CONCEPT BOTTLENECK MODEL FOR TRUSTWORTHY ANEMIA DIAGNOSIS

Long Hoang Vu
Trustworthy AI
George Washington University
G23407544
long.vu@gwu.edu

1 Introduction

Clinical laboratories confirm anemia by reviewing blood smears under a microscope and a complete blood count (CBC) test. Deep networks can automate this task with high accuracy, yet they behave fully as black boxes: they classify "anemic" or "healthy" without providing explainable evidence that a pathologist would cite (e.g. *microcytic, hypochromic cells + low HGB*).

Lack of **explainability** is a direct trust barrier in medical tasks: When an AI system cannot justify its final verdict, clinicians would hesitate to rely on it in real triage workflows.

I propose a hybrid **Concept-Bottleneck Model (CBM)** (Fig. 1) that couples a vision network backbone with an explicit layer of defined, explainable concepts–words that pathologists use in the morphology reports–and augment them with ground truth CBC test results. The model has to predict these concepts at the bottleneck before an MLP gives the final diagnosis, allowing for clear and verifiable explanations. Accuracy, AUROC, and F1 will be used as quantitative trust measures to determine the quality of your explainable concepts and performance of the hybrid CBM.

The contributions:

- **Hybrid CBM**–ResNet34 predicts 15 morphology flags, concatenated with 5 CBC values, then a 4-layer MLP head outputs the diagnosis.
- **Data cleaning**–I parsed morphology flag keywords from morphology reports. Then, I identified and removed 124 slides with implausible CBC values and used the five CBC indices that appear in every report.
- **Quantitative trust measurement**–The model achieves **91.7% test accuracy**, $F_1 = \mathbf{0.92}$, and high concept fidelity (**AUROC** $\geq$ **0.9** on microcytic & hypochromic) while providing explanations with explicit concepts that clinicians can audit.

2 Related Works

Concept Bottleneck Models. [Koh et al., 2020] proposed *Concept Bottleneck Models* (CBMs), an architecture that forces the network to predict a user-defined vector of interpretable concepts c before producing the final task label y. Because c lies in a human-understandable space (e.g. “has wings”, “is red” for birds), CBMs enable *faithful post-hoc explanations*: one can inspect the predicted concepts or manually intervene by flipping them. I adopt a custom *hybrid* CBM architecture—predicted visual concepts *plus* ground-truth non-visual features (CBC indices)—to mirror the real diagnostic workflow and to yield explanations clinicians can verify directly.

The AneRBC benchmark. [Nishat et al., 2025] introduce the *Anemic Red-Blood-Cell* (AneRBC) benchmark to facilitate computer-aided anemia research. The release contains two subsets: *AneRBC-I* (1,000 full-field smears at 1224×960px, each paired with a CBC report and a narrative morphology assessment) and *AneRBC-II* (12,000 cropped tiles produced by uniformly subdividing AneRBC-I for CNN compatibility). They benchmark four canonical CNNs, MobileNetV2, ResNet152V2, VGG16, and InceptionV3, reporting up to **91**% (VGG16, trained from scratch), but leave

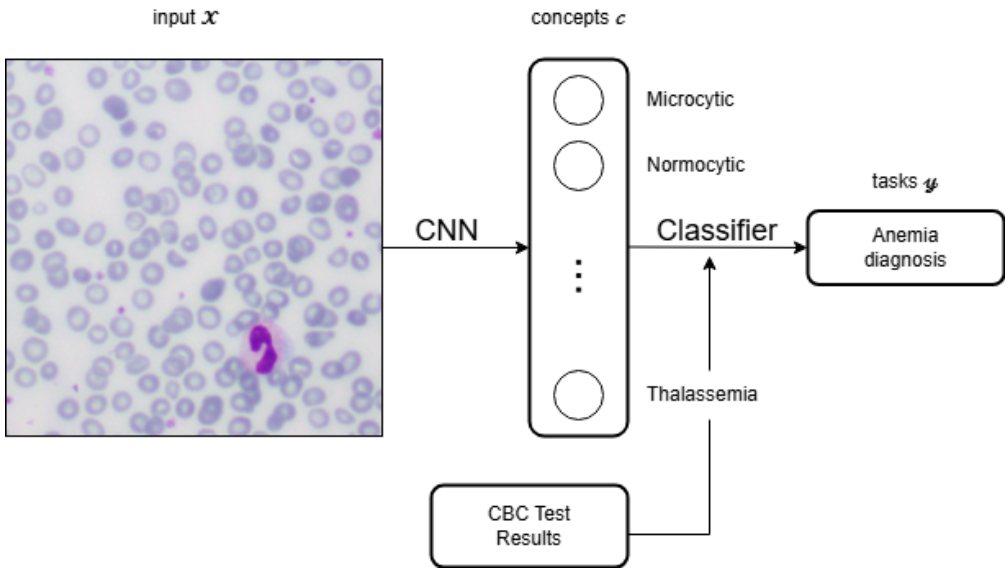

Figure 1: Diagram of the hybrid CBM architecture. The model uses a ResNet34 backbone to predict morphological flags, which are combined with CBC annotations before being passed to an MLP for final classification.

