were conducted on an NVIDIA Jetson Orin Nano developer kit (embedded ARM CPU + CUDA-capable GPU, 8 GB unified memory). The edge runtime hosts containerized services for the EHR backend, screening microservices, and the web application.

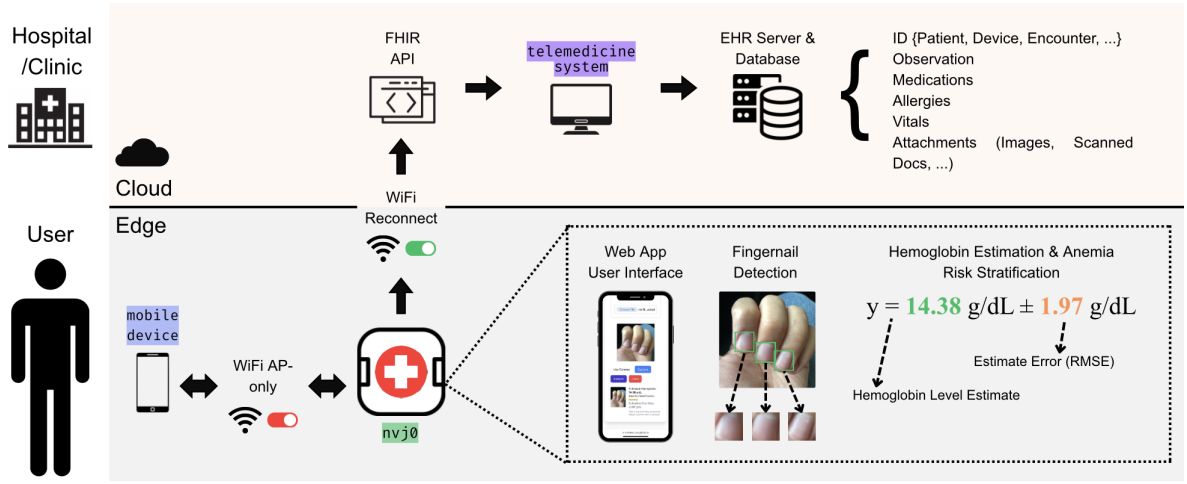


Figure 1. Constraint-aware, offline-first edge telemedicine system with best-effort interoperability: a healthcare professional uses a local web UI on a mobile device connected via Wi-Fi to an edge device that hosts encrypted EHR storage, clinical services (registration/record retrieval), and modular on-device screening microservices. The anemia screening service shown converts captured fingernail images into hemoglobin estimates with uncertainty (e.g., RMSE) and anemia risk labels, which are stored as structured observations and linked image artifacts in the patient record; when connectivity is available, queued data synchronize asynchronously through an HL7 FHIR API to cloud telemedicine services and downstream facility EHR systems.

The screening stack supports (i) a detector exported to an accelerated runtime (e.g., ONNX/TensorRT) and (ii) a classical ML estimator executed locally for hemoglobin prediction.

2.3. Data Source, Splits, and Learning Targets

We use a public fingernail-image hemoglobin dataset (250 cases) from Yakimov et al. [17] with per-case hemoglobin values and derived clinical labels. The dataset was balanced by gender and contained 128 male and 122 female patients with an age demographic that as on average 56 ± 20 y.o. spanning from 18 to 95 y.o. [17] The quality of the haemoglobin reference values was ensured by following standard clinical blood-collection protocols performed by qualified medical personnel [18,19]. Venous-blood Hb was measured using a certified haematology analyser in the ISO 15189:2012–standardized clinical diagnostic laboratory at City Clinical Hospital No. 67 (Moscow, Russia) [17].

Imaging was conducted in an enclosed, light-isolated box under constant illumination to eliminate ambient light. Camera exposure time and white-balance settings were adjusted so that reflected intensities from the nail and surrounding skin lay within the camera's usable dynamic range; specifically, the nail plate and skin regions of interest avoided pixel values near saturation (255) or near zero [17].

Each case includes RGB images and pre-specific ROI metadata (bounding boxes for nail and skin regions). We used a fixed-patient level train/test split stored as row indices in [name of json file] (Section 2.8); this file enumerates training and test row patient IDs per strategy variant to ensure reproducibility across balancing/augmentation conditions.

Learning targets include: (i) hemoglobin regression ($y \in \mathbb{R}$, g/dL), and (ii) triage labels derived from hemoglobin (binary anemia and/or ordinal severity) for reporting sensitivity/specificity at clinically relevant thresholds.

2.4. Region-of-Interest (ROI) Extraction

We define a consistent set of ROI "slots" per case: three nail ROIs, three skin ROIs, plus a small fixed "white reference" patch extracted from a predetermined region of each image. ROI slots provide a fixed input schema for feature computation and downstream classical ML. At evaluation time we support an on-device detector (YOLOv8n) [20] to predict nail ROIs; predicted crops are mapped into the same ROI-slot schema used by the feature pipeline.

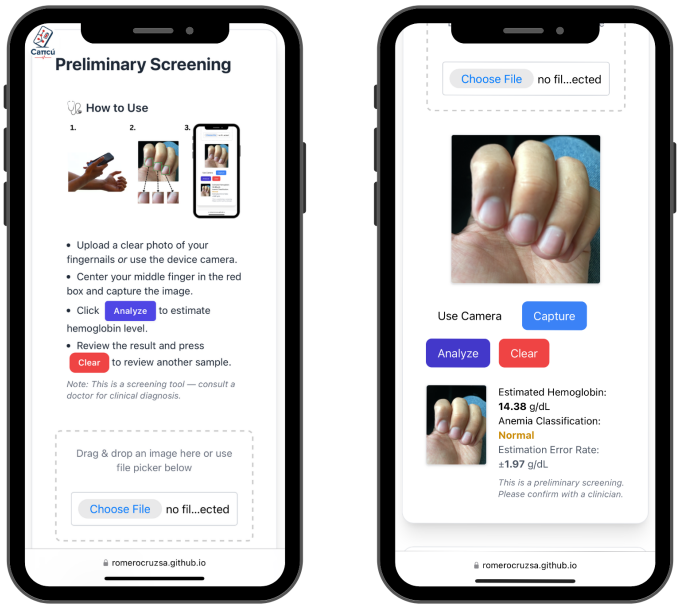


Figure 2. Mobile web interface for preliminary fingernail-image anemia screening: the left screen provides step-by-step instructions and an upload/capture entry point, while the right screen shows the live capture workflow and the returned screening output (estimated hemoglobin with uncertainty and an anemia classification) alongside the input image preview.

2.5. Feature Construction (Per-ROI, Concatenated Across Slots)

Features are computed *per ROI slot* and concatenated across slots into a fixed-length vector. Let P_r denote the cropped patch for ROI slot r (one of 6 slots: nails and skin), and let W denote the `WHITE REF` patch. We compute:

- Per-image reference RGB medians: $\tilde{w}c = \text{median}(W_c)$ for channel $c \in R, G, B$.
- Per-ROI mean RGB: $\mu r, c = \text{mean}((P_r)c)$.
- Channel-wise normalized means: $xr, c = \mu_{r,c} / (\tilde{w}c + \epsilon)$.

The final feature vector is the concatenation across ROI slots:

$$x = \left[x_{\texttt{NAIL 1}}, x_{\texttt{NAIL 2}}, x_{\texttt{NAIL 3}}, x_{\texttt{SKIN 1}}, x_{\texttt{SKIN 2}}, x_{\texttt{SKIN 3}}\right] \in \mathbb{R}^{18}, \quad (1)$$

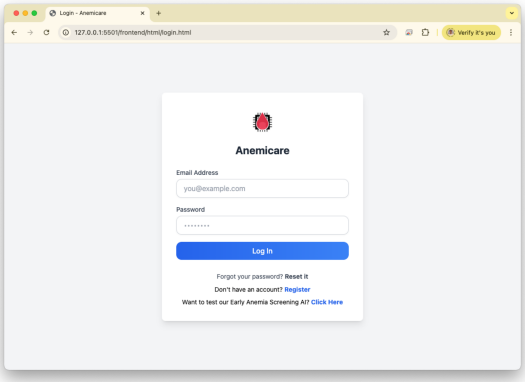
where each $x_r \in \mathbb{R}^3$ corresponds to the normalized (R,G,B) means for slot r. For cases with multiple views, we compute x per image and aggregate to case-level features by averaging across views.

2.6. Patch-Level Augmentation

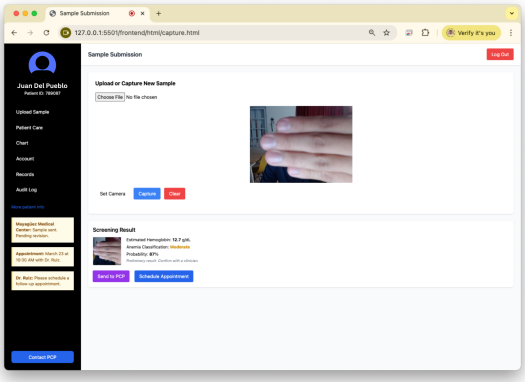
To increase effective sample diversity given limited dataset size, we augment *cropped patches* (nail/skin ROIs and reference patch) using a set of photometric and mild geometric transforms designed to preserve anatomical plausibility. The training pipeline applies: resize and crop to a canonical input size, horizontal flips (no vertical flips), color jitter (brightness/contrast/saturation/hue), small affine perturbations (rotation/translate/scale/shear), occasional perspective distortion, and low-variance Gaussian noise. Augmentation is applied by generating $n_{\text{aug}} = 2$ augmented copies per original patch while retaining the original (total factor $\approx 3\times$ per case).

2.7. Distribution-Aware Balancing via KDE Downsampling

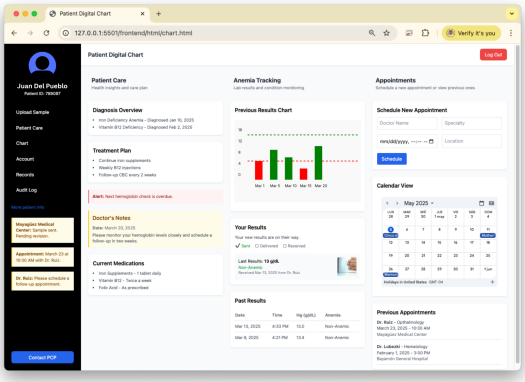
Hemoglobin distributions are typically skewed toward non-anemic values, risking poor learning signal in clinically important low-Hb ranges. We evaluate KDE-based balancing that uses kernel



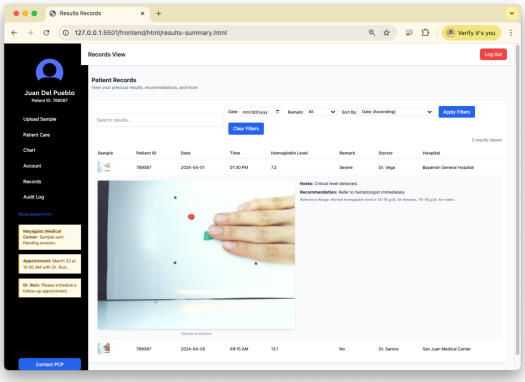
(**a**) Login.



(**b**) Sample submission.



(**c**) Records summary.



(**d**) Patient Chart.

Figure 3. Key point-of-care user interfaces: (a) authenticated login; (b) sample submission with real-time capture and on-device AI screening output; (c) records summary including hemoglobin results and visual analysis; (d) patient chart with results, plan, appointments, and history.

density estimation to compute sample inclusion probabilities and then *downsample* to produce a flatter effective training distribution within each label grouping.

We evaluate six data strategies:

1. Unbalanced (raw).
2. Unbalanced + augmentation.
3. Balanced by Remark (raw) via KDE downsampling.
4. Post-augmentation then balance by Remark via KDE downsampling.
5. Balanced by Severity (raw) via KDE downsampling.
6. Post-augmentation then balance by Severity via KDE downsampling.

Concretely, for a chosen grouping variable g (Remark or Severity), we estimate a KDE over hemoglobin values within each group and assign inclusion probabilities inversely proportional to the estimated density; we then sample to a target per-class size (scaled appropriately in post-augmentation balancing). This yields balanced training subsets without synthetic label generation and allows us to explicitly compare "balance before augmentation" vs. "augment then balance".

2.8. Classical ML Models and Training Protocol

We train hemoglobin estimators on the fixed 18-D feature representation using classical regression models (including tree ensembles, linear baselines, and kernel methods). For each of the six data

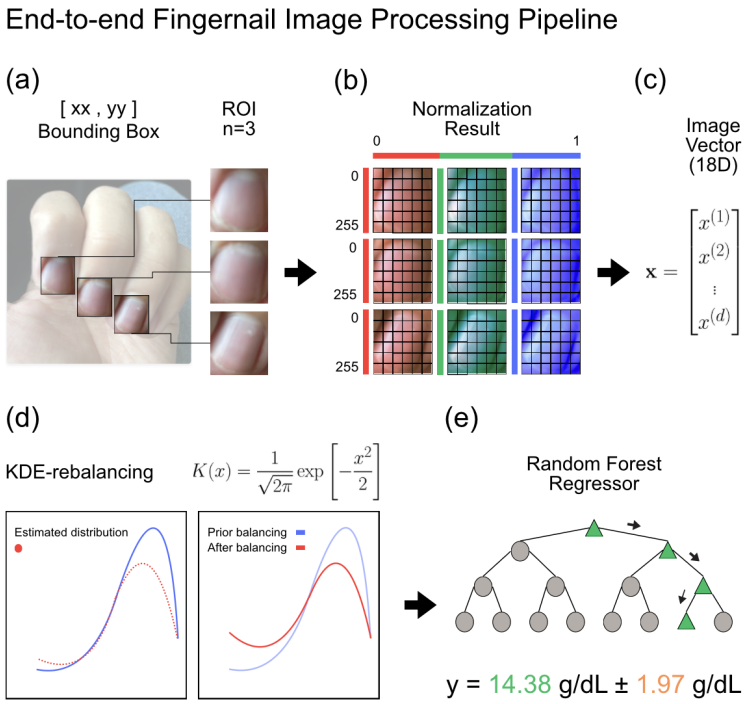


Figure 4. End-to-end hemoglobin estimation pipeline from fingernail images: (a) detected nail-bed bounding boxes are cropped into $n = 3$ regions of interest, (b) each normalized and converted into a (c) fixed-length feature vector. The training distribution is then rebalanced via (d) kernel density estimation (KDE) to emphasize clinically relevant hemoglobin ranges before (e) fitting a ML model (e.g., a Random Forest regressor), which outputs a hemoglobin estimate with an associated error metric (e.g., $y = 14.38$ g/dL $\pm$ 1.97 g/dL).

strategies (Section 2.7), we perform K-fold cross-validation ($K = 7$) on the training split to tune hyperparameters using RMSE as the primary selection metric. For each strategy we:

1. Run cross-validated hyperparameter search (RMSE-based) on the training portion.
2. Fit the best model on the full strategy-specific training set.
3. Evaluate once on the held-out test set.

Model selection on CV. Within this revision, we adopt a simple, reproducible selection rule: choose the (strategy, model, hyperparameters) triple that minimizes mean CV RMSE on the training split; ties within a small tolerance are broken in favor of lower inference cost (smaller ensemble / simpler estimator) to match edge constraints. (More elaborate selection criteria—e.g., Hb-bin subgroup constraints and calibrated uncertainty—are deferred.)

2.9. On-Device Inference Definition and Runtime Benchmarking

We define **one screening inference** as the full on-device pipeline from a captured RGB image to a stored hemoglobin estimate and triage label:

1. ROI extraction (detector inference, crop assignment into ROI slots).
2. Feature computation (per-ROI normalized mean RGB, concatenation to 18-D vector).
3. Hemoglobin prediction (classical ML regressor) and optional triage labeling.
4. Write-back to the local EHR (structured observation + linked artifacts).

On the Jetson Orin Nano [21], the detector stage runs locally with mean latency of **22 ms** per image for ROI extraction, and the hemoglobin prediction stage runs locally with mean latency of **34 ms** per case for feature computation + regression output. These timings exclude clinician interaction time and represent compute latency under typical screening operation.

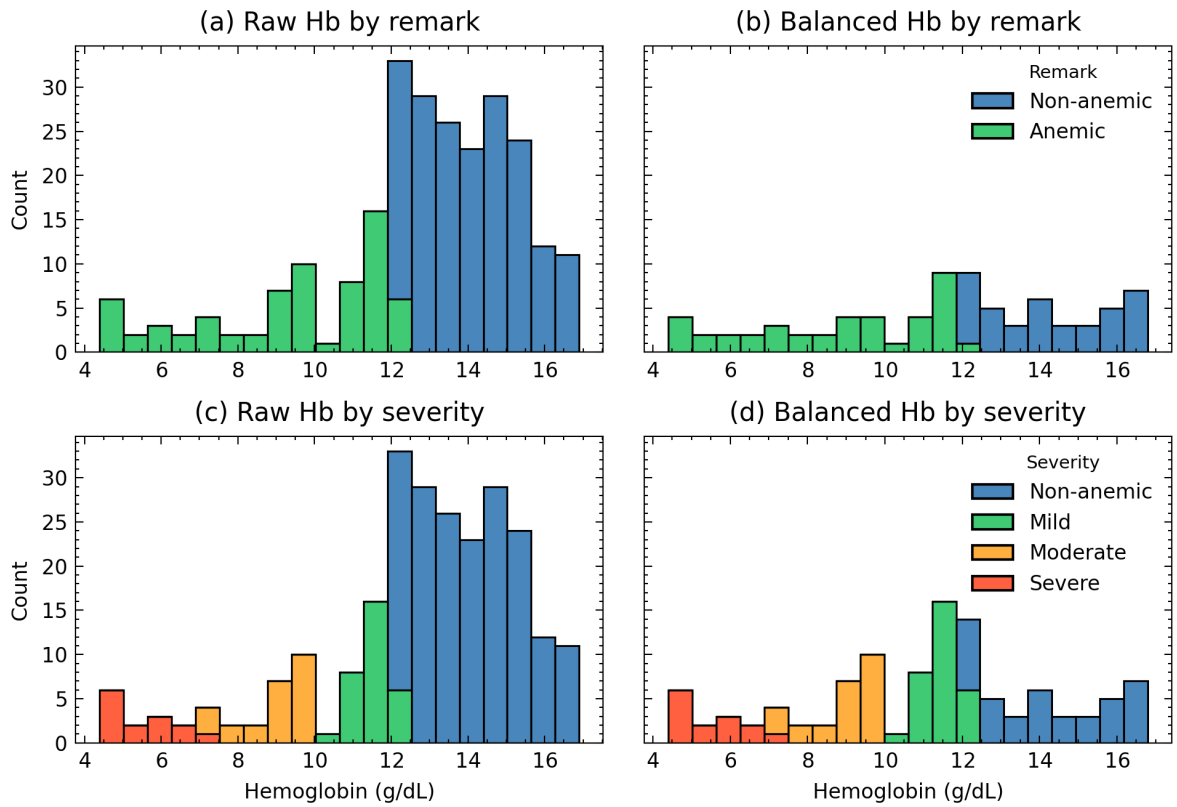


Figure 5. Hemoglobin label distributions before and after KDE-based rebalancing. (a) Raw (unbalanced) hemoglobin (Hb) histogram grouped by binary remark (non-anemic vs. anemic). (b) Hb histogram after KDE-guided rebalancing by remark, increasing representation of lower-Hb/anemic samples relative to the dominant non-anemic mode. (c) Raw Hb histogram grouped by anemia severity (non-anemic, mild, moderate, severe). (d) Hb histogram after KDE-guided rebalancing by severity to reduce class imbalance across clinically defined Hb strata. Axes report Hb (g/dL) versus sample count.

2.10. Offline-First EHR, Security, and Interoperability Scope

Scope. We implement a minimal, intuitive security and interoperability scope intended to mirror a standard EHR deployment pattern: a modular service boundary, REST-first local access, RBAC-based access control, and optional best-effort synchronization through a dedicated gateway.

Local patient record model The local EHR stores (i) patients, (ii) encounters/visits, and (iii) observations produced by screening microservices. Each screening result is stored as a structured observation (e.g., hemoglobin estimate, units, timestamp, model version, and a triage label) and linked to image artifacts (raw capture and/or ROI visualization).

Base security implementation We enforce authenticated access to the clinician UI and API using role-based access control (RBAC). Roles (e.g., nurse, clinician, admin) map to permissions on operations (create/read/update/export). For data-at-rest protection on the edge device, sensitive record fields and stored artifacts are encrypted using authenticated encryption (AES-GCM) with per-record nonces and integrity tags. Keys are not stored in plaintext; the runtime unlocks an encrypted keystore at initialization for authorized use.

Base interoperability implementation The edge device exposes a local REST API used by the web user interface and screening services. For best-effort interoperability, a separate gateway service transforms a minimal subset of local entities (Patient, Encounter, Observation, and DocumentReference-like image links) into a standards-shaped payload for downstream systems. When connectivity is present, the gateway uploads queued records asynchronously; when connectivity is absent, records remain locally durable and the gateway retries later. This “store-and-forward” design preserves offline operation while enabling eventual consistency when networks permit.

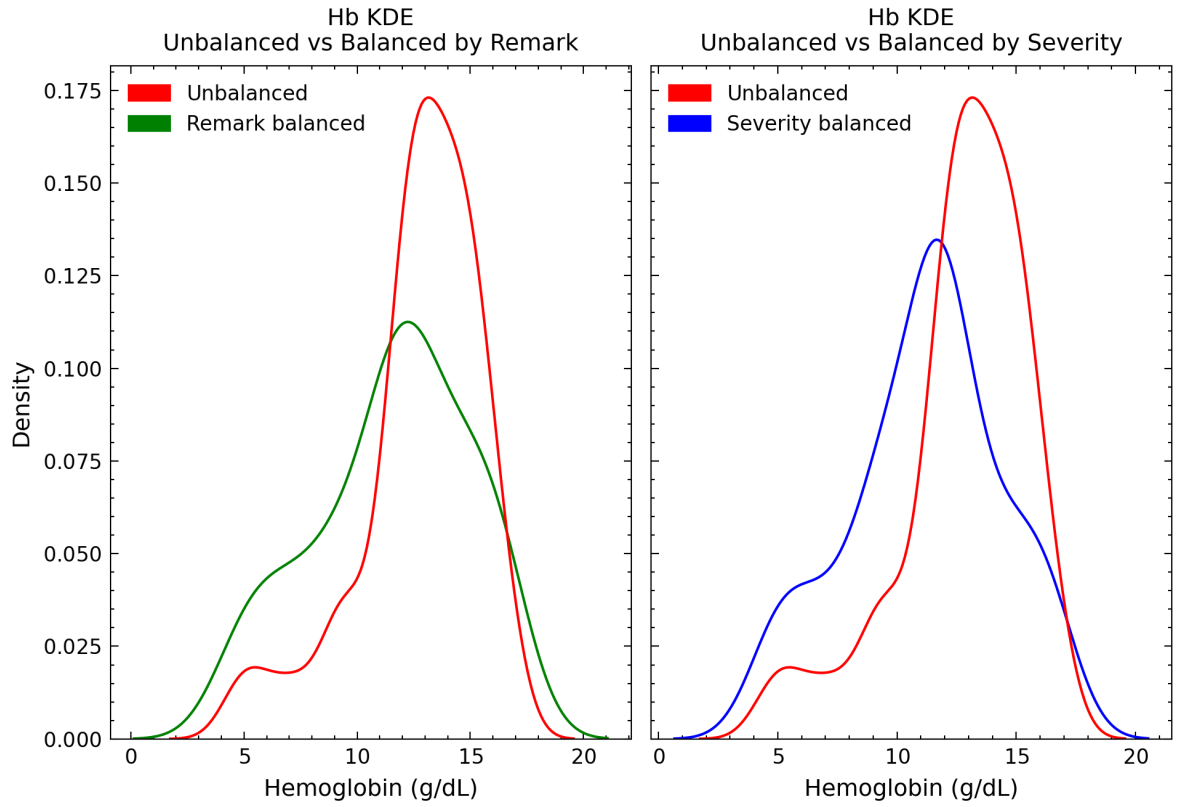


Figure 6. Kernel density estimates (KDEs) of Hb distributions for unbalanced vs. balanced sets. (Left) KDE of the unbalanced dataset compared to the remark-balanced dataset, illustrating redistribution toward the lower-Hb tail when balancing by anemia remark. (Right) KDE of the unbalanced dataset compared to the severity-balanced dataset, illustrating redistribution across severity-defined Hb ranges. Axes report Hb (g/dL) versus estimated density.

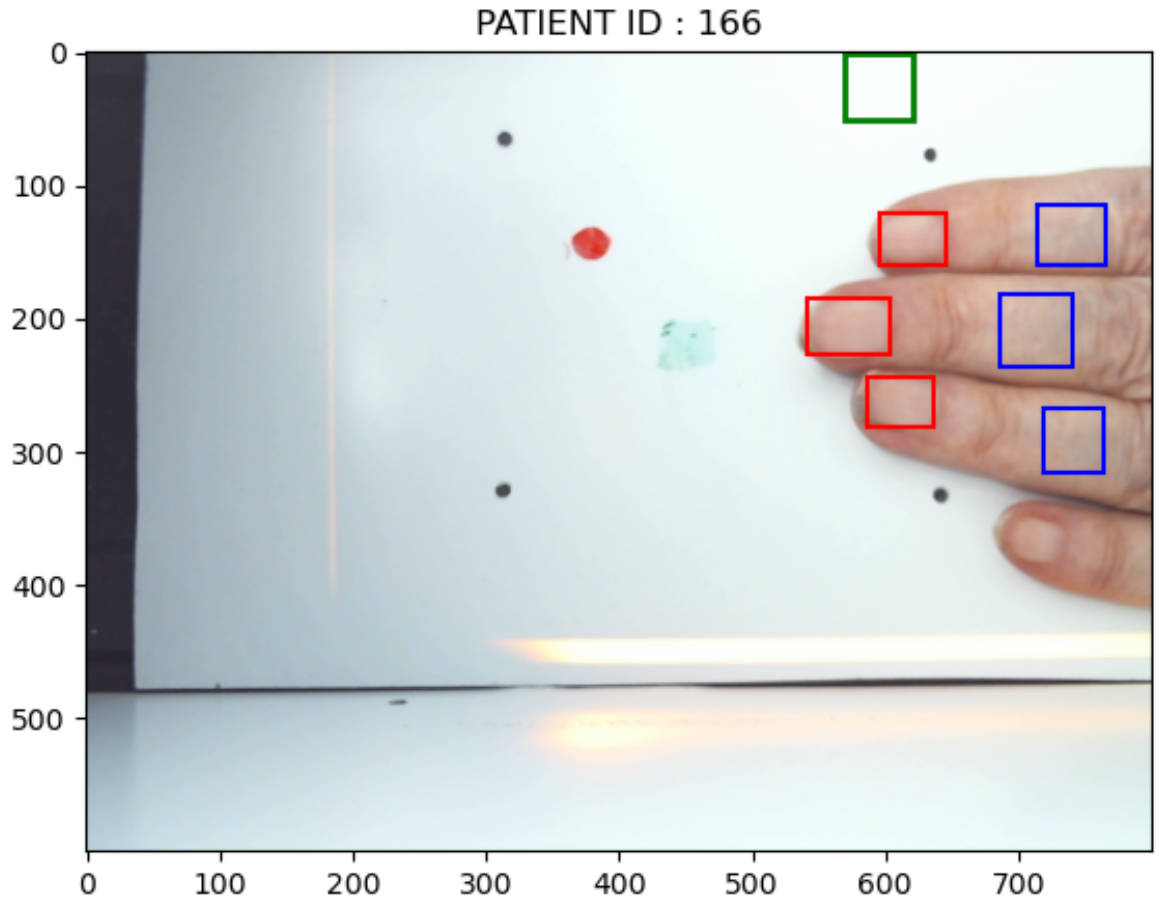


Figure 7. Example ROI slotting for fingernail patch extraction. A representative hand image with detected nail regions overlaid as fixed "ROI slots" (color-coded boxes). Detected/assigned slots are cropped to produce per-ROI patches used for feature computation; missing slots are handled consistently (e.g., zero-padding or masked features) so that a fixed-length feature vector can be formed by concatenating features across ROI slots. This visualization illustrates the detection-to-feature "contract" used by the screening microservice.

2.11. User Interface and Clinical Workflow

The clinician-facing interface is a local web application served by the edge device. A typical encounter proceeds as: patient registration or lookup, image acquisition guided by capture instructions, on-device screening, results review (numeric estimate + triage label), and record update. When enabled and when connectivity is available, the gateway synchronizes queued de-identified summaries or patient records according to deployment policy.

2.12. Evaluation Metrics and Reporting

We report standard regression metrics (MAE, RMSE) for hemoglobin estimation on the held-out test set. For triage performance, we report sensitivity and specificity at clinically relevant anemia thresholds and include confusion matrices for binary and ordinal labelings. For localization, we report detector performance (e.g., mAP at a chosen IoU threshold) and measure ROI extraction latency as part of the end-to-end screening pipeline.

3. Results

We evaluated the proposed offline-first edge EHR platform and its integrated anemia screening service using (i) detection accuracy and on-device latency for fingernail ROI localization, (ii) hemoglobin (Hb) regression accuracy under multiple training-data strategies, and (iii) end-to-end runtime of the deployed, fully local screening pipeline. All Hb results are reported in g/dL. The

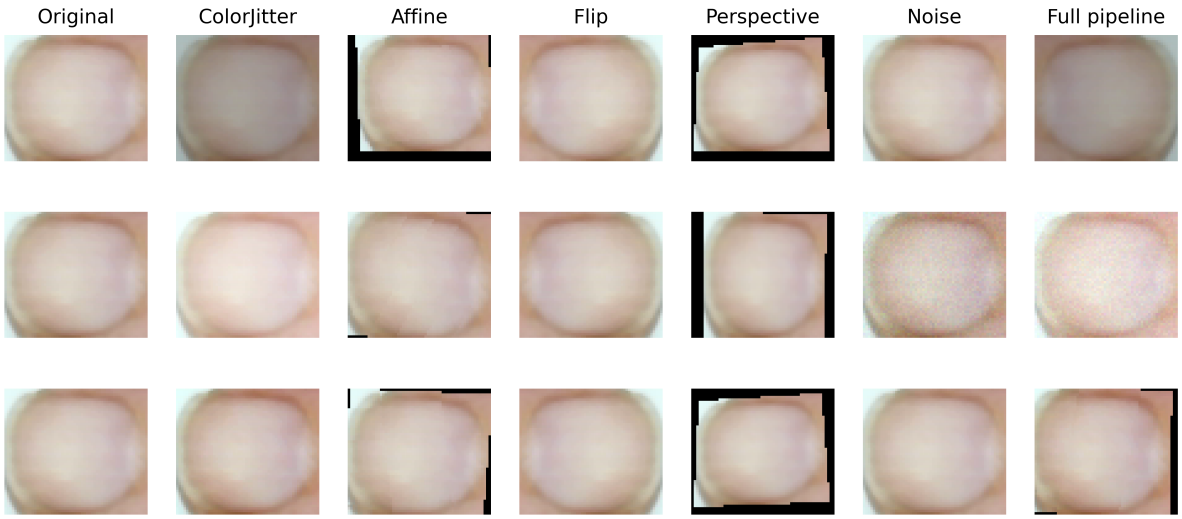


Figure 8. Examples of the cropped ROI augmentation pipeline used for model training. Each row shows a representative nail ROI patch; columns visualize individual transforms applied during training augmentation (ColorJitter/photometric perturbation, Affine transform, Horizontal/Vertical Flip, Perspective warp, and additive Noise), followed by the combined "Full pipeline" output. These augmentations are intended to simulate real-world variability in capture conditions (illumination, pose, and sensor noise) while preserving the underlying anatomical ROI.

evaluation uses the public fingernail/skin image dataset described in Section 2, with a fixed held-out test partition of $N_{\text{test}} = 50$ cases and strategy-specific training partitions.

3.1. Nail-bed localization accuracy and edge optimization

3.1.1. Detection quality.

The YOLOv8n [20] object detection model was succesfully used to localize fingernail regions for pallor analysis. After applying post-training quantization (PTQ) [22] to 8-bit integers, the model's inference latency was reduced by 54.2% from 46.96 ms down to 21.50 ms) while preserving detection accuracy (mAP@0.5 remained 0.995). The quantized model's precision and recall were 0.9999 and 1.000, respectively, virtually identical to the full 32-bit model (Table 2). This optimization effectively halved the model's memory footprint without degrading performance, confirming the feasibility of deploying the detector on edge hardware.

3.1.2. Quantization and on-device latency.

To support real-time use on embedded hardware, we exported the detector and deployed an INT8-quantized engine. Quantization reduced detector latency from 46.96 ms (FP32) to 21.50 ms (INT8), a 54.2% reduction, while preserving mAP@0.5 (0.995) and near-identical precision/recall (Table 2). These results confirm that the ROI localization stage can be executed locally within interactive workflows.

Table 2. YOLOv8n detector performance and latency before and after INT8 post-training quantization (PTQ).

Metric	FP32	INT8 (PTQ)
Precision	0.993	0.999
Recall	1.000	1.000
mAP@0.5	0.995	0.995
mAP@0.5:0.95	0.692	0.638
Model artifact size	6.2 MB	6.2 MB
Inference latency	46.96 ms	21.50 ms

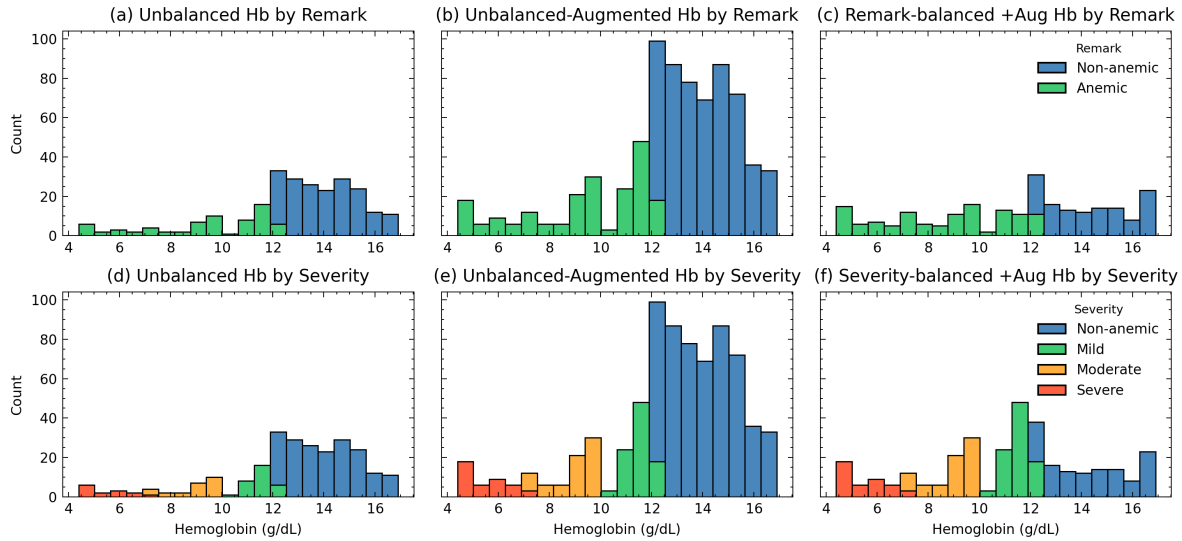


Figure 9. Post-augmentation label distributions across evaluated dataset strategies. Top row: Hb histograms grouped by remark for (a) unbalanced, (b) unbalanced + augmentation, and (c) remark-balanced after augmentation (rebalancing performed on the augmented pool). Bottom row: Hb histograms grouped by severity for (d) unbalanced, (e) unbalanced + augmentation, and (f) severity-balanced after augmentation. This figure summarizes how augmentation changes sample density and how post-augmentation KDE balancing restores targeted class/severity representation. Axes report Hb (g/dL) versus sample count.

3.2. *Hemoglobin estimation accuracy across data strategies*

We compared multiple classical regressors (Elastic Net, Ridge, Lasso, Random Forest, Gradient Boosting, SVR, Huber, RANSAC) under six training-data strategies reflecting the pipeline decisions in Section 2: *unbalanced*, *unbalanced_aug*, *remark_balanced*, *remark_balanced_aug*, *severity_balanced*, and *severity_balanced_aug*. For each strategy, hyperparameters were selected using cross-validation on the constructed training set, and the final model was evaluated once on the fixed held-out test set.

3.2.1. Overall "best"-performing configuration by lowest RMSE.

The lowest test-set RMSE was obtained by a Random Forest trained on the *unbalanced* (no KDE downsampling) training distribution, achieving a test MAE of 1.493 and RMSE of 1.881 (Table 3). This setting preserves the largest effective training set size, which appears to benefit aggregate regression accuracy.

3.2.2. Effect of KDE balancing.

KDE-based downsampling to construct more uniform Hb coverage (*remark_balanced* or *severity_balanced*) produced test RMSE values in the 2.05–2.18 range for the best models (SVR, Gradient Boosting, or Random Forest depending on the strategy). In other words, balancing introduces a modest increase in global RMSE compared with *unbalanced*, consistent with a trade-off: these strategies deliberately reduce the dominance of high-prevalence Hb ranges in training.

3.2.3. Effect of patch augmentation in this pipeline.

Across all three base distributions (*unbalanced*, *remark_balanced*, *severity_balanced*), adding augmentation on cropped ROIs increased test error for the best model within each family. For example, the best *unbalanced_aug* model increased test RMSE from 1.881 to 2.062 (+0.18), and *severity_balanced_aug* increased test RMSE from 2.048 to 2.719 (+0.67). This suggests that, for the current engineered-feature representation and small-data regime, naive ROI augmentation can introduce distribution shift that does not translate into improved regression generalization.

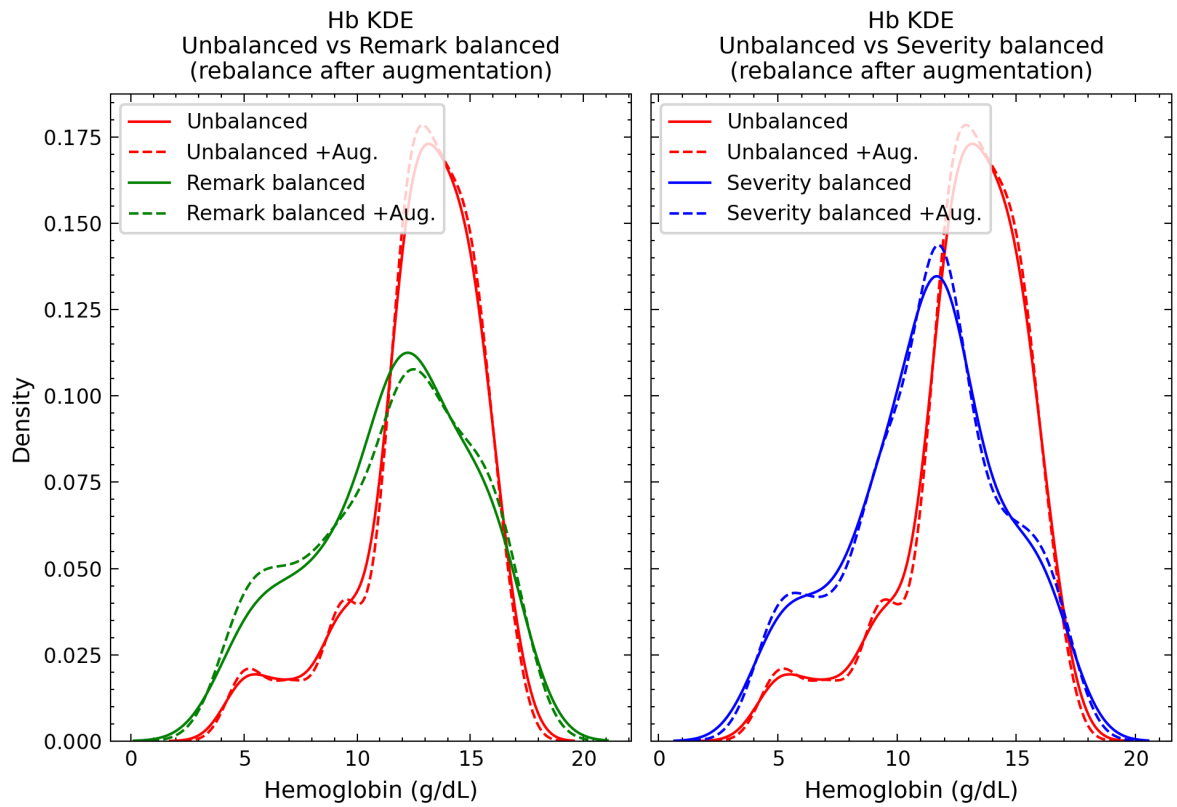


Figure 10. KDE comparison including post-augmentation rebalancing. (Left) Hb KDEs for unbalanced, unbalanced + augmentation, remark-balanced, and remark-balanced + augmentation (where "+Aug." denotes augmentation applied prior to rebalancing). (Right) Corresponding Hb KDEs for unbalanced, unbalanced + augmentation, severity-balanced, and severity-balanced + augmentation. Solid vs. dashed styling distinguishes pre- vs. post-augmentation distributions, emphasizing the effect of applying KDE balancing after augmentation to shape the final training distribution.

3.3. Error Distributions on True-Anemic Test Cases (Raincloud Analysis)

To complement aggregate MAE/RMSE scores, we examined the distribution of hemoglobin prediction errors restricted to the *true-anemic* subset of the held-out test set. For each regression model and training-data strategy, we compute the signed prediction error

$$e = \hat{y} - y \quad (\mathrm{g/dL}), \tag{2}$$

where y is the laboratory hemoglobin value and $\hat{y}$ is the model estimate. In this subset, large *positive* errors indicate hemoglobin *over-estimation* on anemic individuals, while negative errors indicate underestimation. Figures 13 and 14 summarize the error distributions using raincloud plots (density + boxplot + per-sample scatter); red "×" markers denote *false negatives* for anemia classification induced by the screening rule (i.e., predicted as non-anemic despite ground-truth anemia).

3.3.1. Consistent right-shift under unbalanced training.

Across all eight models, the **Unbalanced** strategy exhibits a pronounced right-shift (median error typically > 0), with a visible cluster of red "×" markers at positive errors. This indicates that, on true-anemic test cases, models trained on the original skewed distribution tend to *overestimate* hemoglobin, and these overestimates are frequently large enough to flip the downstream anemia decision into a false negative.

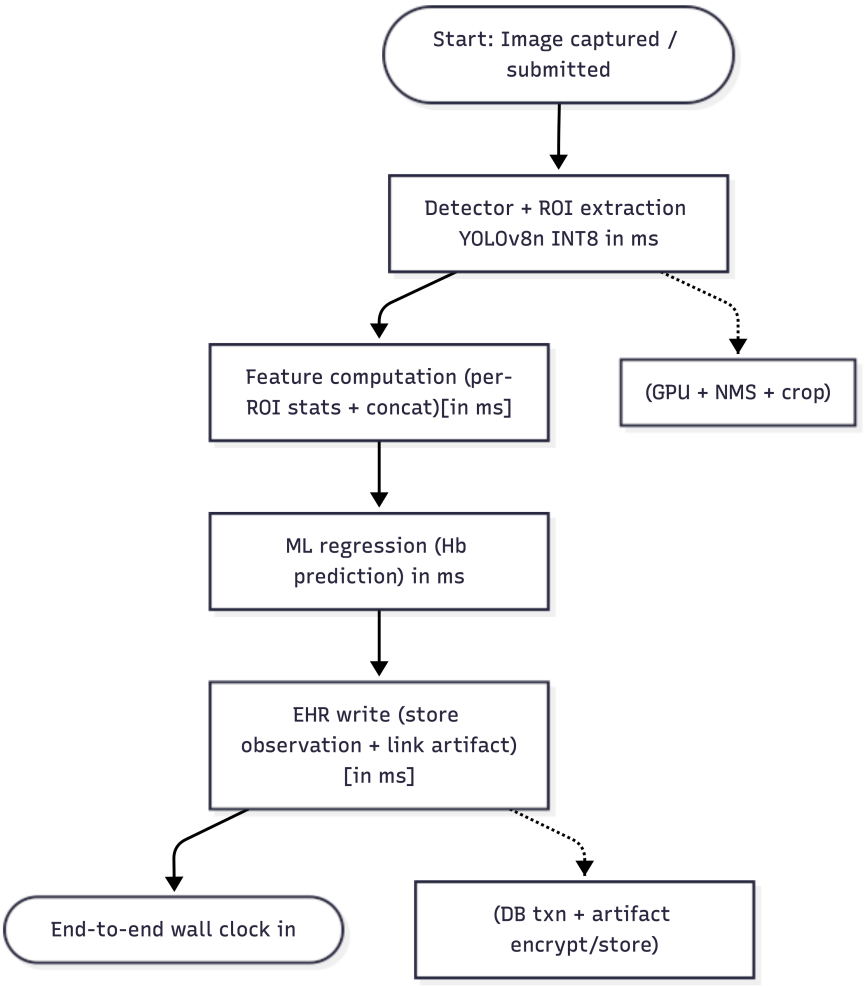


Figure 11. Runtime breakdown ("waterfall") for a single on-device screening transaction, from image submission through structured write-back. The pipeline is decomposed into (i) detector/ROI extraction, (ii) feature computation, (iii) ML regression for hemoglobin estimation, and (iv) EHR write (including linking encrypted artifacts), with the final bar representing end-to-end wall-clock time. Dashed callouts indicate internal sub-steps commonly dominated by GPU inference/NMS/cropping in the detector stage and database transaction plus artifact encryption/storage in the write-back stage

3.3.2. Balancing shifts errors toward zero and reduces the positive tail.

Both **Remark Balanced** and **Severity Balanced** strategies generally shift the center of the error distribution leftward (closer to $e = 0$) relative to Unbalanced, and reduce the extent of the right tail. This pattern is especially clear in the ensemble learners (Gradient Boosting, Random Forest), where the boxplots contract and move toward zero under balanced strategies. False negatives persist in several models, but they are less concentrated in the extreme positive-error region compared to Unbalanced.