model decisions unexplained. No concept supervision or CBC integration is attempted. My work builds directly on AneRBC-I, retains its raw resolution, and augments it with an interpretable hybrid Concept-Bottleneck architecture that exposes both morphology keywords and CBC indices.

Benchmark CNNs on AneRBC-I. The original AneRBC paper reports four off-the-shelf CNNs— MobileNetV2, ResNet152V2, VGG16 and InceptionV3 from scratch on the AneRBC-I smears. All four reach >*98% training accuracy*, yet the best *test* score is only **91%** (VGG16, $F_1 = 0.93$), with the others ranging from 77–89%. While these CNNs set a high numeric baseline, they provide no *clinically meaningful* explanation of their decisions, creating a trust deficit as these models are untrustworthy on medical tasks. My hybrid CBM attains a comparable **91.7%** test accuracy and $F_1 =$ **0.92** while flagging clinically meaningful concepts, closing the performance and trust gaps.

3 Data & Pre-processing

AneRBC-I is used to train the hybrid CBM. Full details are in Table 1.

4 Method

4.1 Hybrid Concept–Bottleneck Architecture

Our pipeline (Fig. 2) splits the task into two stages inside a single end-to-end network:

1. **Concept predictor** g_θ — a vision backbone that maps an 224×224 smear image $\mathbf{x}$ to $\hat{\mathbf{c}}_{\text{morph}} \in (0,1)^K$, where each dimension corresponds to a clinically meaningful morphology keyword (Table 1, K=15).
2. **Label head** h_ϕ — an MLP that receives the concatenation of predicted morphology flags and *observed* CBC indices $\mathbf{c}_{\text{cbc}} \in \mathbb{R}^5$ and outputs logits for the binary label y (anemic / healthy).

The full forward pass is

$$\hat{\mathbf{c}}_{\text{morph}} = g_\theta(\mathbf{x}), \qquad \hat{y} = h_\phi\big([\hat{\mathbf{c}}_{\text{morph}} \parallel \mathbf{c}_{\text{cbc}}]\big).$$

Table 1: Summary of the AneRBC-I dataset characteristics and processing.

Characteristic	Details
Dataset Source	AneRBC-I dataset
Initial Size	1,000 slides, each with dimensions 1224×960 pixels.
Cleaning	Removed 124 slides identified as having biologically implausible CBC values (i.e., outside the 5 % to 95 % range).
Final Size	876 slides remaining after cleaning. The final dataset is approximately balanced with 50 % anemic and 50 % healthy samples.
Morphological Concepts	Utilized 15 keyword flags derived from pathologist reports. Flags that show up <5% in all of the reports are pruned. The final list of flags: anisocytosis, elliptocytes, hypochromic, microcytic, monocytosis, neutrophilia, normochromic, normocytic, platelets decreased, platelets increased, polychromasia, reactive lymphocytes, target cells, tear drop cells, thalassemia
CBC Indices	Included numeric values for Hemoglobin (HGB), Hematocrit (HCT), Red Blood Cell count (RBC), Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin (MCH). These were treated as unsupervised features.
Data Split	The dataset was partitioned into training (70 %), validation (15 %), and testing (15 %) sets. The split was stratified based on the anemic/healthy label to maintain class distribution.