3.3.3. Augmentation alone does not consistently correct overestimation on anemic cases.

The **Unbalanced Aug.** strategy typically remains right-shifted and continues to show multiple false negatives across models, suggesting that augmentation without rebalancing does not systematically eliminate overestimation on true-anemic cases. In several panels, Unbalanced Aug. retains a similar median error to Unbalanced and can preserve (or only modestly reduce) the positive-error spread.

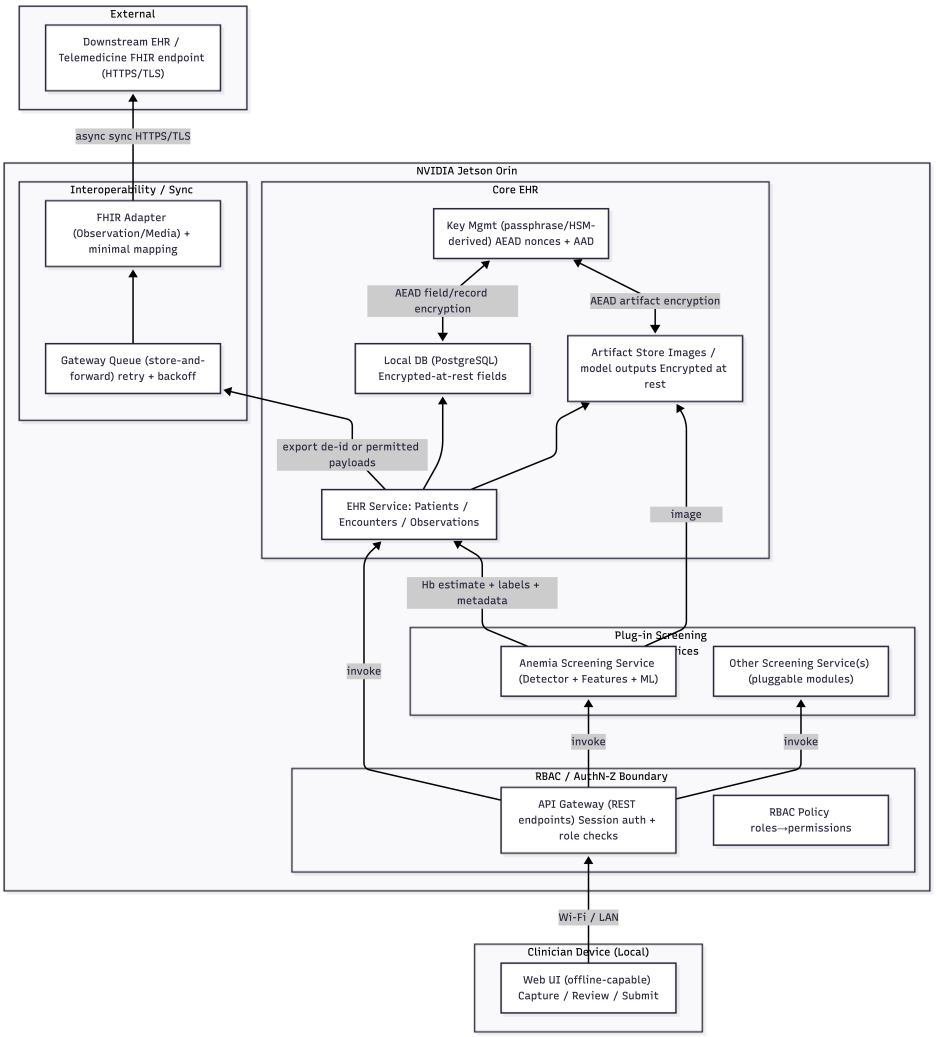


Figure 12. Security and interoperability architecture for an offline-first edge EHR with plug-in screening microservices. A clinician-facing web UI connects over a local network to an API gateway that enforces authentication and role-based access control (RBAC). The core EHR services persist structured patient data and linked image/model artifacts in local stores protected by authenticated encryption at rest; screening modules (e.g., anemia) run as separate, pluggable services and write back results as observations plus linked artifacts. Interoperability is implemented as a store-and-forward gateway queue and a lightweight FHIR adapter, enabling best-effort asynchronous synchronization to an external endpoint over HTTPS/TLS when connectivity is available.

3.3.4. Post-augmentation balancing yields the most conservative error profiles in most models.

The two post-augmentation balancing variants, **Remark Balanced Aug.** and **Severity Balanced Aug.**, most often yield the *closest-to-zero* (and in several cases slightly negative) median errors on true-anemic cases, while also truncating the positive tail. In many models (e.g., ElasticNet, Huber, Random Forest, RANSAC, Ridge), these strategies show fewer red "×" markers than Unbalanced / Unbalanced Aug., indicating fewer false negatives under the same screening rule. An exception is Support Vector Regression, where the balanced-augmented variants can exhibit noticeably wider dispersion (larger IQR), i.e., reduced bias but increased variance on the anemic subset.

3.3.5. Model-specific dispersion and outliers.

The robust and linear models (Ridge/Lasso/ElasticNet/Huber) exhibit broadly similar shifts by strategy: unbalanced training is right-biased, balancing moves toward zero, and post-augmentation balancing tends to be most conservative. RANSAC shows occasional extreme positive outliers (including high-error points that coincide with false negatives), consistent with sensitivity to small-sample id-

Table 3. Best Hb regressor per training-data strategy (selected by cross-validation on the training partition) and evaluated on the fixed test set ($N_{\text{test}} = 50$). N_{train} reflects the constructed training set size after KDE downsampling and/or augmentation.

Strategy	N_{train}	Best model	CV MAE	CV RMSE	Test MAE	Test RMSE
Unbalanced	200	Random Forest	1.817	2.294	1.493	1.881
Unbal+Aug.	600	SVR	2.098	2.601	1.607	2.062
Remark bal.	80	Gradient Boosting	2.304	2.876	1.794	2.091
Remark bal.+Aug.	240	Gradient Boosting	2.300	2.929	2.114	2.524
Severity bal.	80	SVR	1.828	2.373	1.695	2.048
Severity bal.+Aug.	240	Gradient Boosting	2.279	2.860	2.403	2.719

iosyncrasies. SVR displays the widest cross-strategy spread in several panels, with balanced strategies not always reducing dispersion even when centering improves.

Overall, the raincloud plots provide a distribution-level view of how each training strategy changes (i) *bias* (systematic right-shift vs. centering near 0) and (ii) *tail risk* (frequency/magnitude of large positive errors) on the subset of cases where anemia is truly present.

3.4. System-level validation: offline workflow and structured result capture

Finally, we validated end-to-end functionality of the offline-first workflow: patient registration and retrieval, image capture/upload via the local web UI, execution of the screening microservice, and persistence of outputs (predicted Hb and associated artifacts) into the local EHR store. When connectivity is available, the platform can optionally export and synchronize structured observations via a FHIR-compatible interface. Quantitative benchmarking of cryptographic overhead and network synchronization performance is deferred to future work; in this revision, we report on-device screening latency and predictive accuracy as the primary feasibility indicators.

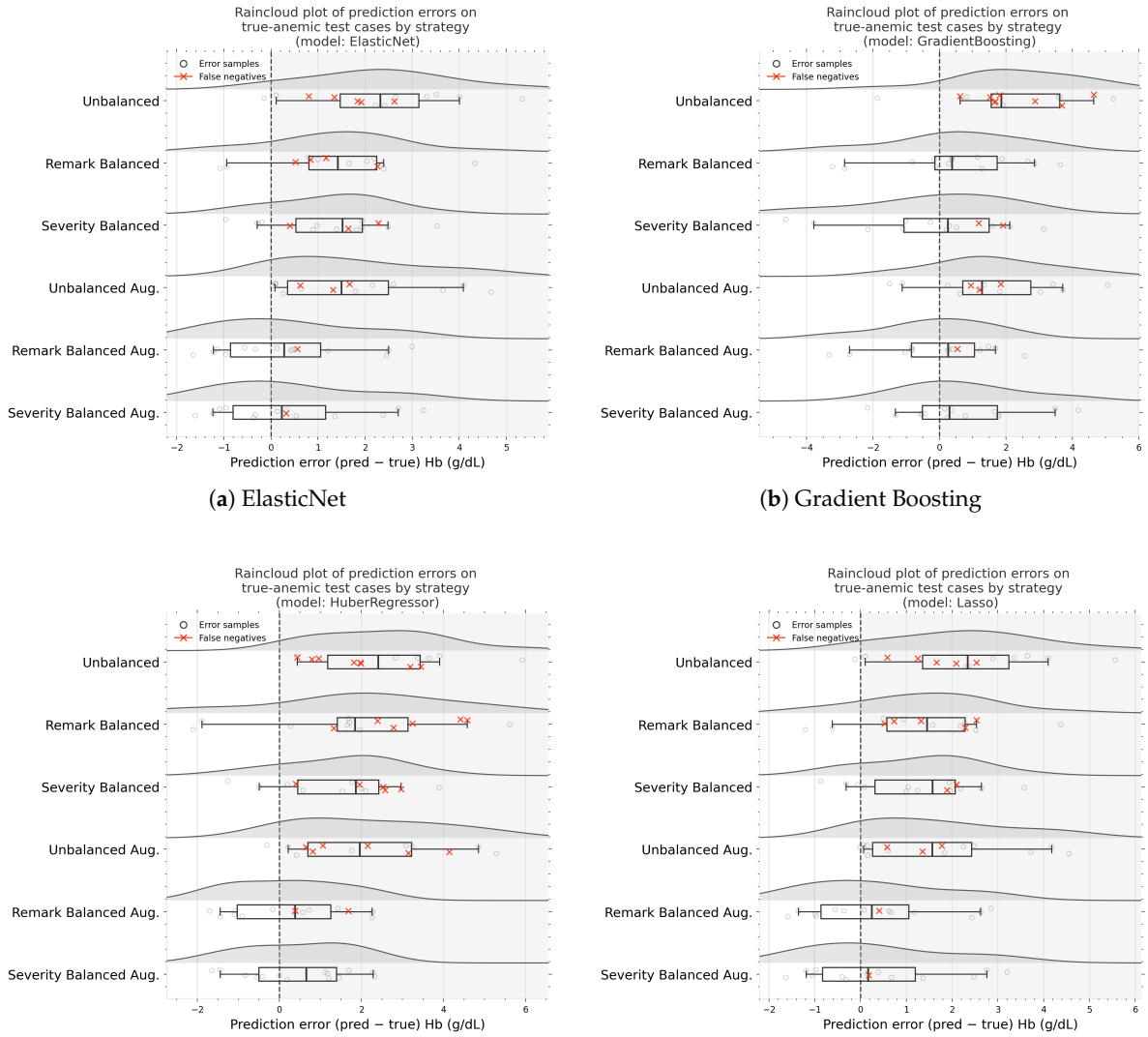
4. Discussion

This study's primary contribution is a constraint-aware, offline-first clinical platform that executes the full point-of-care workflow on a single edge device (patient registration, longitudinal record access, screening invocation, and secure local persistence) while treating cloud connectivity as an opportunistic enhancement rather than a dependency. The anemia module is presented deliberately as a *pluggable* screening microservice within this architecture: it exercises the platform's core contract (capture $\rightarrow$ on-device inference $\rightarrow$ structured write-back to the local EHR) and demonstrates how additional point-of-care modules can be integrated without redesigning the EHR or UI layers.

A key design outcome of formalizing services and constraints is that it forces systems choices to be evaluated against operational budgets, not solely against model metrics. The microservice decomposition (core EHR + API, screening services, and an optional synchronization gateway) isolates latency-critical inference from storage and interoperability concerns, and makes end-to-end performance measurable at clear service boundaries. In practice, this is what enables "offline-first" to be a verifiable property of the deployed system rather than a qualitative aspiration.

Edge feasibility depends on pipeline latency, not single-model speed.

The on-device vision stage provides a concrete example of how platform feasibility is determined by end-to-end runtime. Post-training INT8 quantization reduced nail-bed detector latency from 46.96 ms to 21.50 ms (54.2% reduction) while preserving detection quality at mAP@0.5 = 0.995, with a moderate decrease in mAP@0.5:0.95 (0.692 $\rightarrow$ 0.638). In the deployed screening pipeline, ROI extraction executes locally on average $\sim$22 ms, and the downstream hemoglobin regression stage executes on average $\sim$34 ms on the Jetson Orin Nano, supporting interactive use in clinic workflows. More generally, these results motivate treating quantization and compilation (e.g., TensorRT engines) as first-class design levers in edge clinical systems because they directly determine whether screening can be delivered within throughput constraints.



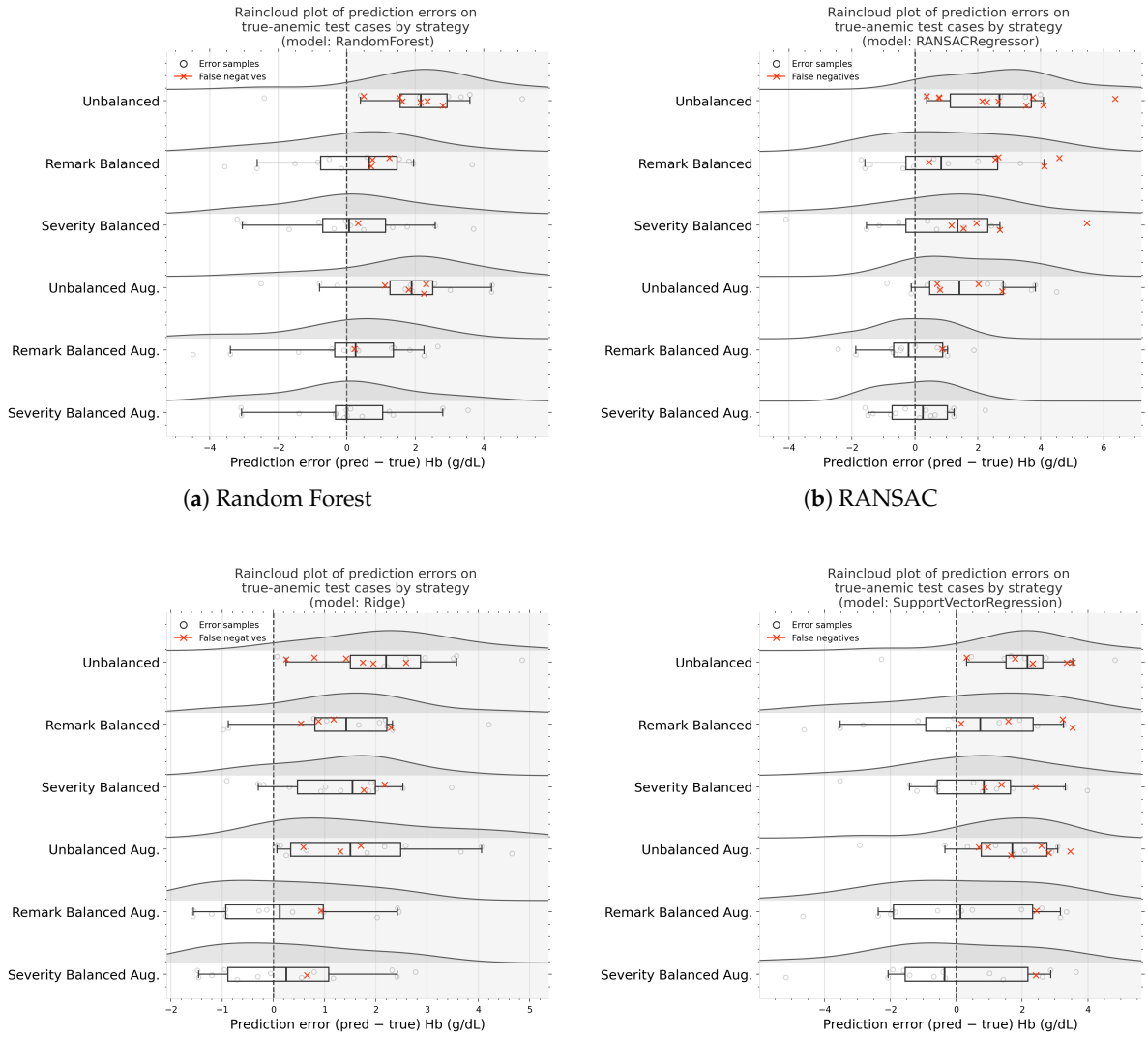
(**a**) ElasticNet (**b**) Gradient Boosting

(**c**) Huber (**d**) Lasso

Figure 13. Raincloud plots of hemoglobin prediction error $e = \hat{y} - y$ on **true-anemic** test cases, stratified by training-data strategy and models: ElasticNet, Gradient Boosting, Huber, and Lasso models. Each row corresponds to a training strategy (Unbalanced, Remark Balanced, Severity Balanced, and their augmented variants). Grey density "clouds" and boxplots summarize the error distribution; light points indicate per-case errors; red "×" markers denote false negatives under the anemia decision rule (anemic by ground truth, predicted non-anemic).

Hemoglobin estimation performance is shaped by training distribution policy.

Across the evaluated regressors and data strategies, the best-performing configurations achieved test-set RMSE values on the order of ∼2.0–2.2 g/dL. In the most recent evaluation, severity-balanced training paired with Support Vector Regression yielded the lowest reported test RMSE (2.048 g/dL; test MAE 1.695 g/dL), while remark-balanced training paired with Gradient Boosting performed comparably (test RMSE 2.091 g/dL; test MAE 1.794 g/dL). Random Forest remained competitive across both balancing regimes (e.g., test RMSE 2.178 g/dL under remark balancing; 2.175 g/dL under severity balancing), but was not consistently the top regressor in the latest sweep. These results reinforce an important operational point: balancing decisions (remark- vs. severity-driven) are not merely statistical conveniences; they are *policy knobs* that change the model's error profile in clinically relevant subpopulations.



(**a**) Random Forest (**b**) RANSAC

(**c**) Ridge (**d**) SVR

Figure 14. Raincloud plots of hemoglobin prediction error $e = \hat{y} - y$ on **true-anemic** test cases, stratified by training-data strategy and models: Random Forest, RANSAC, Ridge, and SVR models. Each row corresponds to a training strategy (Unbalanced, Remark Balanced, Severity Balanced, and their augmented variants). Grey density "clouds" and boxplots summarize the error distribution; light points indicate per-case errors; red "×" markers denote false negatives under the anemia decision rule (anemic by ground truth, predicted non-anemic).

Raincloud error plots reveal systematic behavior on true-anemic test cases.

The raincloud plots summarize prediction error distributions *restricted to true-anemic test cases* across data strategies for each model. Because the horizontal axis is $(\hat{y} - y)$, positive shifts correspond to hemoglobin *overestimation*, which can increase false negatives under threshold-based screening. Across multiple model families, the unbalanced training regime shows a tendency toward positive error shifts on true-anemic cases (overestimation), whereas rebalancing strategies concentrate errors closer to zero and reduce extreme right-tail behavior. A consistent qualitative pattern is that post-augmentation balancing ("Remark Balanced Aug." / "Severity Balanced Aug.") often compresses the central mass of errors on true-anemic cases, but this does not necessarily translate into improved global regression metrics; in the latest results table, augmented variants generally exhibited higher test RMSE than their non-augmented balanced counterparts. Practically, these plots provide a compact way to inspect the *failure mode that matters most for screening* (missed anemic cases) and to compare how strategy choice changes that failure mode across model classes.

Augmentation is not automatically beneficial in small, color-sensitive datasets.

Although patch augmentation is commonly used to increase robustness to pose and illumination variability, the most recent results show that "_aug" strategies increased test error across both remark- and severity-balanced regimes (e.g., remark-balanced test RMSE ≈2.09–2.31 vs. remark-balanced-aug test RMSE ≈2.52–3.13; severity-balanced test RMSE ≈2.05–2.31 vs. severity-balanced-aug test RMSE ≈2.72–2.98 for the reported models). This suggests that, for this dataset and feature pipeline, augmentation as implemented may introduce a distribution shift relative to the held-out test set, amplify nuisance variation that the engineered features do not normalize, or simply increase variance in a small-N setting. The correct takeaway is empirical: augmentation should be treated as a tunable component whose benefit must be demonstrated under the target evaluation protocol, rather than assumed.

Security and interoperability are intentionally scoped and offline-first.

The security and interoperability design is intentionally minimal but technically sound, mirroring the pragmatic architecture used in mature modular EHR ecosystems: a local clinical datastore as the source of truth, modular services for clinical functionality, and standards-based interfaces at integration boundaries. Concretely, the platform enforces a role-based access control (RBAC) boundary at the UI/API layer and encrypts sensitive records and linked artifacts at rest using authenticated encryption (AES-GCM). Interoperability is implemented as a best-effort, store-and-forward pathway: a local gateway queue accumulates outbound events while offline and, when connectivity is available, synchronizes a minimal set of external-facing resources (e.g., patient identifiers and screening outputs mapped to FHIR-style Observation/Media constructs) to a configured endpoint. This approach preserves offline operability and isolates integration concerns from core clinical workflows, while remaining consistent with a module-oriented EHR architecture where new screening services can be added without changing the core record system.

Limitations and next validation steps.

The dominant limitations concern translation rather than feasibility. The anemia evaluation is based on a small public dataset ($N = 250$) and is likely sensitive to domain shift (illumination, camera pipelines, skin-tone distributions, and protocol adherence), which is especially salient for color-derived features. The detector results are strong on the available validation imagery, but real-world acquisition variability may still affect ROI quality and downstream features. Finally, while the platform implements practical encryption, RBAC, and best-effort synchronization, this work does not claim full regulatory compliance or comprehensive security hardening; it demonstrates an implementable baseline appropriate for offline-first deployments. The next step to strengthen clinical credibility is prospective evaluation in the intended setting with standardized capture guidance, subgroup reporting (e.g., by hemoglobin bin and acquisition conditions), and explicit operating-point selection aligned to local referral capacity.

5. Conclusions

We presented an offline-first, layered edge telemedicine platform designed for clinics and outreach environments with intermittent connectivity and limited infrastructure. The system consolidates local EHR storage, application logic, and modular on-device screening microservices on a single embedded device, with a clinician-facing local web UI and optional best-effort synchronization to external systems via a standards-oriented gateway.

As a representative plug-in screening service, we implemented fingernail-image-based anemia screening using on-device nail-bed localization and hemoglobin estimation. On embedded hardware, post-training INT8 quantization reduced detector latency from 46.96 ms to 21.50 ms (54.2% reduction) while maintaining high detection performance (mAP@0.5 = 0.995), supporting interactive point-of-care operation; the deployed pipeline executes ROI extraction in ∼22 ms and hemoglobin regression in

∼34 ms on the Jetson Orin Nano. For hemoglobin estimation, a multi-strategy evaluation showed that training distribution policy materially affects performance: the lowest reported test RMSE in the most recent sweep was achieved by severity-balanced training with Support Vector Regression (2.048 g/dL; MAE 1.695 g/dL), while remark-balanced Gradient Boosting performed comparably (2.091 g/dL; MAE 1.794 g/dL). Augmented variants generally increased test error in this evaluation, highlighting that augmentation must be validated rather than assumed beneficial in small, color-sensitive datasets. Model selection is framed as choosing an operating point rather than a single universally best estimator. On the held-out test set (n=50; anemic n=14), a sensitivity-first configuration (remark-balanced GradientBoosting) achieved 1.00 sensitivity (14/14) but reduced specificity (0.53; 19/36), reflecting a triage-oriented policy that minimizes missed anemia at the expense of additional confirmatory testing. In contrast, an augmented unbalanced GradientBoosting configuration achieved a more balanced trade-off (sensitivity 0.79; specificity 0.81) at comparable RMSE, indicating that the preferred model depends on deployment constraints and the acceptable false-positive burden.

Overall, the results demonstrate the feasibility of secure, offline operation with integrated AI screening and a pragmatic interoperability scope aligned with modular EHR practice: RBAC at the UI/API boundary, AES-GCM encryption at rest for sensitive records and artifacts, and best-effort store-and-forward synchronization to an external endpoint when connectivity permits. Future work will focus on prospective field validation in the intended deployment context, robustness reporting across clinically relevant subgroups and acquisition conditions, and expansion to additional plug-in screening services within the same edge EHR contract.

6. Future Work and Translation Roadmap

1. **Prospective field validation:** Run a prospective study in the intended deployment setting to quantify (i) MAE/RMSE for Hb regression, (ii) sensitivity/specificity at pre-specified Hb thresholds, and (iii) operational metrics (time-to-result, capture failure rate, re-capture rate).
2. **Quality gating:** Add input-quality checks (blur/exposure/nail visibility) and a simple fail-safe policy (re-capture request or "no result").
3. **Deployment hardening:** Package pinned versions; add service health checks, structured logs, and crash recovery; verify consistent latency under concurrent UI usage.
4. **Interoperability increment:** Keep sync best-effort. Implement a minimal export contract (patient/encounter IDs + screening `Observation`-style outputs + optional linked images) via a queued gateway when connectivity is available.

Author Contributions: Conceptualization, S.A.C.R. and W.E.L.B.; methodology, S.A.C.R. and M.J.M.H.; software, S.A.C.R., M.J.M.H., S.Y.A.R. and J.A.S.F.; validation, S.A.C.R. and W.E.L.B.; formal analysis, S.A.C.R.; investigation, S.A.C.R. and M.J.M.H.; resources, W.E.L.B.; data processing, S.A.C.R. and M.J.M.H.; writing—original draft preparation, S.A.C.R.; writing—review and editing, all authors; visualization, S.A.C.R. and M.J.M.H., J.A.S.F; supervision, W.E.L.B.; project administration, S.A.C.R. and W.E.L.B.; funding acquisition, S.A.C.R and W.E.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSF-EPSCoR Center for the Advancement of Wearable Technologies, National Science Foundation grant OIA-184924. Furthermore, this work used Jetstream2 at Indiana University through allocation CIS250603 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296

Institutional Review Board Statement: The study used a publicly available, de-identified dataset and did not involve new data collection from human subjects; therefore, institutional review board approval was not required for this analysis. Planned future data collection in Puerto Rican communities will be conducted under an IRB-approved protocol at the University of Puerto Rico at Mayagüez.

Informed Consent Statement: Not applicable. The dataset used consists of anonymized records collected and shared under prior ethical approval by its original authors.

Data Availability Statement: The fingernail image dataset analyzed in this study is available from Yakimov et al. [17]. Code for feature extraction, KDE reweighting, model training, and edge deployment will be made available in a public repository upon publication.

Acknowledgments: The authors thank the Center for Research and Development at the University of Puerto Rico at Mayagüez for administrative support, and Agus E. Marrero Marrero, BSN, for guidance on community health workflows and planning for future field evaluations.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LMICs	Low- and Middle-Income Countries
WHO	World Health Organization
Hb	Hemoglobin
API	Application Programming Interface
REST	Representational State Transfer
UI	User Interface
RBAC	Role-Based Access Control
AES	Advanced Encryption Standard
AES-GCM	Advanced Encryption Standard in Galois/Counter Mode
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven International
KDE	Kernel Density Estimation
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
mAP	mean Average Precision
IoU	Intersection over Union
RANSAC	RANdom SAmple Consensus
YOLO	You Only Look Once
PTQ	Post-Training Quantization

Appendix A. Supplementary Model Results

Additional predicted-vs-true plots, Bland–Altman agreement, and confusion matrices for alternative regressors/classifiers. Moreover, model performance across different strategies is shown in Table A1.

idx	Strategy	Model	Val MAE	Val RMSE	Test MAE	Test RMSE
4	unbalanced	GradientBoosting	1.744805	2.376128	1.657123	2.064130
3	unbalanced	RandomForest	1.752682	2.364310	1.492580	1.881038
2	unbalanced	Lasso	1.755545	2.354426	1.619063	2.055294
1	unbalanced	Ridge	1.741066	2.309620	1.571398	2.000839
0	unbalanced	ElasticNet	1.729408	2.311786	1.624841	2.075797
5	unbalanced	SupportVectorRegression	1.695306	2.324324	1.518676	1.931470
6	unbalanced	HuberRegressor	1.768820	2.355575	1.707790	2.160689
7	unbalanced	RANSACRegressor	1.836750	2.432478	1.872492	2.328339
28	unbalanced aug	GradientBoosting	1.833560	2.444962	1.686235	2.092567
27	unbalanced aug	RandomForest	1.783460	2.399692	1.707425	2.071885
26	unbalanced aug	Lasso	1.909720	2.526389	1.818504	2.259129
25	unbalanced aug	Ridge	1.889845	2.498913	1.826541	2.273066
24	unbalanced aug	ElasticNet	1.889873	2.499263	1.830098	2.273897
29	unbalanced aug	SupportVectorRegression	1.812684	2.410276	1.602212	2.062206
30	unbalanced aug	HuberRegressor	1.903560	2.553992	1.843167	2.296513
31	unbalanced aug	RANSACRegressor	1.897090	2.544045	1.886539	2.344043
12	remark balanced	GradientBoosting	2.304291	2.875903	1.793817	2.091317
11	remark balanced	RandomForest	2.202529	2.786093	1.789606	2.178235
10	remark balanced	Lasso	2.403675	3.188720	1.773397	2.192696
9	remark balanced	Ridge	2.301757	3.039550	1.784307	2.192869
8	remark balanced	ElasticNet	2.293114	3.032196	1.802253	2.207815
13	remark balanced	SupportVectorRegression	1.906287	2.546419	1.836397	2.312082
14	remark balanced	HuberRegressor	2.680457	3.497003	2.343244	2.894390
15	remark balanced	RANSACRegressor	2.819317	3.647884	2.193600	3.046019
36	remark balanced aug	GradientBoosting	2.300030	2.928760	2.114073	2.524201
35	remark balanced aug	RandomForest	2.417166	3.013901	2.302483	2.694730
34	remark balanced aug	Lasso	2.511109	3.122034	2.487610	2.914302
33	remark balanced aug	Ridge	2.449691	3.076051	2.434495	2.847980
32	remark balanced aug	ElasticNet	2.465140	3.080420	2.514393	2.969631
37	remark balanced aug	SupportVectorRegression	2.198505	2.780235	2.368737	2.767604
38	remark balanced aug	HuberRegressor	2.564206	3.226937	2.314693	2.834734
39	remark balanced aug	RANSACRegressor	2.667010	3.359722	2.450812	3.131512
16	severity balanced	ElasticNet	2.071649	2.699673	1.771473	2.209205
17	severity balanced	Ridge	2.071691	2.695815	1.767760	2.224609
18	severity balanced	Lasso	2.110076	2.738590	1.776973	2.241110
19	severity balanced	RandomForest	1.958795	2.532324	1.776798	2.174799
20	severity balanced	GradientBoosting	2.039298	2.603057	1.877912	2.308078
21	severity balanced	SupportVectorRegression	1.828047	2.372644	1.694641	2.048299
22	severity balanced	HuberRegressor	2.204231	2.912189	1.827314	2.346841
23	severity balanced	RANSACRegressor	2.380310	3.442254	2.463980	4.201977
40	severity balanced aug	ElasticNet	2.277669	2.904024	2.531066	2.984675
41	severity balanced aug	Ridge	2.277767	2.904399	2.445523	2.845367
42	severity balanced aug	Lasso	2.296876	2.946031	2.530658	2.982150
43	severity balanced aug	RandomForest	2.207324	2.778476	2.382975	2.785850
44	severity balanced aug	GradientBoosting	2.278527	2.859685	2.403471	2.718525
45	severity balanced aug	SupportVectorRegression	2.110615	2.656923	2.359506	2.802018
46	severity balanced aug	HuberRegressor	2.337203	2.967802	2.293366	2.752779
47	severity balanced aug	RANSACRegressor	2.405756	3.061265	2.548659	3.298975

Table A1. Validation (Val) and test (Test) mean absolute error (MAE) and root mean squared error (RMSE) across all strategies (including unbalanced and unbalanced aug) and model families. Lower values indicate better predictive accuracy. In these results, the unbalanced RandomForest achieves the best held-out performance with $\textit{Test MAE} = 1.493$ and $\textit{Test RMSE} = 1.881$, while augmentation variants generally show higher test error than their non-augmented counterparts.

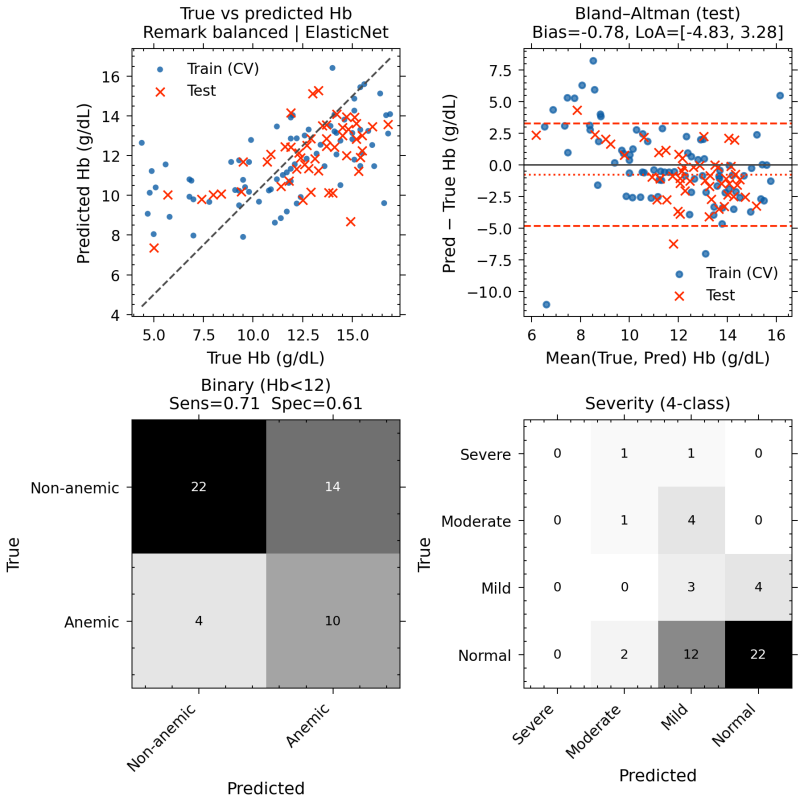


Figure A1. Supplementary evaluation plots for ELASTICNET: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

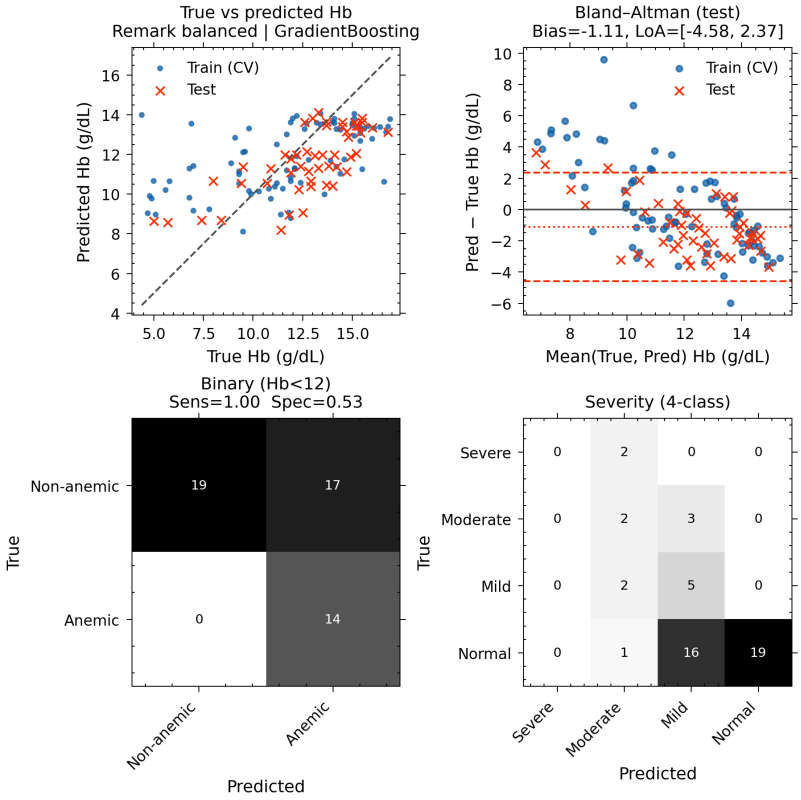


Figure A2. Supplementary evaluation plots for GRADIENT BOOSTING: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

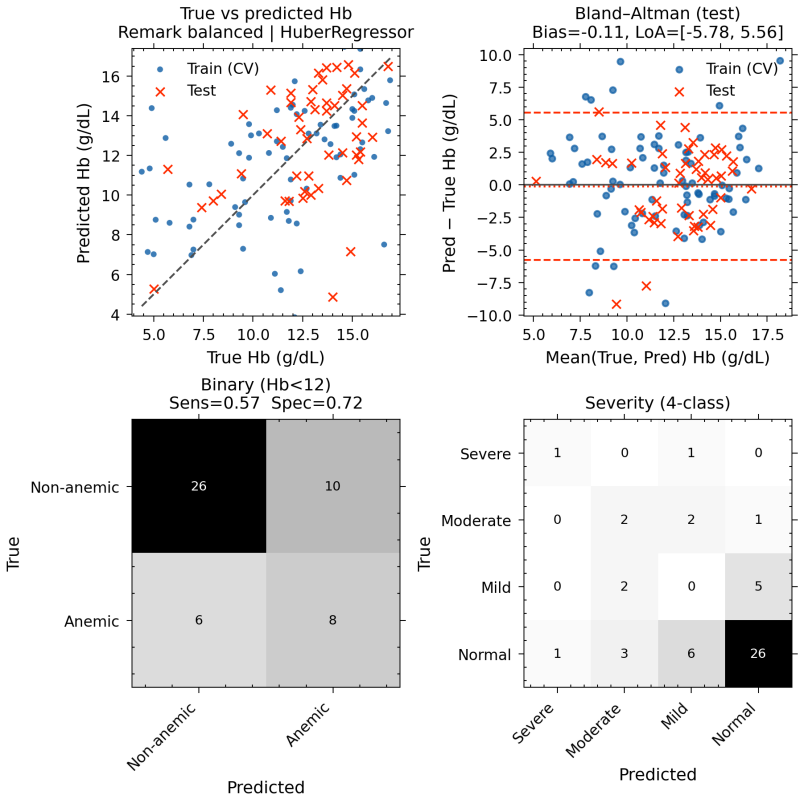


Figure A3. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

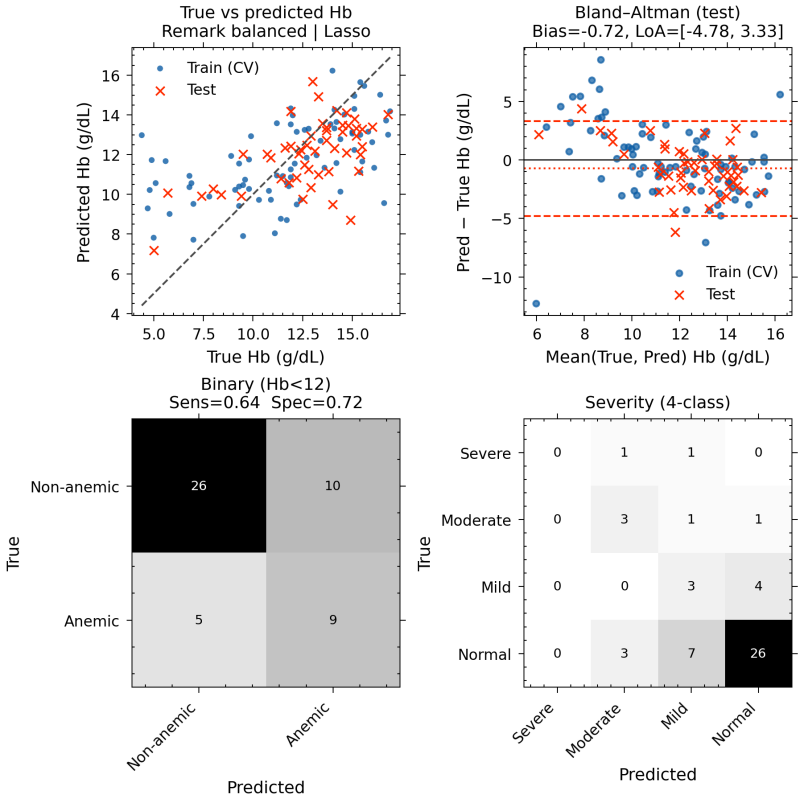


Figure A4. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

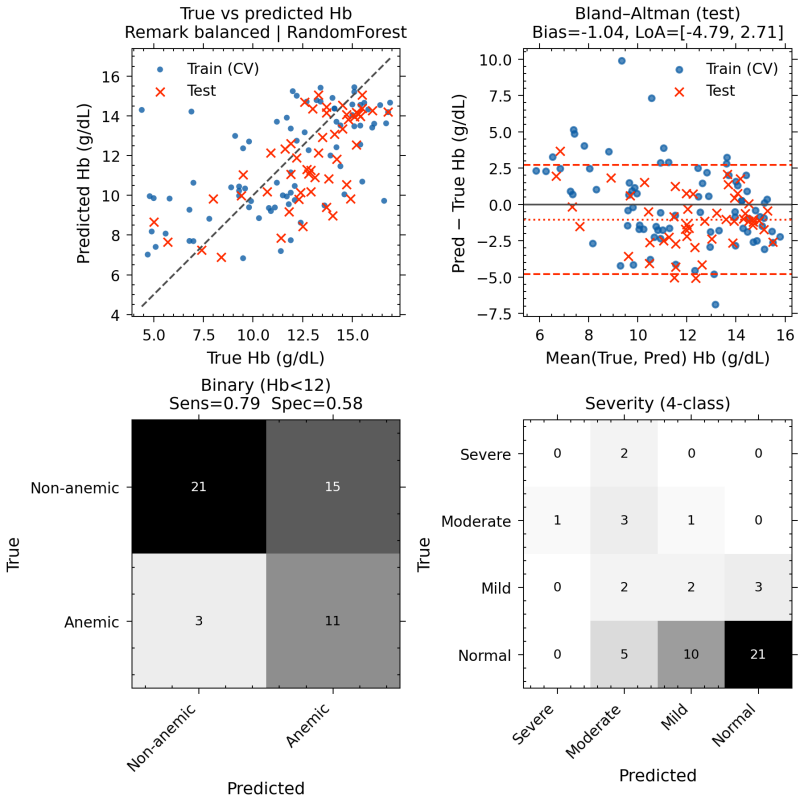


Figure A5. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

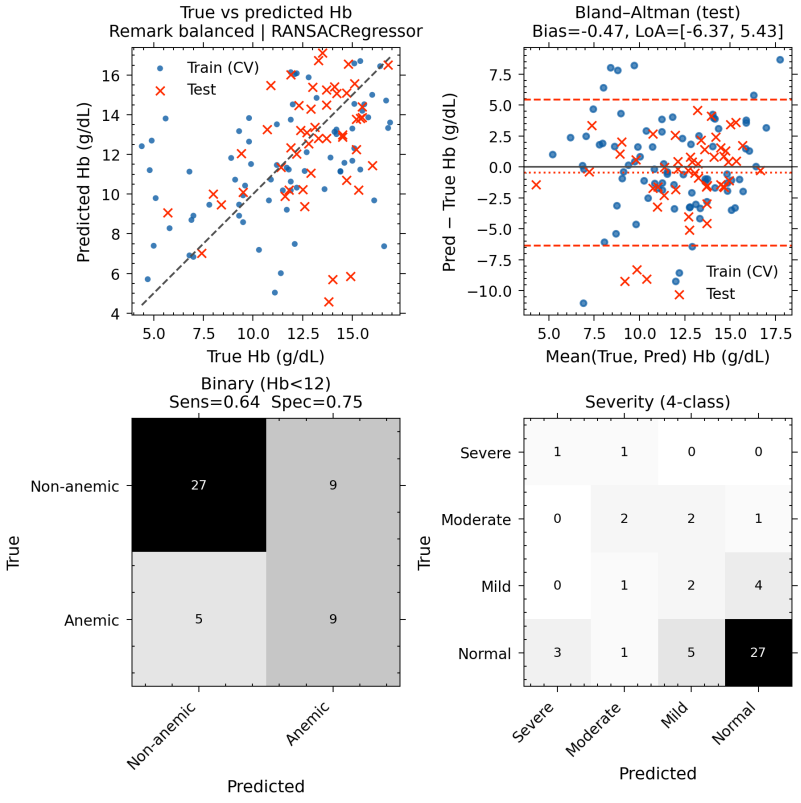


Figure A6. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

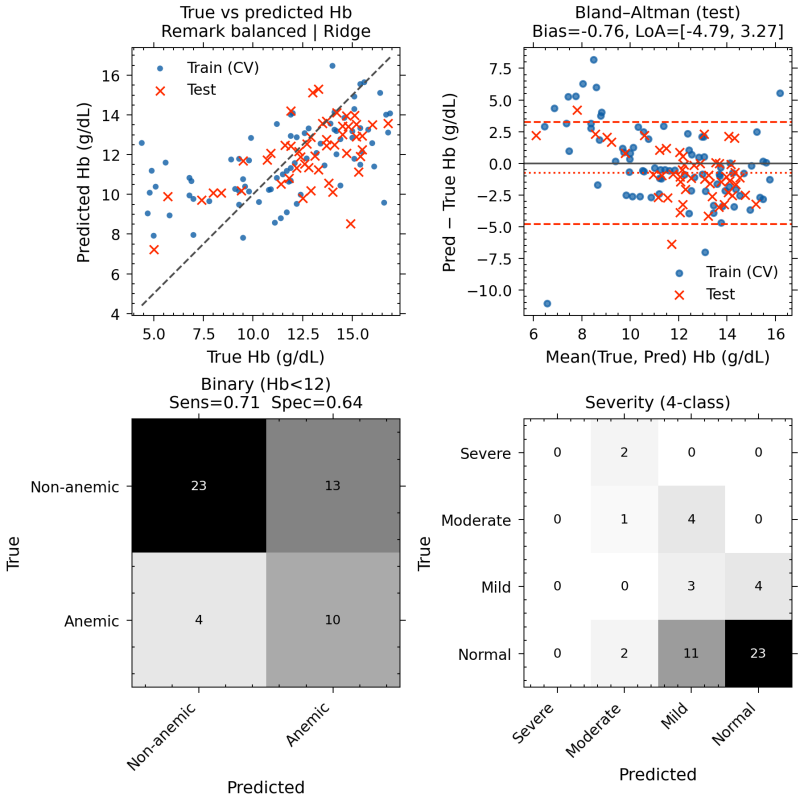


Figure A7. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

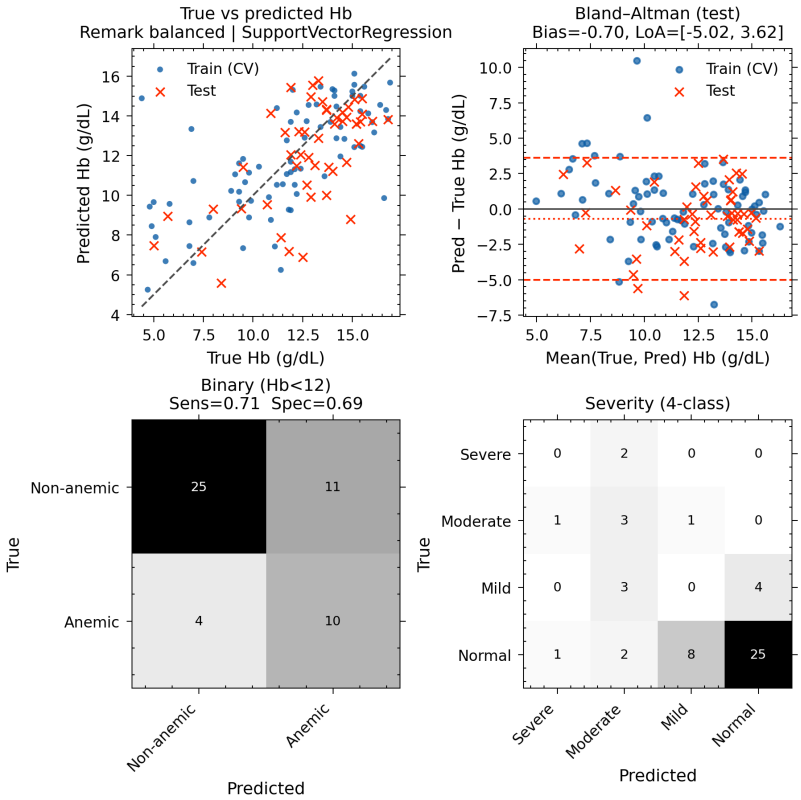


Figure A8. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced` strategy.