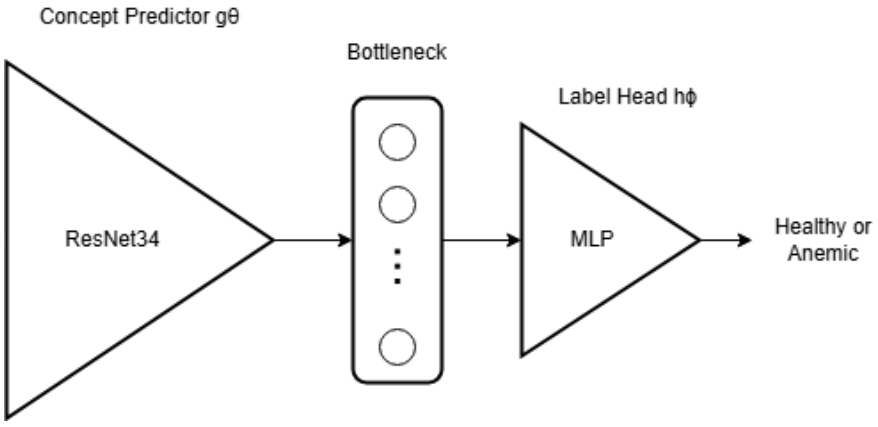

Figure 2: Components of the hybrid CBM. The Concept Predictor is ResNet34, which predicts concepts in the bottleneck. The Label Head is a 4-layer MLP that classifies the anemia diagnosis.

4.2 Concept Predictor g_θ

I use a **ResNet–34** backbone (TORCHVISION "`ResNet34_Weights.DEFAULT`"). The final global-pooling feature (512-D) is fed to a linear layer with 15 sigmoids—one per morphology keyword (refer to Tab. 1). Layers up to `layer2` are frozen.[1]

4.3 Observed CBC Concepts

Five numeric indices (HGB, HCT, RBC, MCV, MCH) are concatenated to the bottleneck layer. They are *not* predicted by the vision network and therefore receive no concept-loss gradient.

4.4 Label Head h_ϕ

We use a **four-layer MLP** $256 \to 128 \to 64 \to 2$ with BatchNorm and ReLU activations:

`Linear - BN - ReLU - Linear - BN - ReLU - Linear - BN - ReLU - Linear`.

A Dropout layer ($p = 0.20$) follows the first two ReLUs to mitigate over-fitting; no dropout is used on the final projection.

4.5 Training Objective

Let $\mathbf{m} \in \{0,1\}^{K+5}$ be a mask whose first K entries are 1 (morphology) and last five entries are 0 (CBC)[2]. The loss is

$$\mathcal{L} \;=\; \lambda \underbrace{\big\langle \mathbf{m} \odot \mathrm{BCE}\big(\hat{\mathbf{c}}, \mathbf{c}\big) \big\rangle}_{\text{concept loss } \mathcal{L}_c} \;+\; \underbrace{\mathrm{CE}\big(\hat{y}, y\big)}_{\text{label loss } \mathcal{L}_y},$$

with $\lambda = 1$. I experimented with *pos-weighted* BCE to emphasise rare concepts; the best checkpoint uses uniform weighting.

4.6 Optimisation & Schedule

- **Optimiser:** AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.999$, weight-decay 10^{-4}).
- **Learning rates:** 10^{-3} for the label head, 3×10^{-4} for the concept head, 10^{-4} for fine-tuned backbone layers.
- **Scheduler:** REDUCELRONPLATEAU (monitor val-label-loss, factor 0.3, patience 3).
- **Weighted Concept Loss:** Assign higher weights to rarer flags (e.g., tear drop cells) to boost the label head. The formula is $loss = weight * BCE$ where $weight = 1/count_f$ and f is a morphological flag.

4.7 Trust metrics

To capture both predictive *performance* and *trustworthiness*, three complementary metrics are reported:

Table 2: Evaluation Metrics and Their Justifications.

Metric	Justification
Accuracy / F1(on held-out test set)	Baseline diagnostic safety. F1 is particularly important for a medical imaging task because we want to minimize false negatives, where a person's health is on the line.
Per-concept AUROC	Fidelity of explanations (comparing concept ground truth vs. model prediction). AUROC helps us understand whether we can trust its classification for a certain concept flag.

4.8 Implementation Details

Code is implemented in PyTorch 2.2 and runs on a single NVIDIA L4. Training for 30 epochs with batch size 32. Source code is available at `https://github.com/vulong2505/concept-bottleneck-anemia`.

[1] Freezing lower blocks preserves generic ImageNet features while preventing early over-fitting on the small smear dataset.

[2] CBC entries are masked as 0 so it returns no loss gradient

5 Experiment

Table 3 summarizes the performance of our **hybrid CBM**.[3]

Table 3: Performance results for the Hybrid Concept Bottleneck Model (CBM) on the test set.