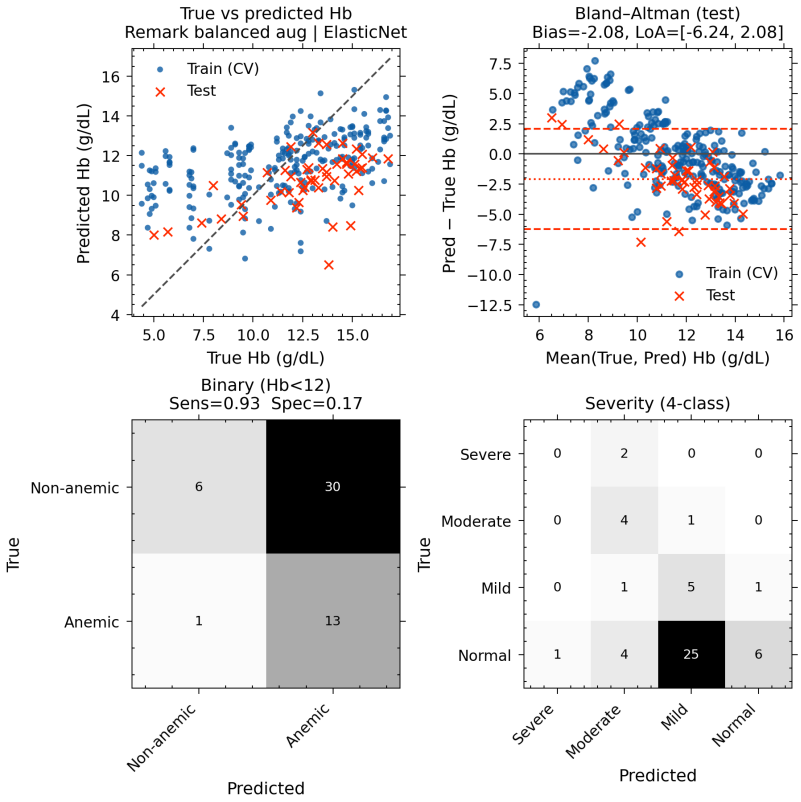


Figure A9. Supplementary evaluation plots for ElasticNet: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

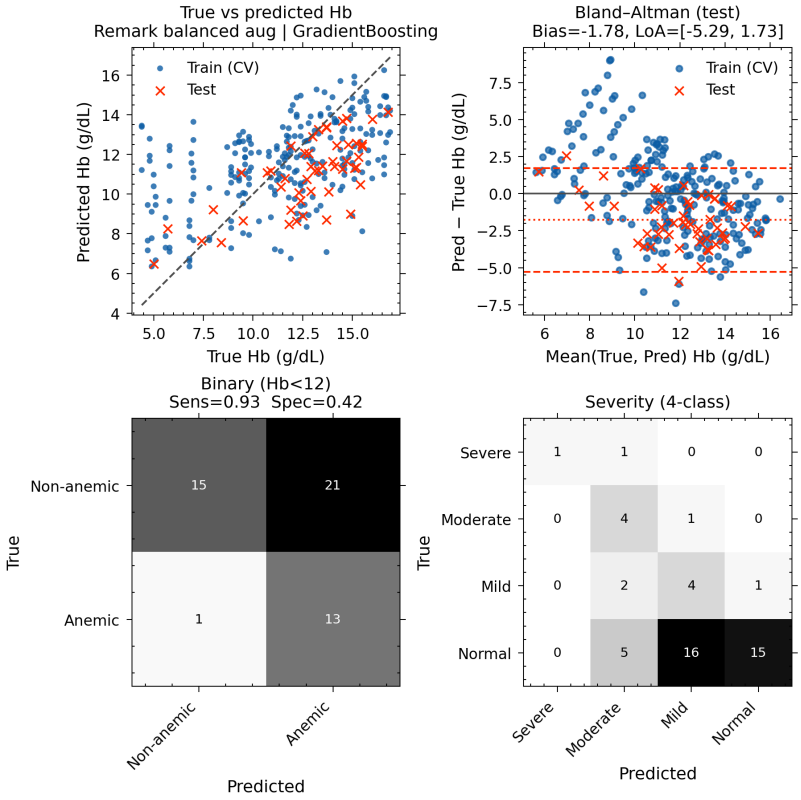


Figure A10. Supplementary evaluation plots for Gradient Boosting: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

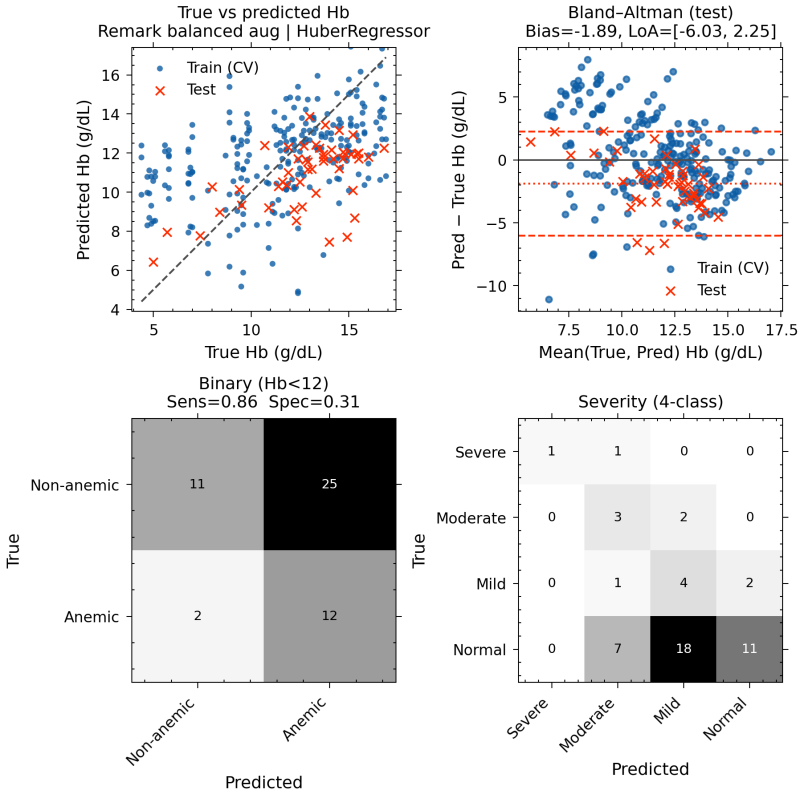


Figure A11. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

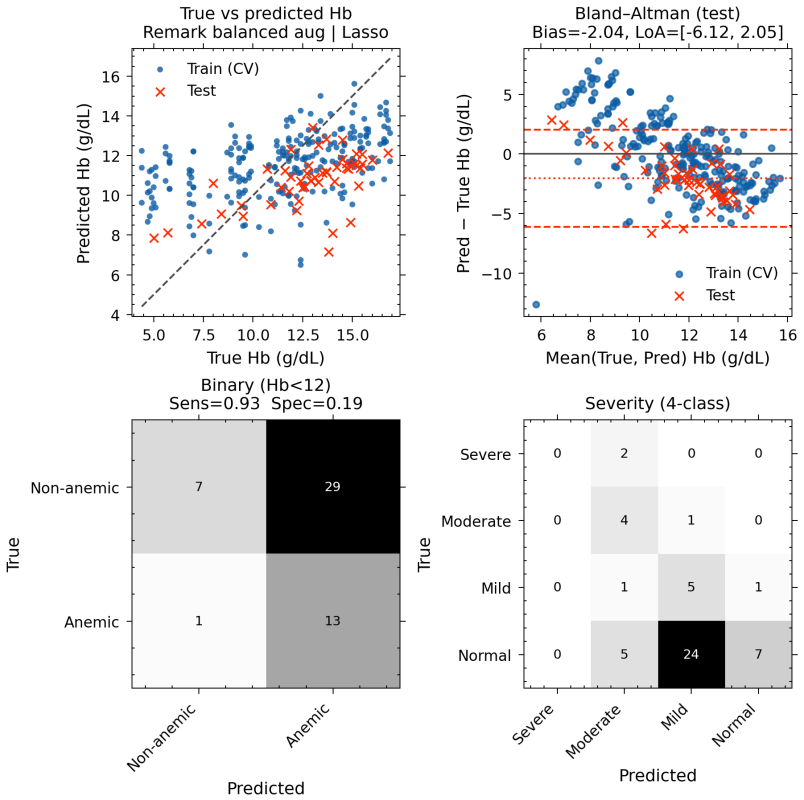


Figure A12. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

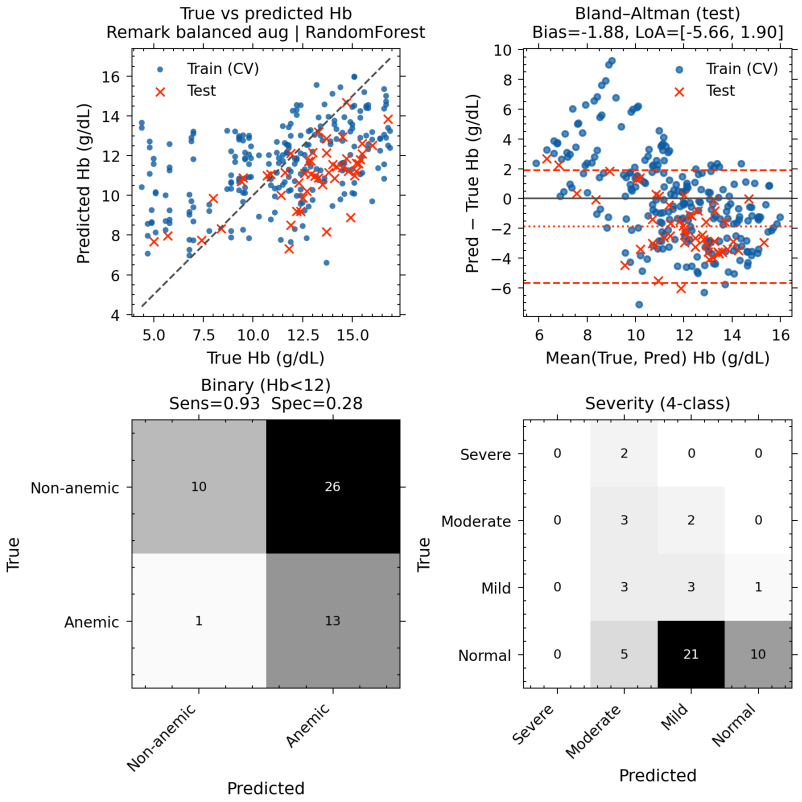


Figure A13. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

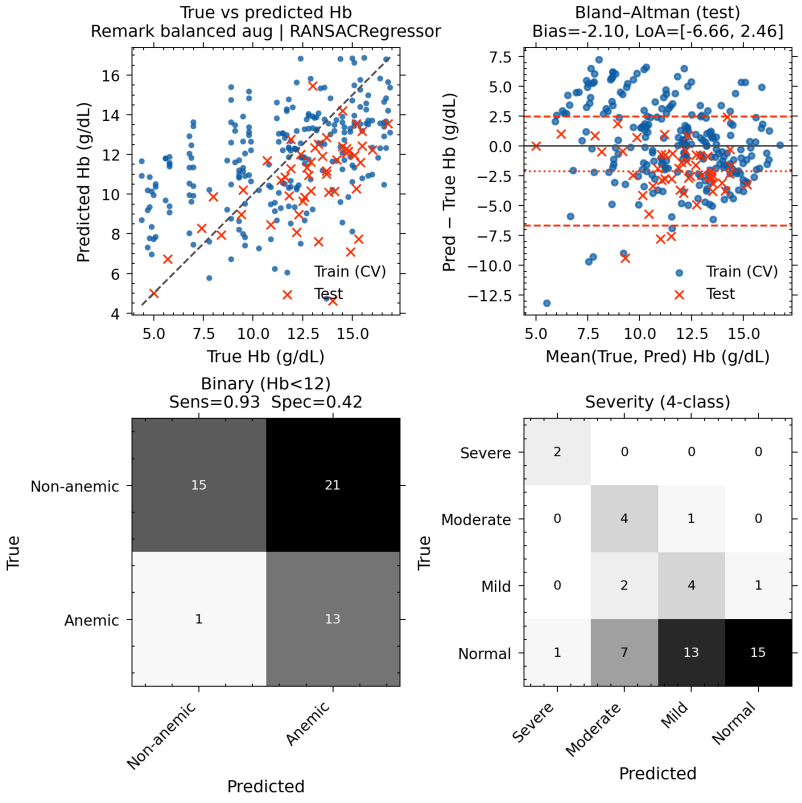


Figure A14. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

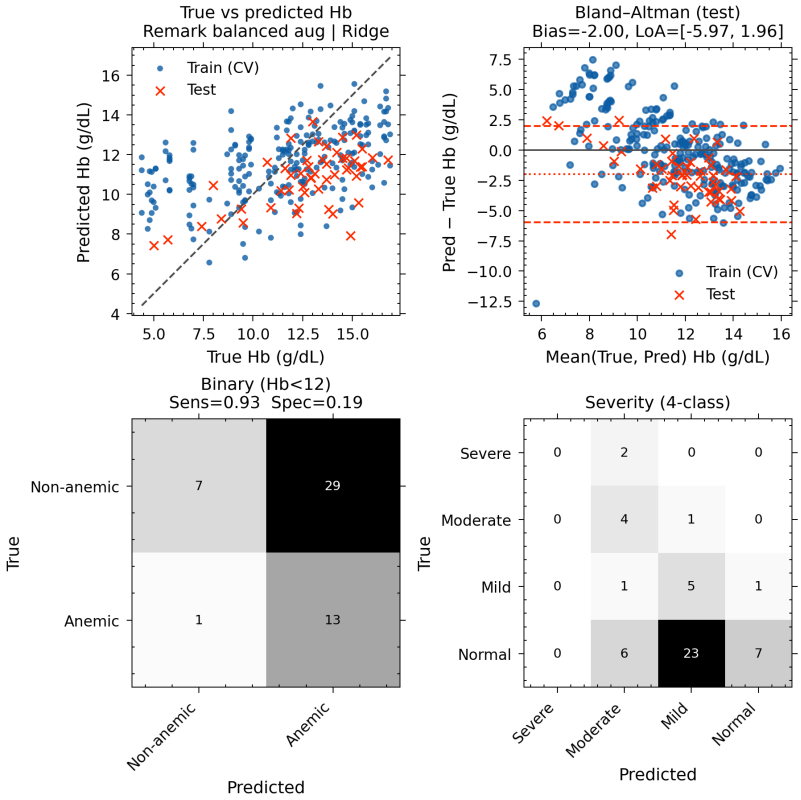


Figure A15. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

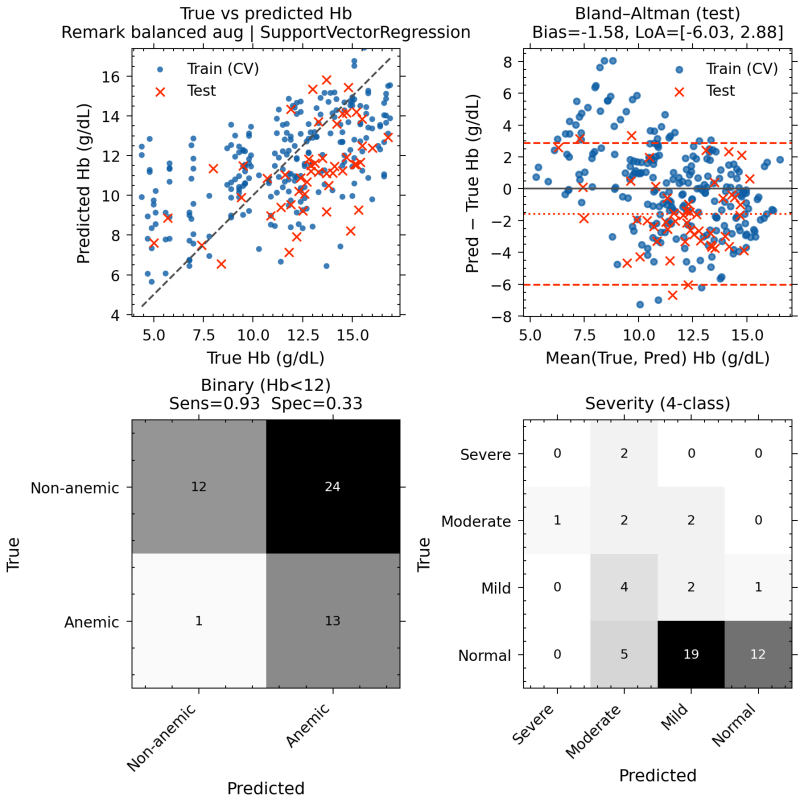


Figure A16. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `remark_balanced_aug` strategy.

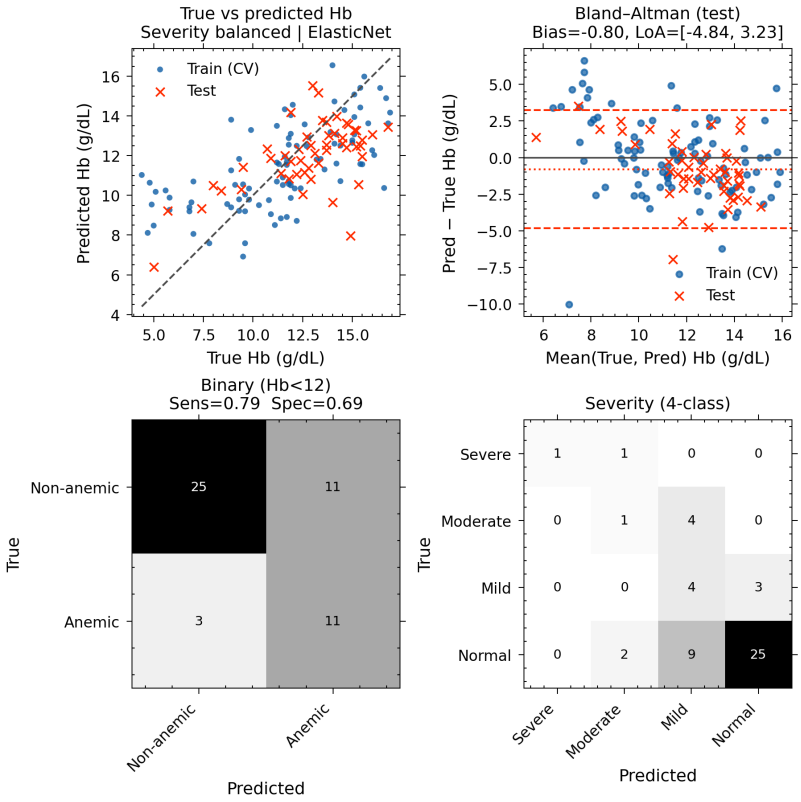


Figure A17. Supplementary evaluation plots for ElasticNet: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

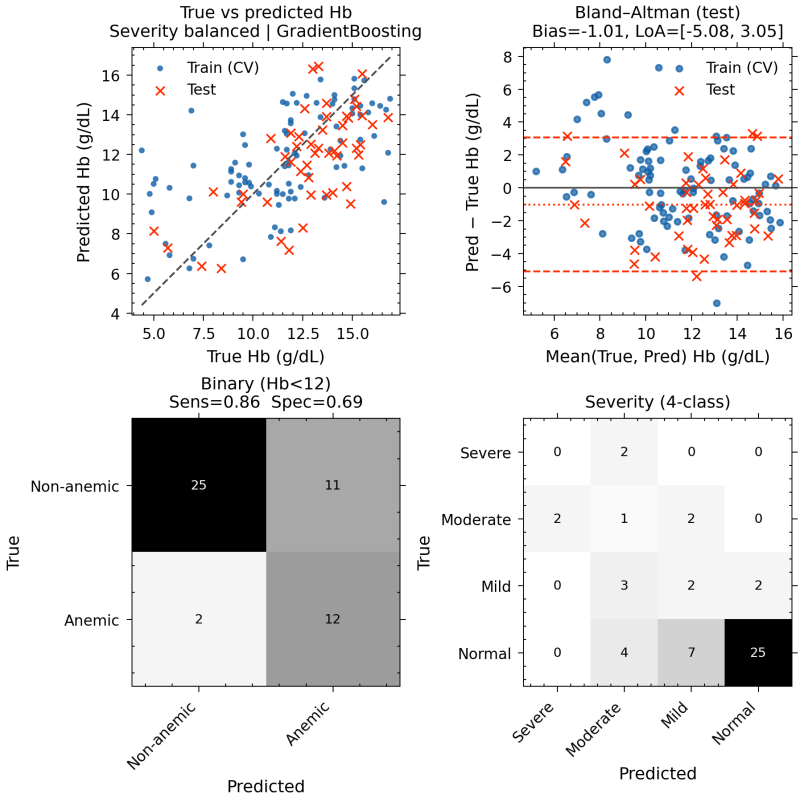


Figure A18. Supplementary evaluation plots for Gradient Boosting: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

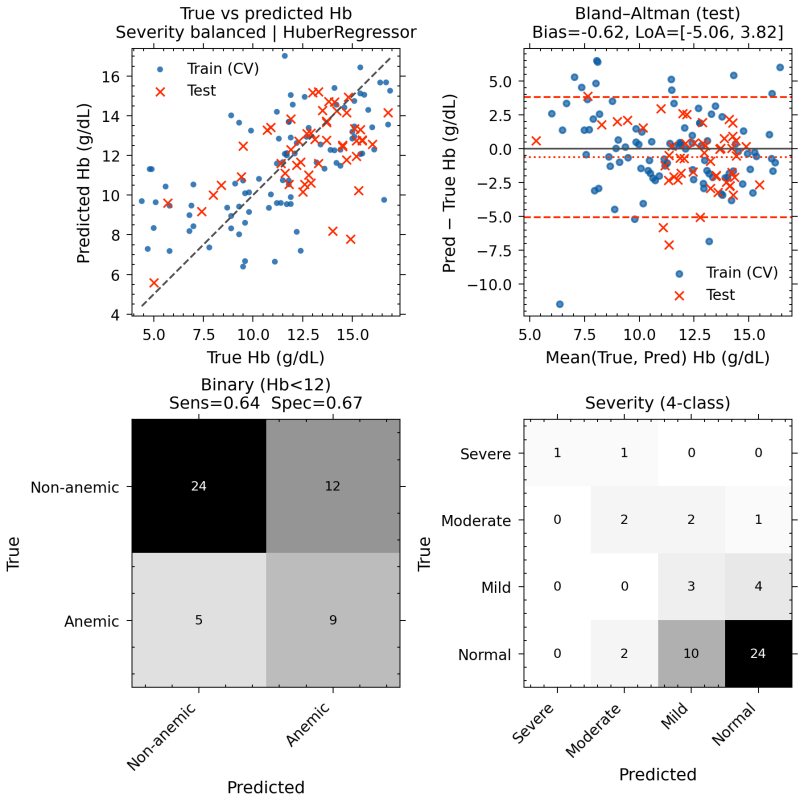


Figure A19. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

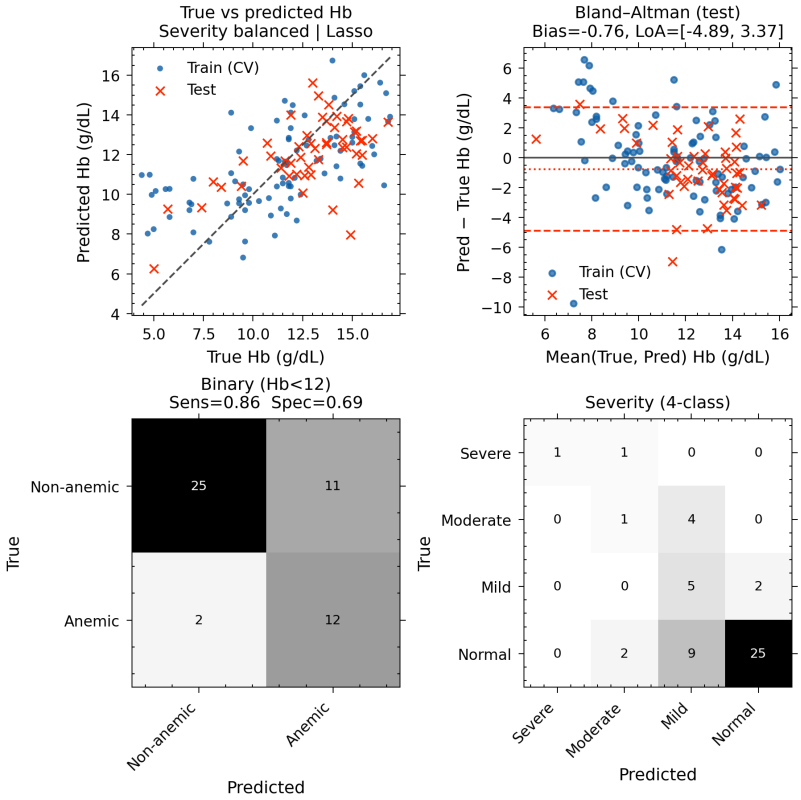


Figure A20. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

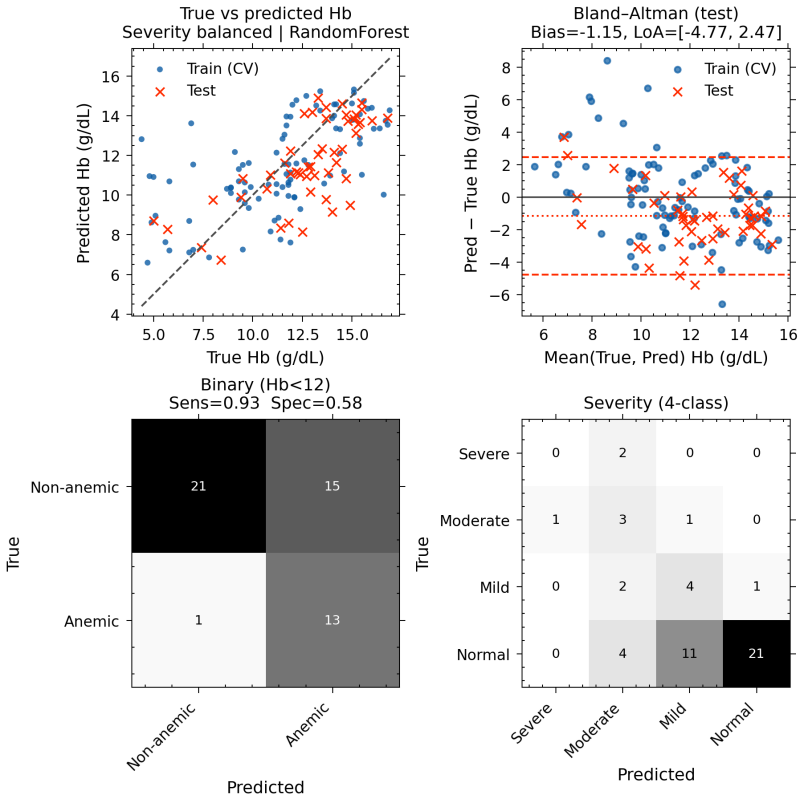


Figure A21. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

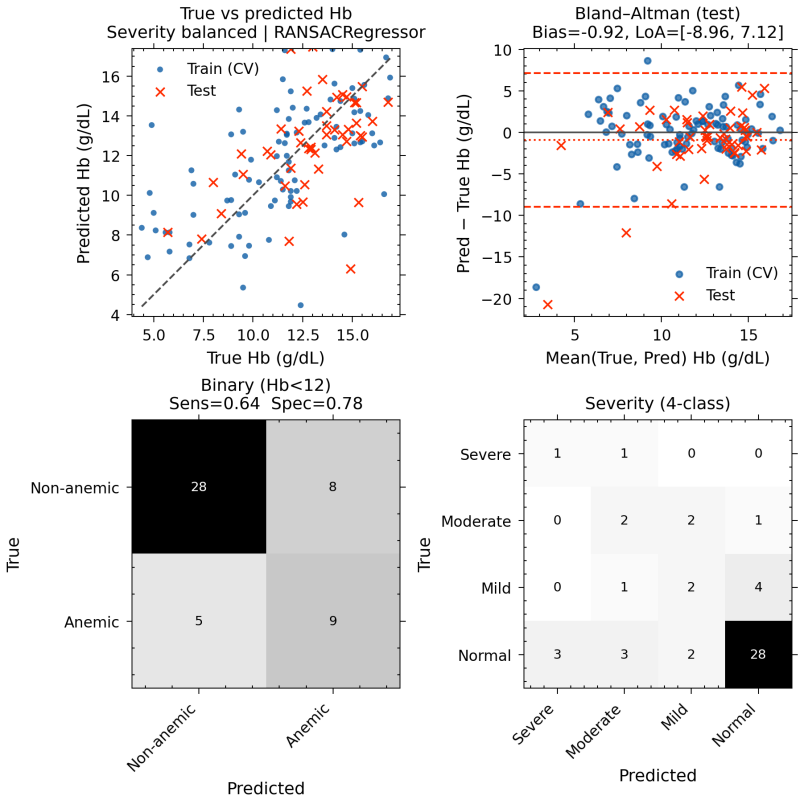


Figure A22. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

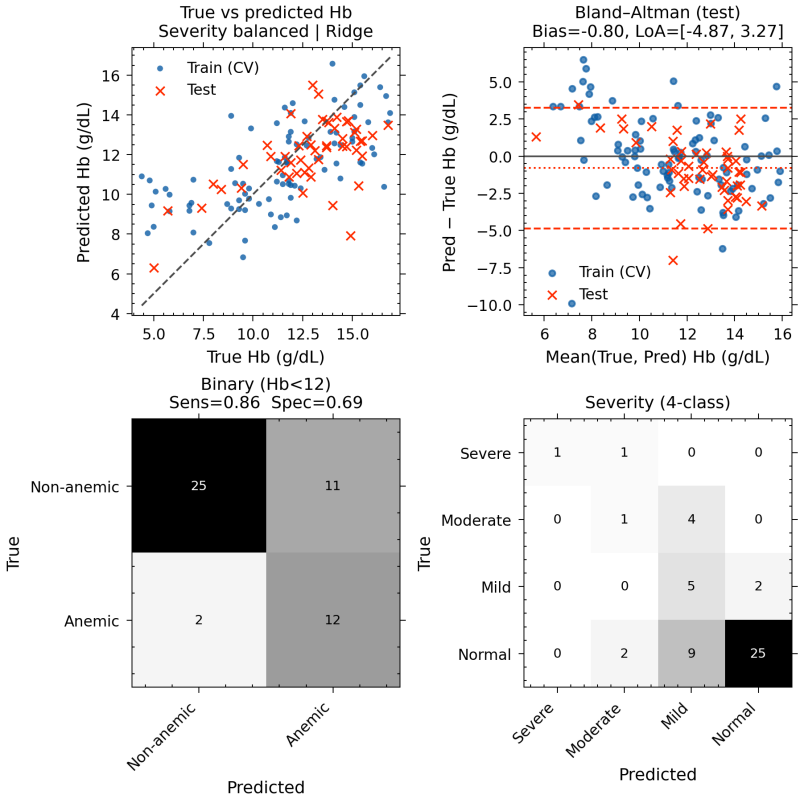


Figure A23. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

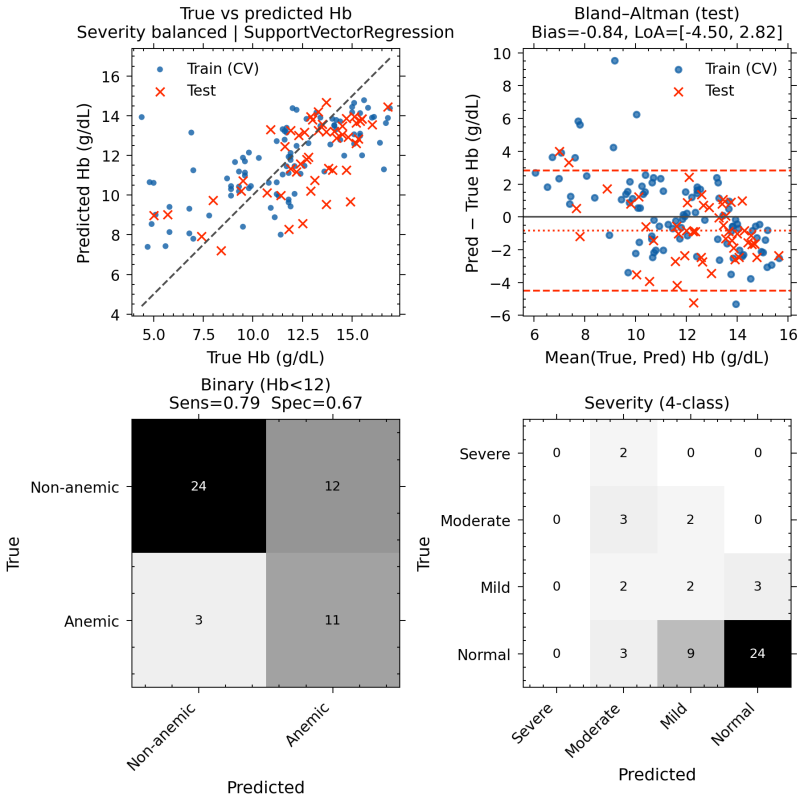


Figure A24. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced` strategy.

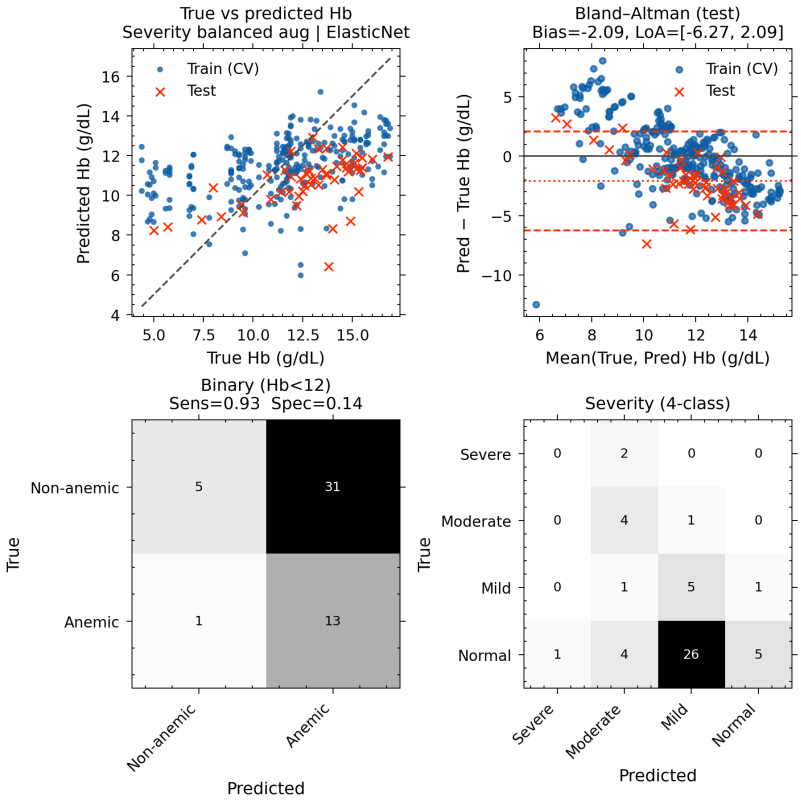


Figure A25. Supplementary evaluation plots for ELASTICNET: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

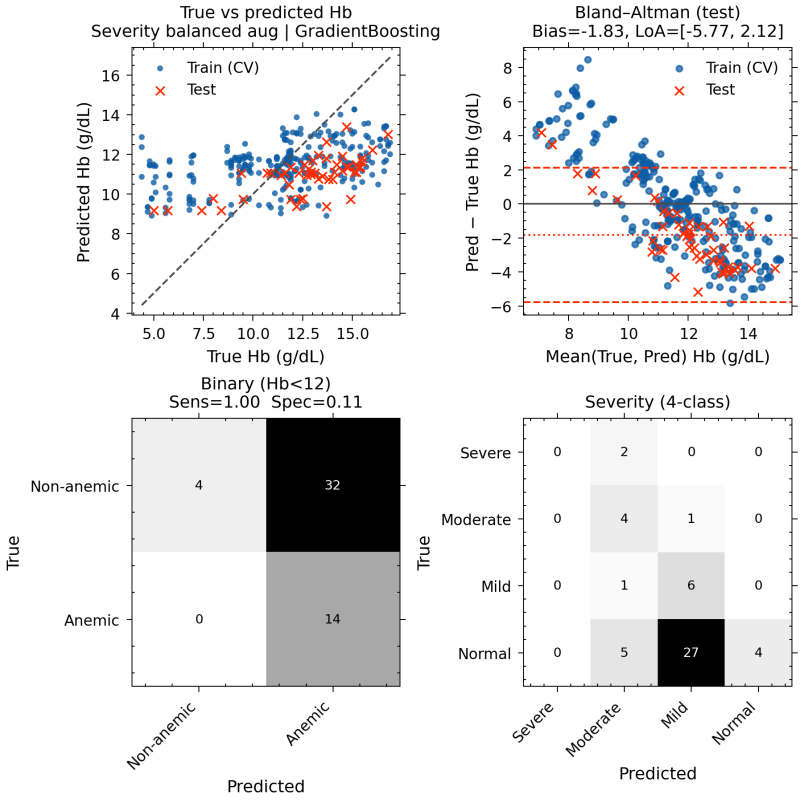


Figure A26. Supplementary evaluation plots for GRADIENT BOOSTING: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