Model	Test Acc	Test F1 Macro	Avg. AUROC
Hybrid CBM	0.917	0.920	0.58

6 Discussion

6.1 How the Hybrid CBM Improves Trust

1. **Evidence–aligned explanations.** The hybrid CBM outputs the same morphology descriptors (*microcytic*, *hypochromic*, *elliptocytes*...) that a haematologist writes in the morphology report, together with the numeric CBC indices from lab testing that trigger standard anaemia thresholds. A clinician can therefore verify at a glance *why* a slide is flagged—e.g. "low HGB ($8\,\mathrm{g/dL}$) + high probability of *microcytic* and *hypochromic* $\rightarrow$ iron-deficiency pattern"—rather than taking a raw probability at face value.
2. **Intervenability.** Because concepts are explicit, a user can flip a flag ("what if *microcytic*=0?") or overwrite a CBC value to obtain a counterfactual diagnosis—useful in double-reading or edge cases.
3. **Performance *and* trustworthiness.** The 91.7% test accuracy matches or exceeds benchmark, black-box CNNs, while the AUROC on key concepts exceeds 0.90, demonstrating that interpretability is achieved *without* sacrificing accuracy. The result is a performant model that has a highly explainable diagnosis.

6.2 Limitations

- **Rare-concept fidelity.** Rare flags on the long-tail end such as *tear-drop cells* ($6.7\,\%$ prevalence) yield AUROCs < 0.25, dragging the macro average to 0.58. Clinically, these findings are secondary, but their poor fidelity reveals the data scarcity problem common in medical AI. The weighted concept loss was used to combat the imbalanced dataset, but poor fidelity still persists.
- **Single-centre dataset.** AneRBC-I comes from one institution, one staining protocol, and one scanner. External validity across labs remains untested. Additionally, the dataset was noisy with erroneously annotated data–approximately 12.4% of the dataset had erroneous CBC reports and was pruned for a cleaner dataset.

6.3 Qualitative Analysis

To illustrate *how* the Hybrid CBM communicates its reasoning, three representative test slides are shown: a **true–positive** (TP), **true–negative** (TN) and **false–negative** (FN). For each case, I present (i) the smear image, (ii) the raw CBC report, (iii) the *predicted* concept vector ($\hat{\mathbf{c}}_{\text{morph}}$) and (iv) the final anemia probability $p(\text{anemic})$ (Figs. 3–5).

Take-aways. **(i)** For confident cases (TP, TN) the top morphology scores align with clinical reading and the CBC values reinforce the verdict, supporting user trust. **(ii)**Borderline errors (FN) are explicable: the probability is low because the model expresses uncertainty in key concepts rather than silently failing. **(iii)**With a CBM, clinicians can override or investigate by editing concept flags or CBC numbers, something that isn't possible basic CNN.

Interpretation workflow. For each slide the pathologist (i) confirms or rejects the top-scoring concepts, (ii) compares the raw CBC numbers against reference intervals, and (iii) accepts or overrides the model verdict (e.g. "model missed subtle hypochromia" in Fig. 5).

Because CBC thresholds act as hard clinical constraints, a contradictory combination—such as a low HGB but a high predicted *normocytic/normochromic* score—immediately signals an unreliable explaination and prompt for a manual review.

Limitations highlighted by examples. The FN slide (Fig. 5) illustrates two open issues. First, concept AUROC is weakest for *low-prevalence* flags such as *tear_drop_cells* ($7\,\%$), *monocytosis* ($6\,\%$), and *reactive_lymphocytes* ($6\,\%$)

[3] Per-concept AUROC are produced in Appendix A.

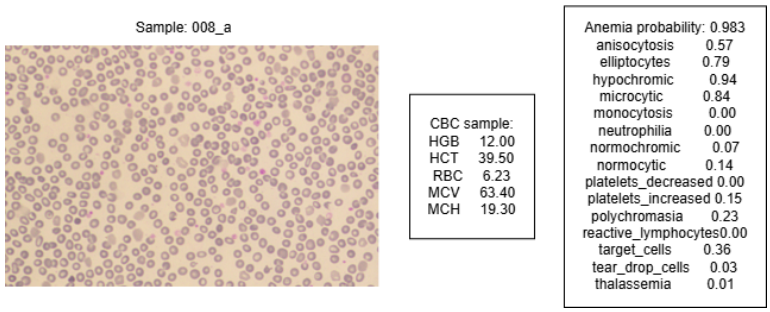

Figure 3: **True–positive** (slide `008_a`). The model assigns $p(\text{anemic}) = \mathbf{0.983}$. Top concept probabilities are *hypochromic* 0.94, *microcytic* 0.84, *anisocytosis* 0.57 and *elliptocytes* 0.79, matching the visual impression of pale, small, shape-variant cells. The CBC confirms the diagnosis (HGB=$12.0\,\mathrm{g/dL}$ is borderline, but MCV=$63.4\,\mathrm{fL}$ and MCH=$19.3\,\mathrm{pg}$ are markedly low), reinforcing the model's explanation.