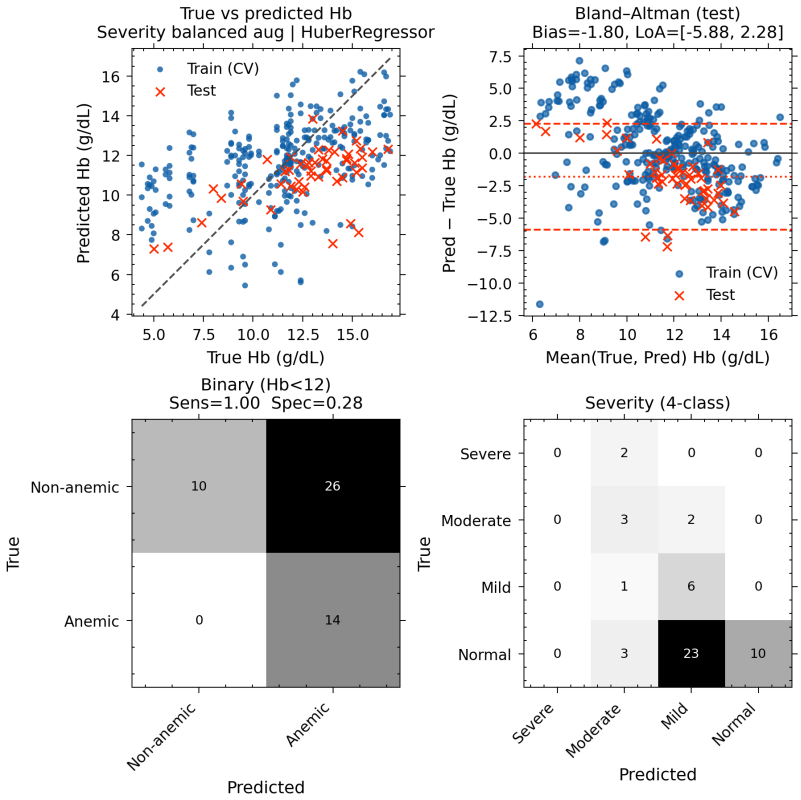


Figure A27. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

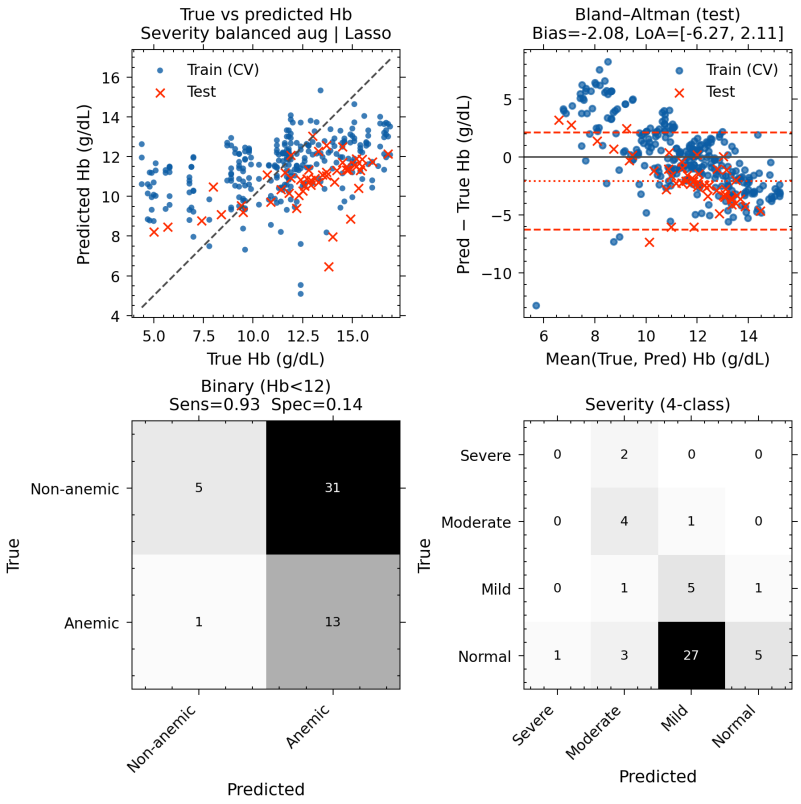


Figure A28. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

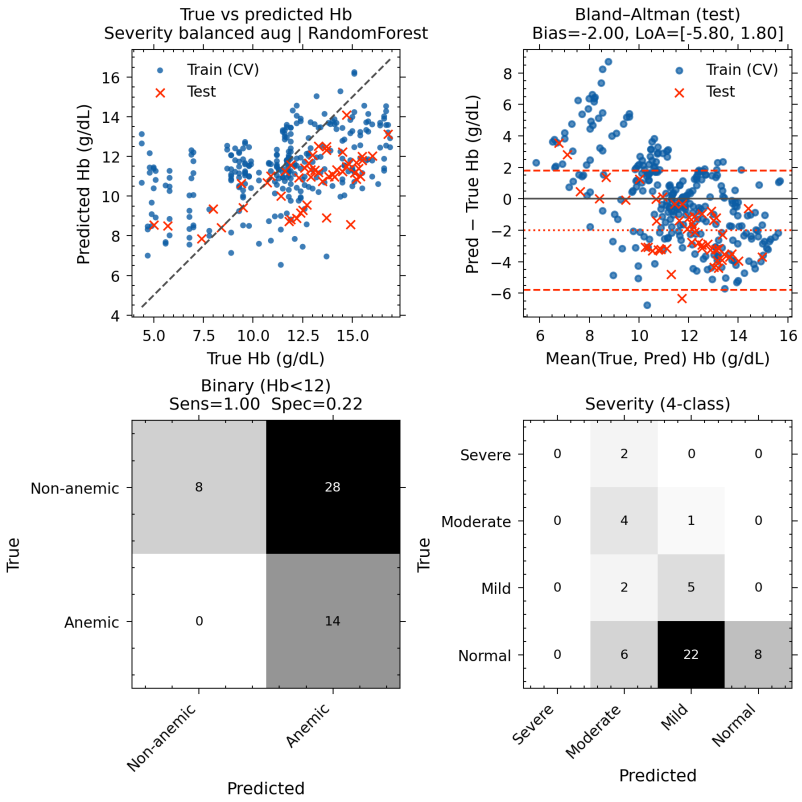


Figure A29. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

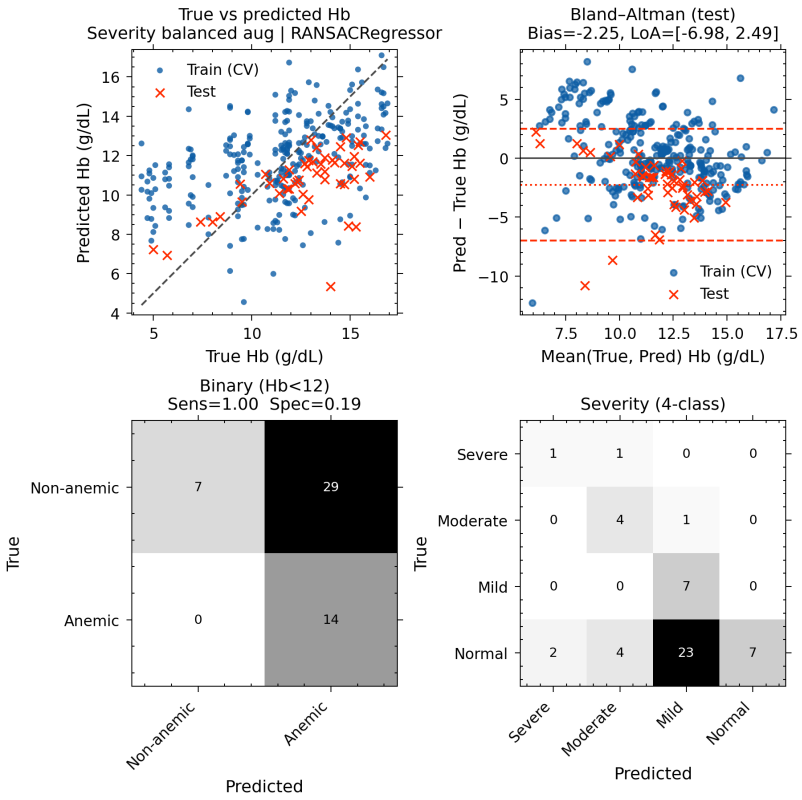


Figure A30. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

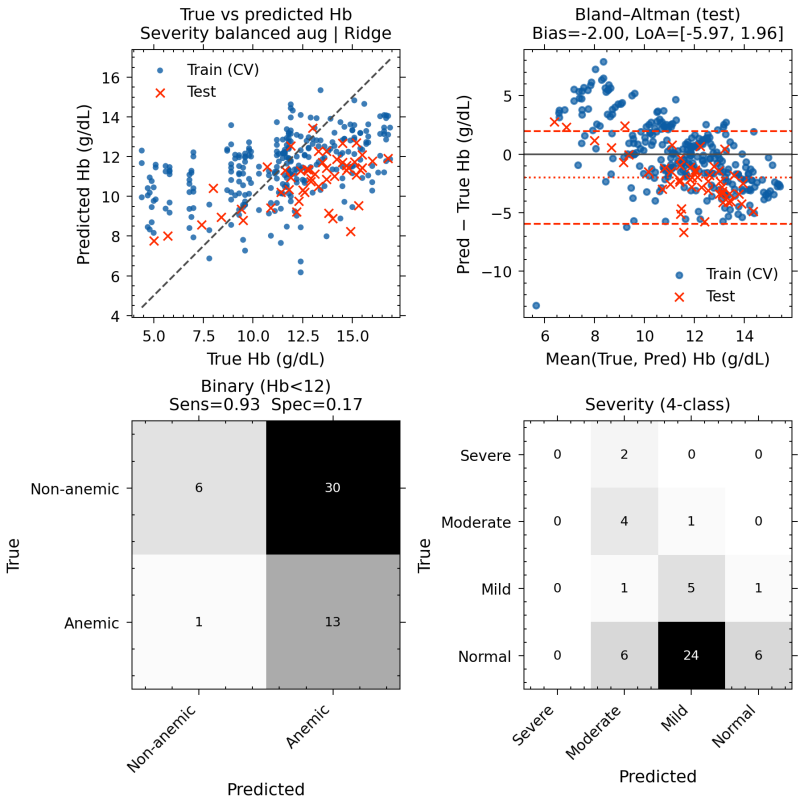


Figure A31. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

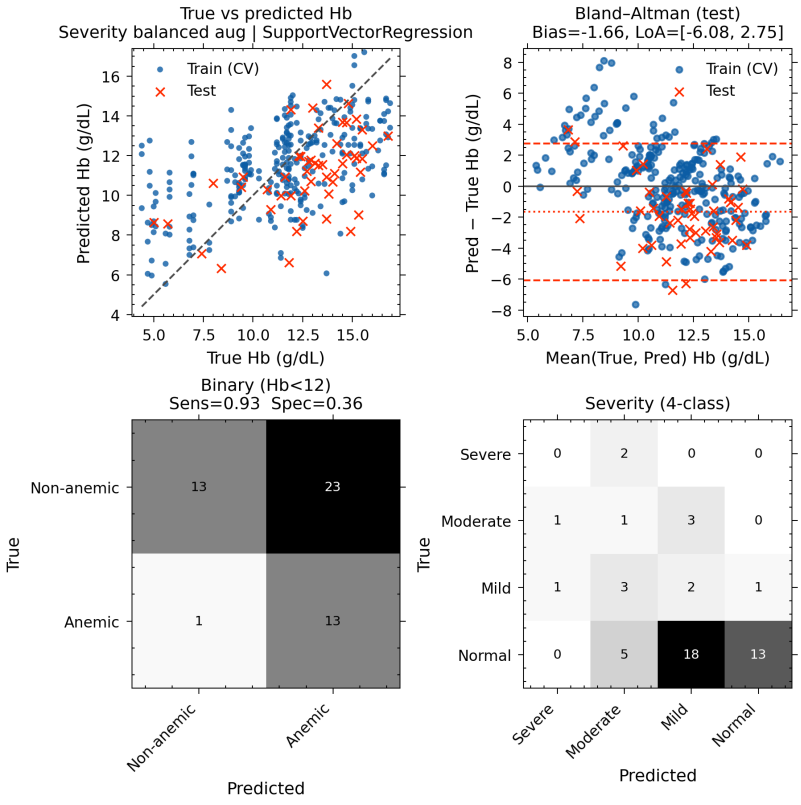


Figure A32. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `severity_balanced_aug` strategy.

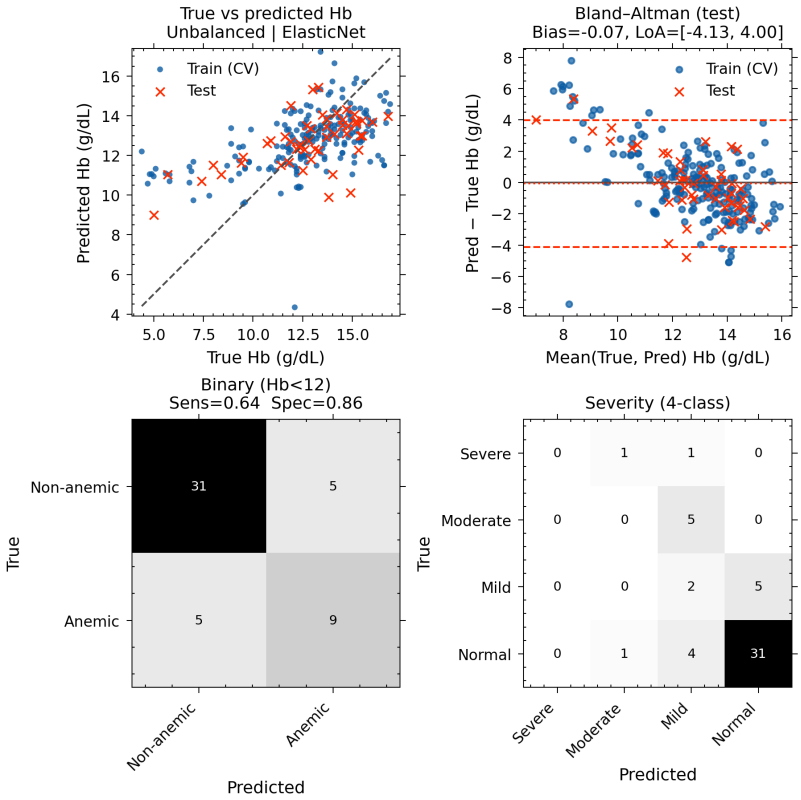


Figure A33. Supplementary evaluation plots for ElasticNet: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

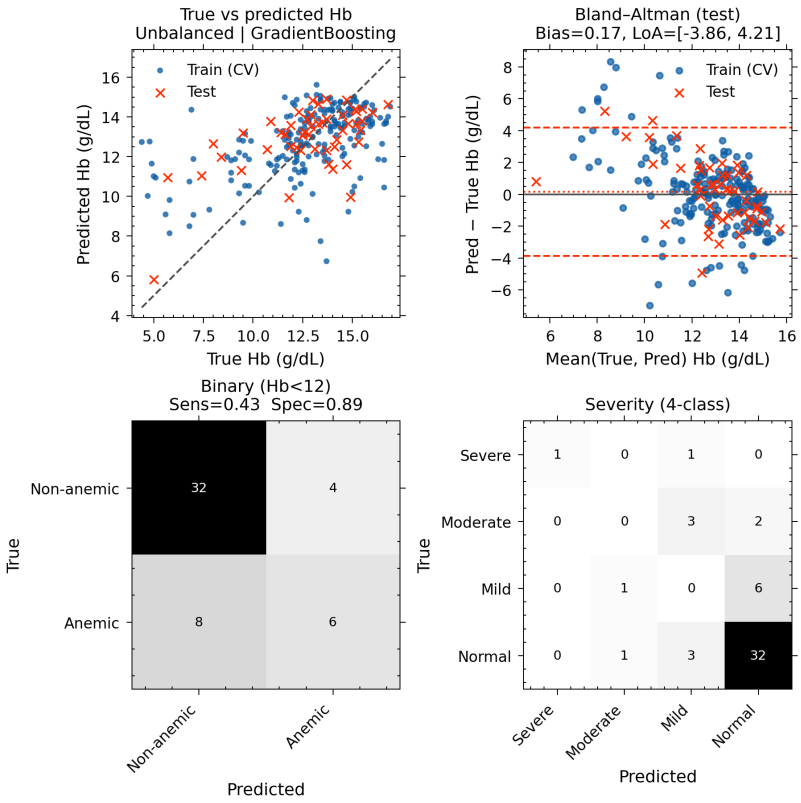


Figure A34. Supplementary evaluation plots for Gradient Boosting: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

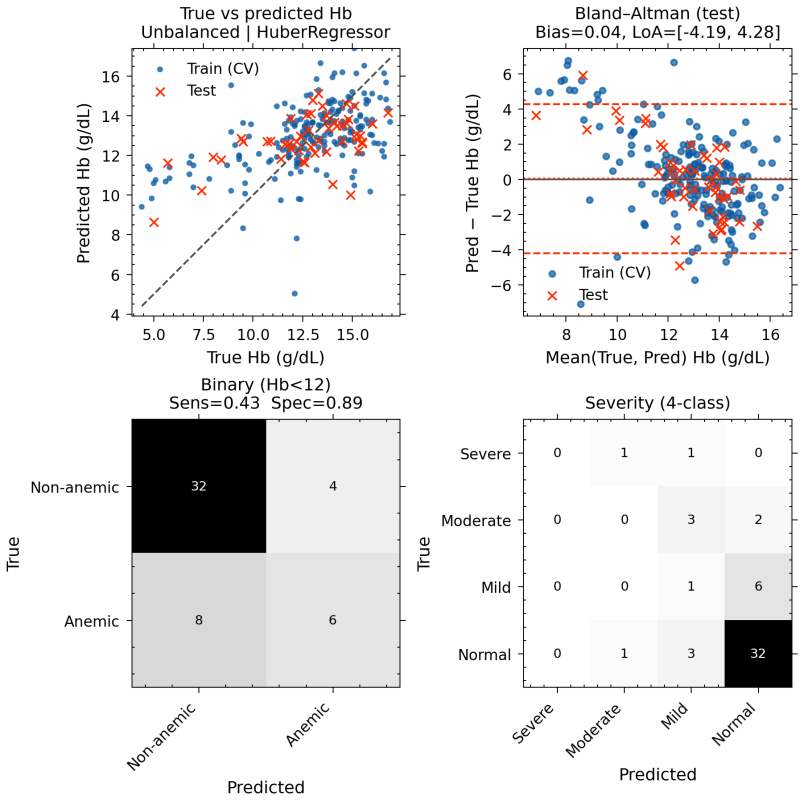


Figure A35. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

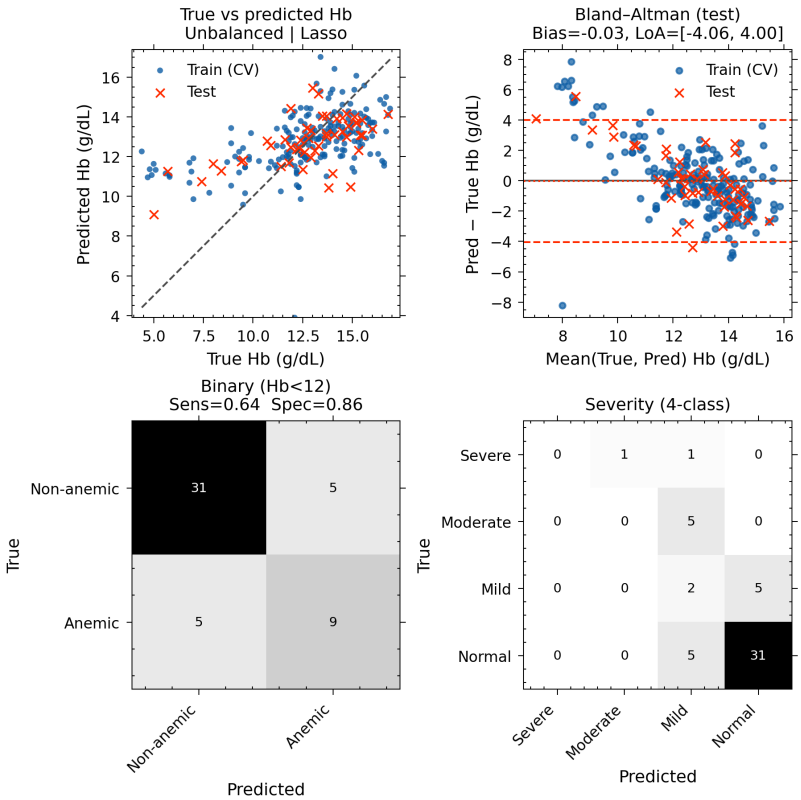


Figure A36. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

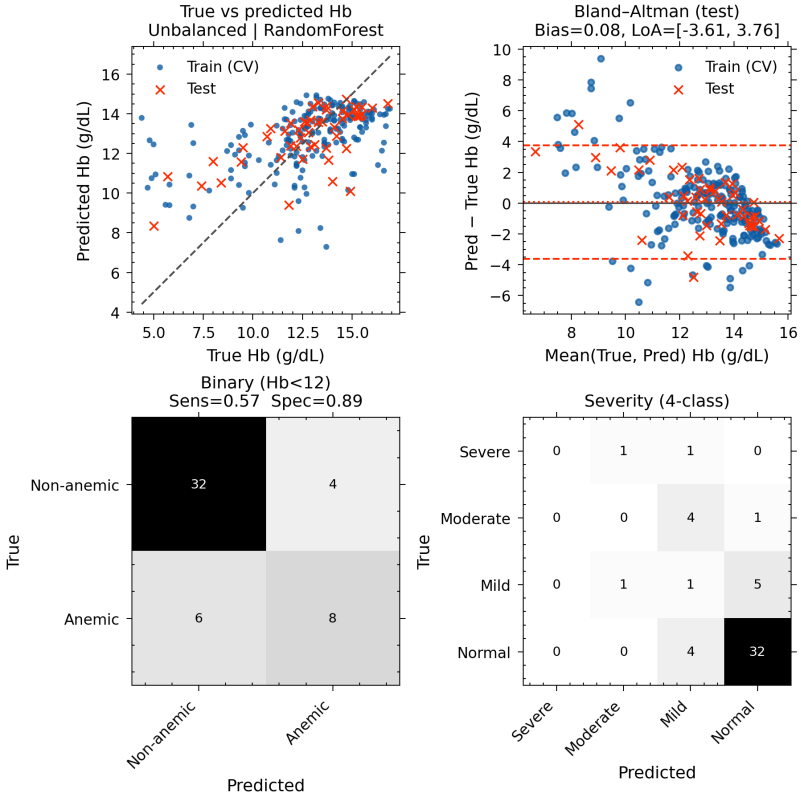


Figure A37. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

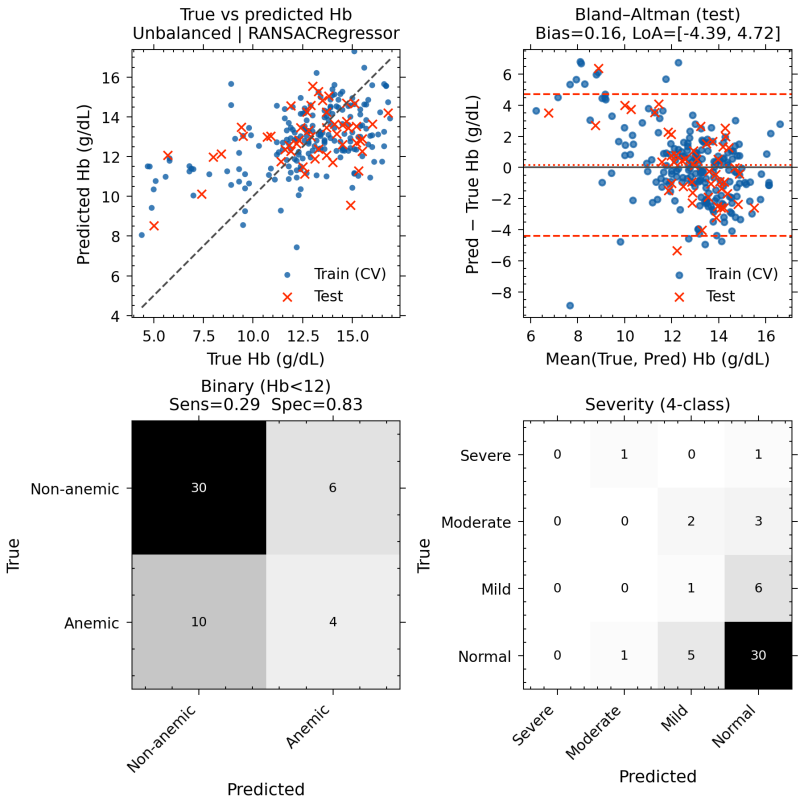


Figure A38. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

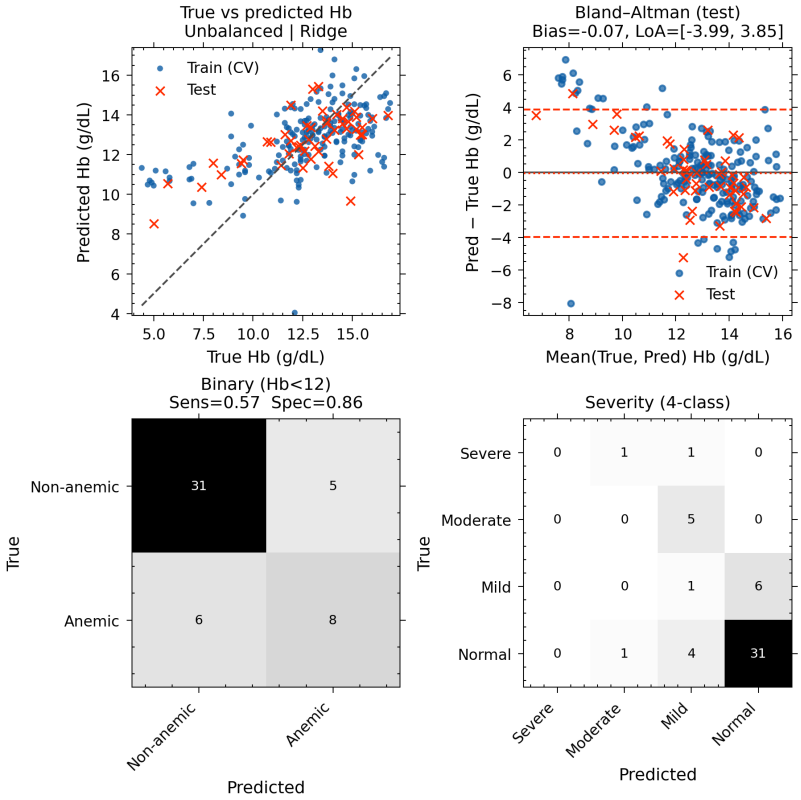


Figure A39. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

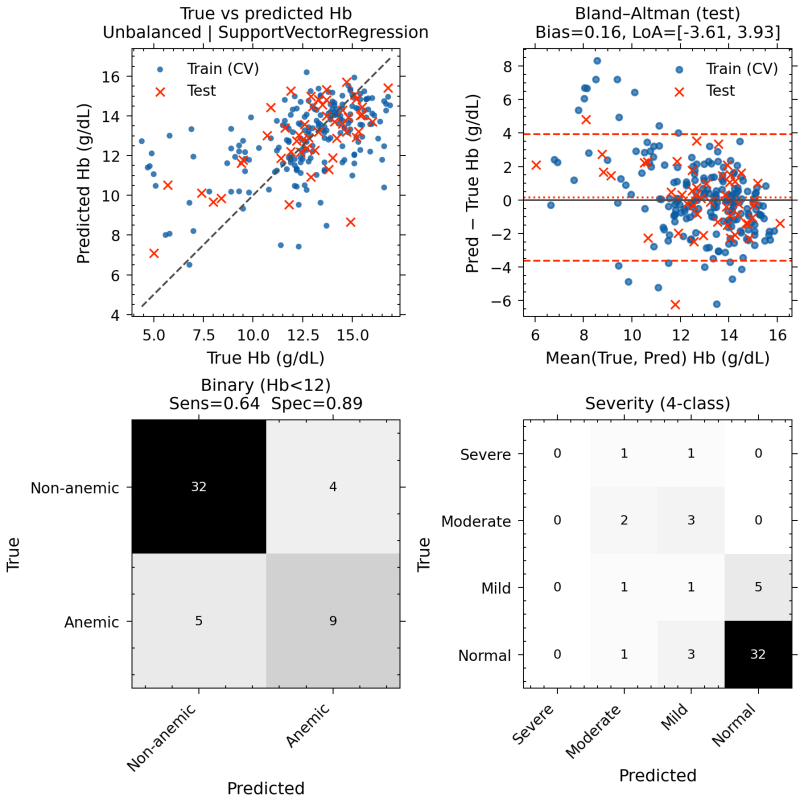


Figure A40. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced` strategy.

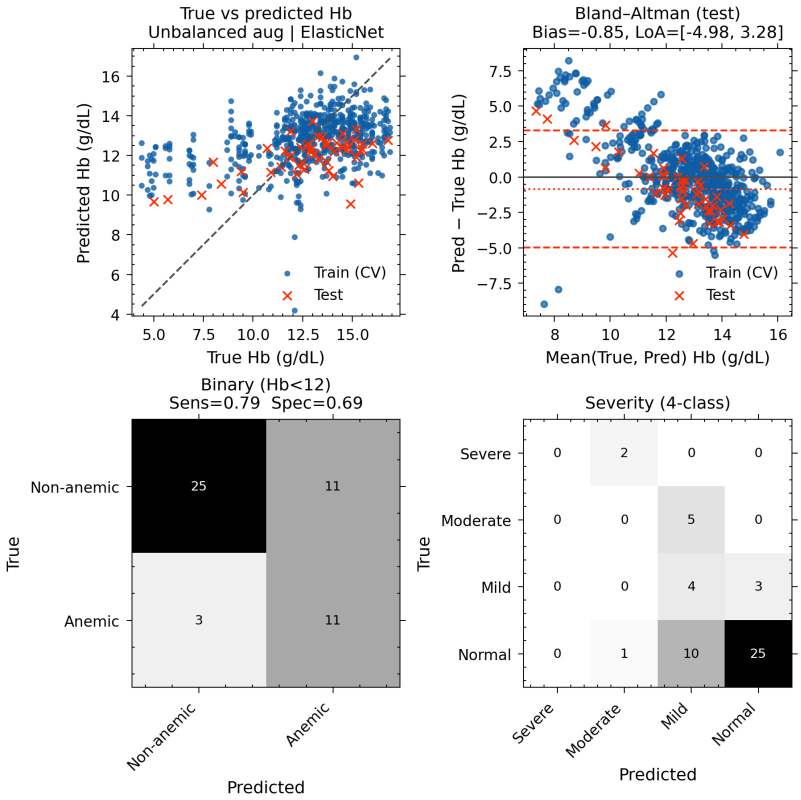


Figure A41. Supplementary evaluation plots for ElasticNet: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

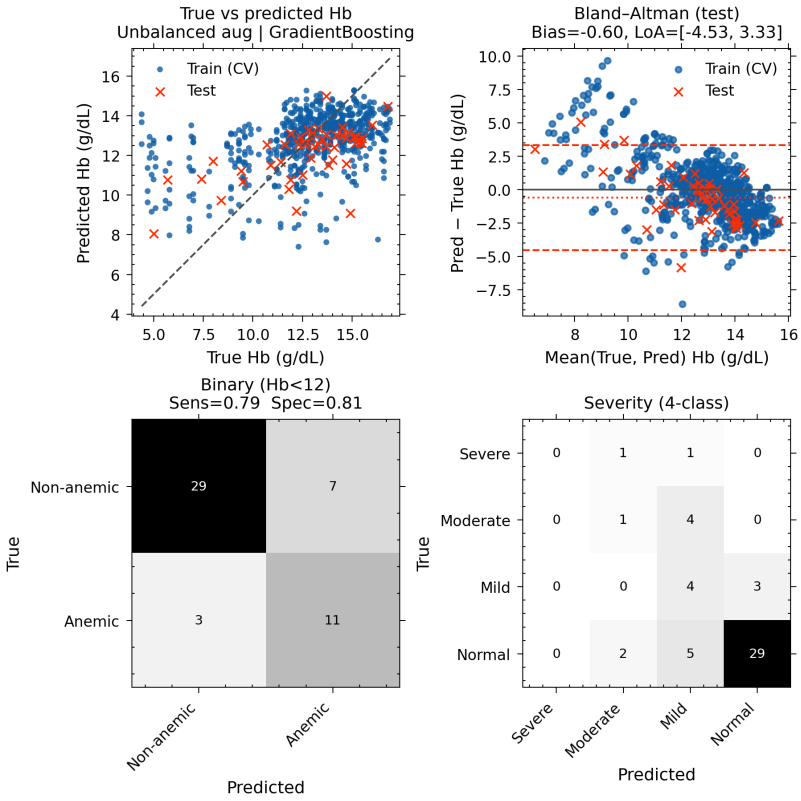


Figure A42. Supplementary evaluation plots for Gradient Boosting: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

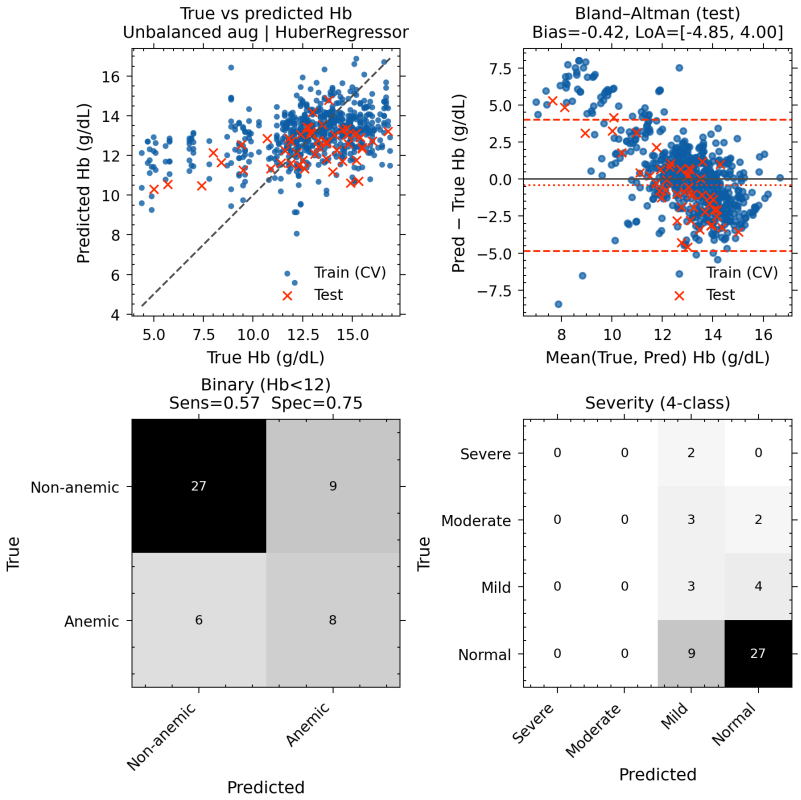


Figure A43. Supplementary evaluation plots for HUBER REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

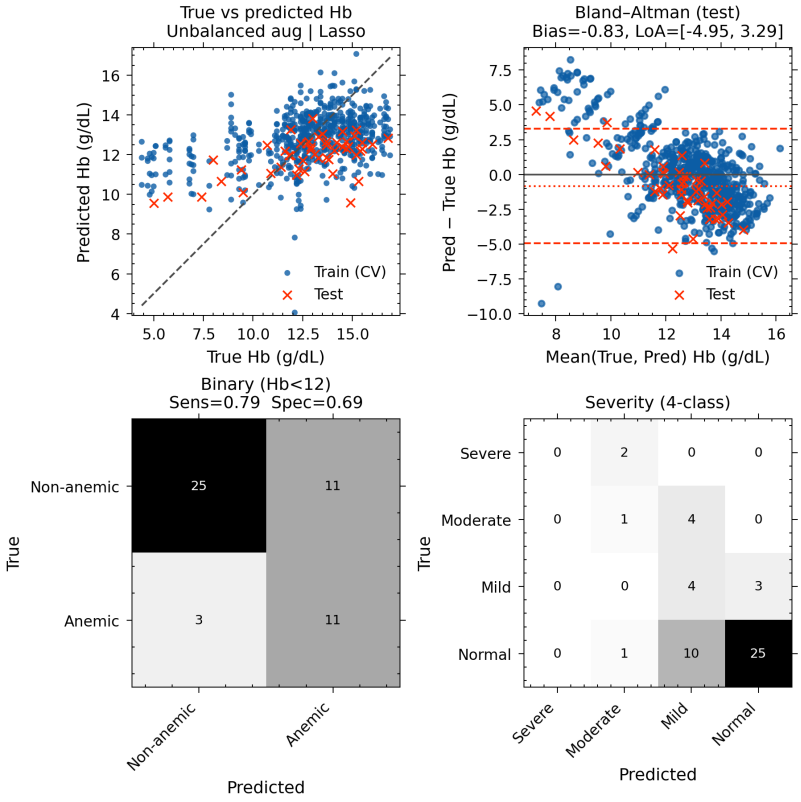


Figure A44. Supplementary evaluation plots for LASSO: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

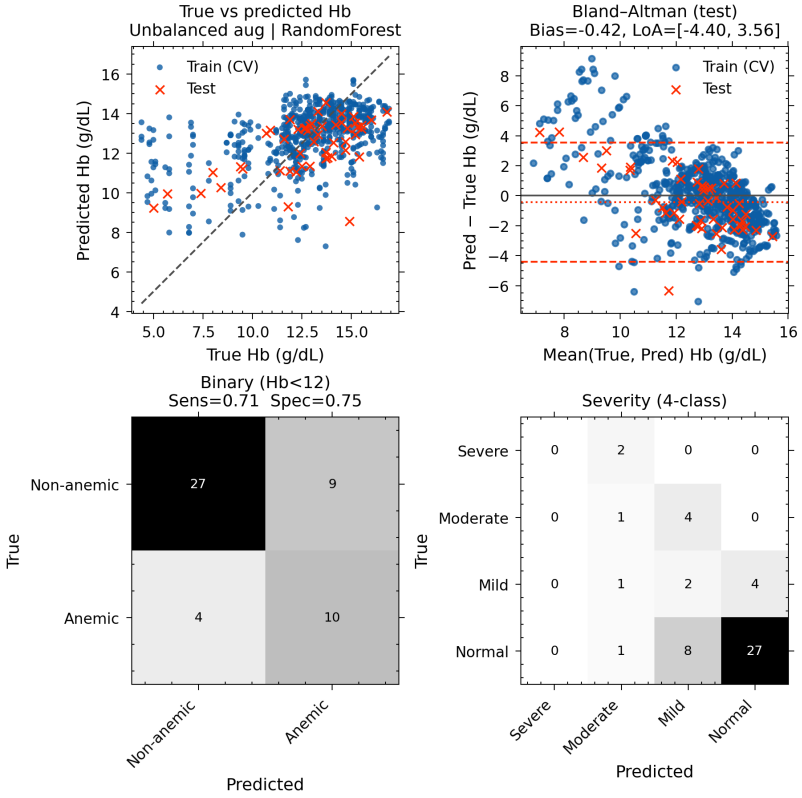


Figure A45. Supplementary evaluation plots for RANDOM FOREST: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

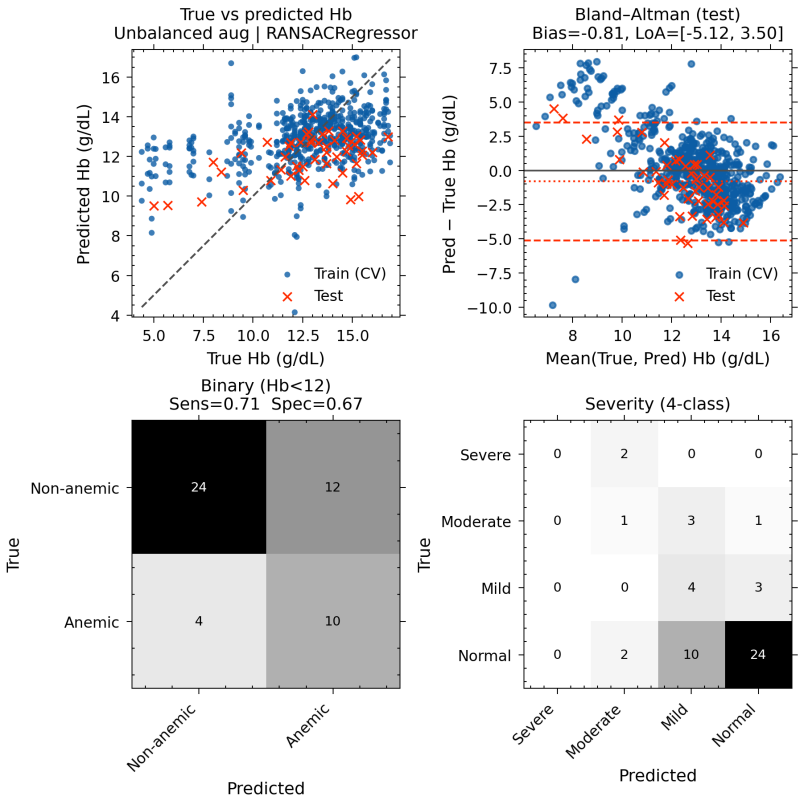


Figure A46. Supplementary evaluation plots for RANSAC REGRESSOR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

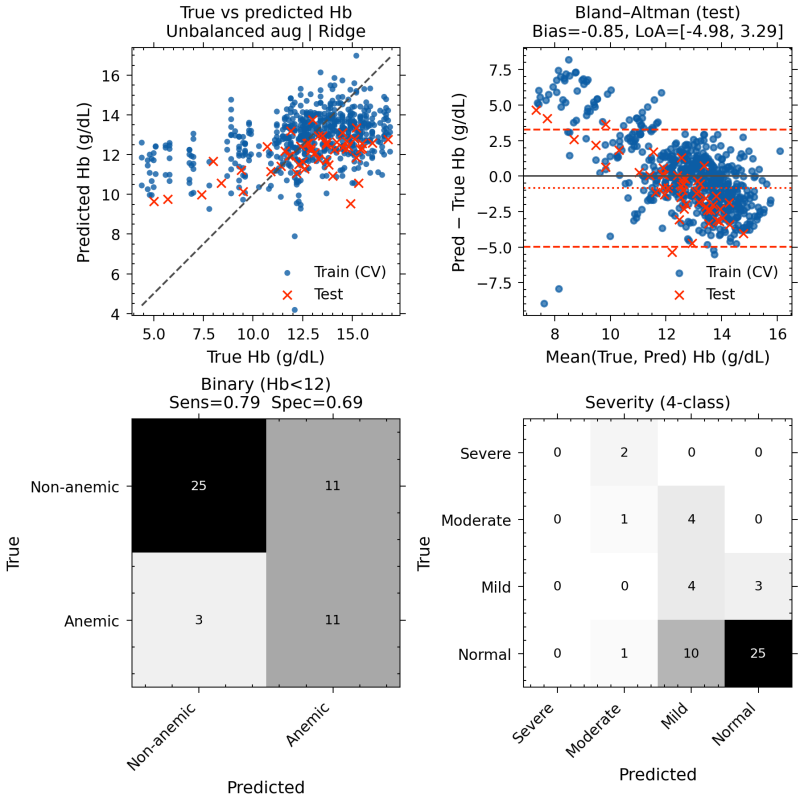


Figure A47. Supplementary evaluation plots for RIDGE: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

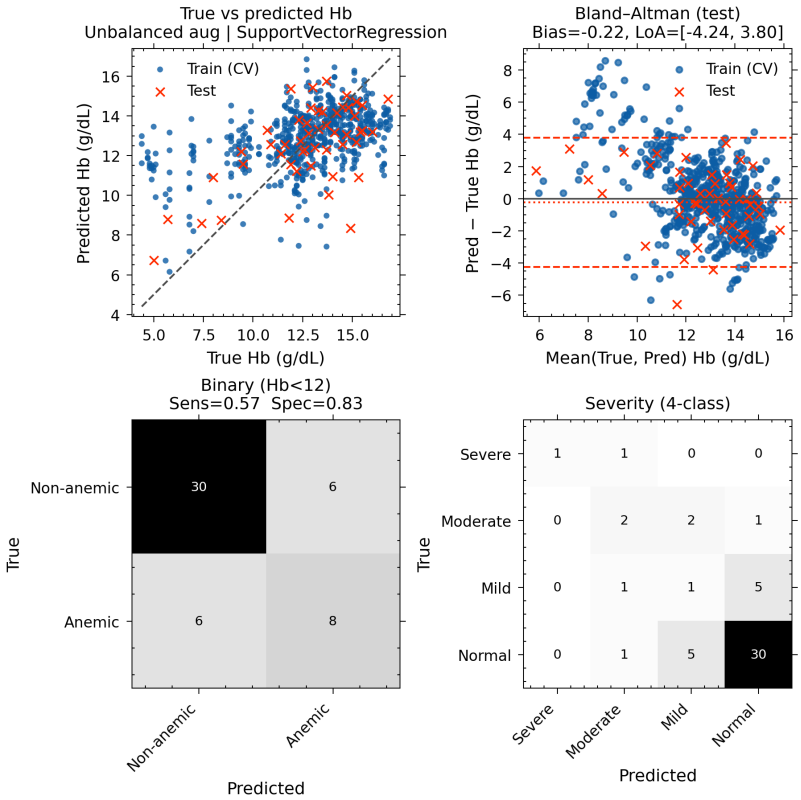


Figure A48. Supplementary evaluation plots for SVR: predicted vs. true hemoglobin, Bland–Altman agreement, and confusion matrices for binary and multiclass anemia classification under `unbalanced_aug` strategy.

References

1. World Health Organization. Operational Framework for Primary Health Care: Transforming Vision into Action. https://www.who.int/publications/i/item/9789240017832, 2020.
2. World Health Organization. Electricity in health-care facilities. https://www.who.int/news-room/fact-sheets/detail/electricity-in-health-care-facilities, 2023.
3. World Health Organization.; United Nations Children's Fund (UNICEF). Essential Services for Quality Care: Water, Sanitation, Hygiene, Health Care Waste and Electricity Services in Health Care Facilities: Global Progress Report. https://cdn.who.int/media/docs/default-source/wash-documents/wash-in-hcf/who_unicef_washwasteelectricityinhcfglobalprogressreport2025_web.pdf, 2025. Accessed 2026-02-10.
4. Ajuebor, O.; Boniol, M.; McIsaac, M.; Onyedike, C.; Akl, E.A. Increasing access to health workers in rural and remote areas: what do stakeholders' value and find feasible and acceptable? *Hum. Resour. Health* **2020**, *18*, 77. https://doi.org/10.1186/s12960-020-00519-2.
5. International Monetary Fund. Digital Interventions in the Health Sector: Country Cases and Policy Discussions. https://www.imf.org/-/media/files/publications/imf-notes/2023/english/insea2023004.pdf, 2023. IMF Notes; accessed 2026-02-10.
6. World Health Organization. GIDH Webinar 27 (2025): Global Digital Health Atlas overview and statistics. https://cdn.who.int/media/docs/default-source/digital-health-documents/gidh-webinar-27---11--2025-for-publishing.pdf, 2025. Slides; accessed 2026-02-10.
7. Global Digital Health Monitor. The State of Digital Health 2023. https://static1.squarespace.com/static/5ace2d0c5cfd792078a05e5f/t/656f97969301e337ada15270/1701812128734/State%2Bof%2BDigital%2BHealth_2023.pdf, 2023. Accessed 2026-02-10.
8. Siyam, A.; et al. The burden of recording and reporting health data in primary health care facilities in five low- and lower-middle-income countries. *BMC Health Services Research* **2021**, *21*, 901. https://doi.org/10.1186/s12913-021-06652-5.
9. Maita, K.C.; Maniaci, M.J.; Haider, C.R.; Avila, F.R.; Torres-Guzman, R.A.; Borna, S.; Lunde, J.J.; Coffey, J.D.; Demaerschalk, B.M.; Forte, A.J. The impact of digital health solutions on bridging the health care gap in rural areas: A scoping review. *Perm. J.* **2024**, *28*, 130–143. https://doi.org/10.7812/TPP/23.134.
10. Mwogosi, A.; Mambile, C. Digital ecosystems for healthcare communication and collaboration: A scoping review. *Digit. Health* **2025**, *11*, 20552076251377933. https://doi.org/10.1177/20552076251377933.
11. Kim, M.K.; Rouphael, C.; McMichael, J.; Welch, N.; Dasarathy, S. Challenges in and opportunities for electronic health record-based data analysis and interpretation. *Gut Liver* **2024**, *18*, 201–208. https://doi.org/10.5009/gnl230272.
12. Brotherton, T.; Brotherton, S.; Ashworth, H.; Kadambi, A.; Ebrahim, H.; Ebrahim, S. Development of an offline, open-source, electronic health record system for refugee care. *Front. Digit. Health* **2022**, *4*, 847002. https://doi.org/10.3389/fdgth.2022.847002.
13. Ahmer, H.; et al. Applying the principles for digital development to improve a primary care digital health intervention at scale: A case study. *PLOS Digital Health* **2024**. https://doi.org/10.1371/journal.pdig.0000434.
14. Haggenmüller, V.; Bogler, L.; Weber, A.C.; et al. Smartphone-based point-of-care anemia screening in rural Bihar in India. *Communications Medicine* **2023**, *3*, 38. https://doi.org/10.1038/s43856-023-00267-z.
15. World Health Organization. Anaemia. https://www.who.int/news-room/fact-sheets/detail/anaemia, 2025. Fact sheet; updated 2025-02-01; accessed 2026-02-10.
16. Health Level Seven International (HL7). FHIR Release 5 (R5): Specification. https://hl7.org/fhir/R5/, 2023. Accessed 2026-02-10.
17. Yakimov, B.; Buiankin, K.; Denisenko, G.; Bardadin, I.; Pavlov, O.; Shitova, Y.; Yuriev, A.; Pankratieva, L.; Pukhov, A.; Shkoda, A.; et al. Dataset of human skin and fingernails images for non-invasive haemoglobin level assessment. *Nature News* **2024**. https://doi.org/10.1038/s41597-024-03895-9.
18. Clinical and Laboratory Standards Institute. Collection of Diagnostic Venous Blood Specimens. https://clsi.org/shop/standards/pre02/, 2025. Date of publication: February 13, 2025. PDF ISBN: 978-1-68440-271-7.
19. Lima-Oliveira, G.; Lippi, G.; Salvagno, G.L.; Picheth, G.; Guidi, G.C. Laboratory Diagnostics and Quality of Blood Collection. *Journal of Medical Biochemistry* **2015**, *34*, 288–294. https://doi.org/10.2478/jomb-2014-0043.
20. Yaseen, M. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector, 2024, [arXiv:cs.CV/2408.15857].
21. NVIDIA Corporation. NVIDIA Jetson Orin Nano Developer Kit Product Brief. https://nvdam.widen.net/s/zkfqjmtds2/jetson-orin-datasheet-anano-developer-kit-3575392-r2, 2023. Accessed: 2025-12-20.

22. Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M.W.; Keutzer, K. A Survey of Quantization Methods for Efficient Neural Network Inference **2021**. [arXiv:cs.CV/2103.13630].