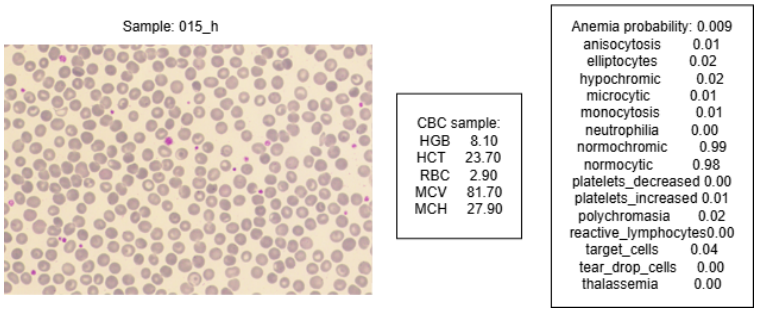

Figure 4: **True–negative** (slide `015_h`). $p(\text{anemic}) = \mathbf{0.009}$. Concepts *normocytic* 0.98 and *normochromic* 0.99 dominate, while *microcytic*/*hypochromic* are near zero. CBC indices lie well inside reference ranges (HGB=$8.1\,\mathrm{g/dL}$, HCT=$23.7\,\%$ would usually flag anemia, but the smear shows normal morphology—illustrating how the model balances visual and lab information to avoid over-calling).

(see Table 4); the model therefore hesitates when those rare features matter. Second, borderline CBC values (e.g. MCV 72 fL) challenge the flagged concepts. Addressing these limitations will likely require additional labelled slides to balance the rare morphology classes.

Overall, these qualitative examples demonstrate that the Hybrid CBM provides *actionable* explanations–linking model decisions to well-understood morphology concepts and lab values.

7 Conclusion

A concept-bottleneck architecture, paired with cleaned CBC inputs, lifts performance to 91.7% while delivering meaningful concept-level explanations. This shows one path toward *building* trust in medical AI: forcing models to "show their work" in terms clinicians already understand.

A Concept Fidelity

References

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020. URL `https://arxiv.org/abs/2007.04612`.

Chinki Nishat, Jhuma Sankar, Manish Kumar, Shobha Singh, Rakesh Lodha, and Sushil K. Kabra. Exploring the role of human metapneumovirus in acute respiratory infections in children in an outpatient setting in North India. *Tropical Doctor*, 55(2):166–171, apr 2025. doi: 10.1177/00494755241285109.

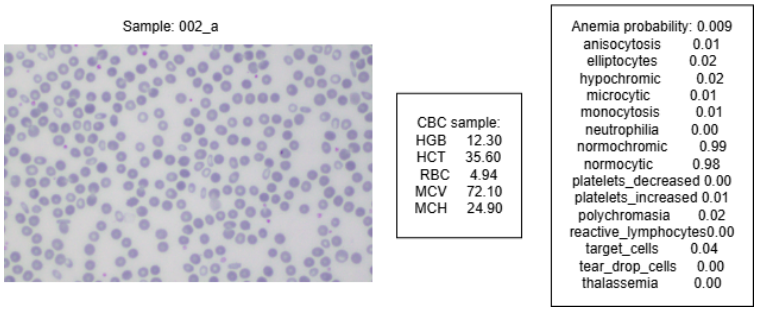

Figure 5: **False–negative** (slide `002_a`). The model outputs $p(\text{anemic}) = \mathbf{0.473}$ (below the 0.50 threshold). Predicted morphology is mixed (*microcytic* 0.32, *hypochromic* 0.29, yet *normochromic* 0.63), leading to an uncertain verdict despite MCV=72.1 fL (low) in the CBC. Visual inspection reveals subtle pallor the model underestimates, highlighting a limitation for borderline cases and motivating future work on cell-percentage concepts.

Table 4: Per-concept and Macro-average Area Under the Receiver Operating Characteristic Curve (AUROC) results for the Hybrid CBM.

Concept Name	**AUROC**
microcytic	0.895
normocytic	0.146
normochromic	0.121
hypochromic	0.944
elliptocytes	0.729
target cells	0.488
tear drop cells	0.224
anisocytosis	0.223
polychromasia	0.536
neutrophilia	0.605
monocytosis	0.706
reactive lymphocytes	0.920
platelets increased	0.644
platelets decreased	0.670
thalassemia	0.900
Macro-average	**0.583**