NutriScreener: Retrieval-Augmented Multi-Pose Graph Attention Network for Malnourishment Screening

Misaal Khan[1,2], **Mayank Vatsa**[1], **Kuldeep Singh**[2], **Richa Singh**[1]

[1]Indian Institute of Technology Jodhpur, Rajasthan, India
[2]All India Institute of Medical Sciences Jodhpur, Rajasthan, India
{khan.9, mvatsa, richa}@iitj.ac.in, singhk@aiimsjodhpur.edu.in

Abstract

Child malnutrition remains a global crisis, yet existing screening methods are laborious and poorly scalable, hindering early intervention. In this work, we present NutriScreener, a retrieval-augmented, multi-pose graph attention network that combines CLIP-based visual embeddings, class-boosted knowledge retrieval, and context awareness to enable robust malnutrition detection and anthropometric prediction from children's images, simultaneously addressing generalizability and class-imbalance. In a clinical study, doctors rated it 4.3/5 for accuracy and 4.6/5 for efficiency, confirming its deployment readiness in low-resource settings. NutriScreener was trained and tested on 2,141 children from AnthroVision and additionally evaluated on diverse cross-continent populations, including ARAN and an in-house collected CampusPose dataset, achieving 0.79 recall, 0.82 AUC, and significantly lower anthropometric RMSEs, demonstrating reliable measurement in unconstrained, pediatric settings. Cross-dataset results show up to 25% recall gain and up to 2.3 cm reduction in head circumference RMSE using demographically matched knowledge bases. NutriScreener offers a scalable and accurate solution for early malnutrition detection in low-resource environments.

Toolkit: https://www.iab-rubric.org/resources/healthcare-datasets/nutriscreener

Extended version — https://arxiv.org/abs/2511.16566

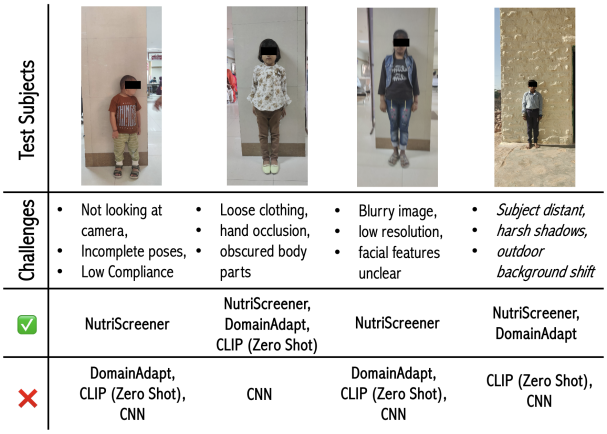


Figure 1: Robustness of *NutriScreener* to detect *malnourishment* across challenging real-world malnutrition scenarios.

Introduction

As of 2024, approximately 150 million children under five years of age globally suffer from stunting, and over 42 million from wasting, both direct outcomes of chronic and acute malnutrition. These conditions are recognized by the World Health Organization as leading causes of irreversible developmental impairment and mortality in early childhood[1] (World Health Organization 2024).

Malnutrition remains a pervasive and underdiagnosed global health crisis, especially in children, leading to long-term developmental issues and increased morbidity. These consequences are further magnified in low-resource settings where timely screening is often inaccessible (Khan et al. 2022). Conventional assessments rely on manual anthropometric measurements, such as mid-upper arm circumference (MUAC) tapes, weight-for-height charts, and questionnaires, which are labor-intensive, error-prone, and often delayed, making them unsuitable for rapid or scalable deployment (Janssen, Bouzembrak, and Tekinerdogan 2024; Khan et al. 2023). There is a pressing need for automated, reliable malnutrition screening that minimizes reliance on manual measurements.

Prior studies have explored malnourishment detection using facial and full-body imagery, but most facial-image approaches (Wang, He, and Long 2023; Tay et al. 2022) target elderly populations and do not generalize to children. Solutions like Microsoft's Child Growth Monitor (Welthungerhilfe 2025) require specialized hardware (infrared depth sensors), limiting accessibility in high-need regions. Most AI models (Janssen, Bouzembrak, and Tekinerdogan 2024) remain experimental and lack validation for routine clinical use, leaving traditional anthropometry as the clinical standard. Current models (Aanjankumar et al. 2025; Khan et al. 2024) trained on limited datasets, often exhibit majority-class bias, reducing sensitivity to malnourished cases. As no single viewpoint captures all diagnostic cues, robust screen-

[1]https://www.who.int/data/gho/data/themes/topics/joint-child-malnutrition-estimates-unicef-who-wb, last viewed: Aug. 1, 2025

ing demands integration across multiple image poses.

Motivated by these insights, we propose *NutriScreener*, a multi-pose malnutrition assessment framework that integrates Graph Attention Networks (GATs)(Veličković et al. 2018) with retrieval-augmented learning for robust classification and anthropometric estimation. Each subject is modeled as a graph whose nodes contain pose-wise embeddings extracted using the Contrastive Language–Image Pretraining (CLIP) image encoder (Radford et al. 2021), augmented with age metadata. We adopt CLIP due to its strong cross-domain generalization, with prior work showing that its pretrained semantic features implicitly encode fine-grained anatomical and morphological cues relevant to medical imaging (Yu et al. 2024), anthropometry (Khandelwal et al. 2024), and few-shot learning (Li et al. 2025). Graph Attention Network (GAT) layers then model inter-pose relationships, promoting cross-view consistency for both classification and regression tasks. To address the significant class imbalance characteristic of malnutrition datasets, we construct a population-level Knowledge Base (KB) and retrieve semantically similar samples using FAISS (Facebook AI Similarity Search) (Douze et al. 2024). Retrieved exemplars are used to boost minority-class predictions through a retrieval-informed augmentation mechanism. Finally, the retrieval output is adaptively fused with GAT predictions using a learned context-aware fusion layer that dynamically balances retrieval information and GAT confidence scores to improve both precision and recall. The contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first effort to integrate retrieval augmentation with multi-pose Graph Attention Networks for malnutrition screening, jointly performing anthropometric regression and binary classification from 2D image inputs.
- Our framework supports generalization to new populations by requiring only a few representative samples in the KB, addressing severe class imbalance and enabling practical deployment in low-resource settings.
- We propose a context-aware gated fusion mechanism to combine GAT and retrieval outputs based on model confidence and local density, enabling robustness under pose variability, domain shifts, and label imbalance.
- We validate *NutriScreener* through a real-world clinician study, confirming its accuracy, efficiency, and deployment readiness. The toolkit, labeled KB and CampusPose dataset are released to support field use and research.

Related Work

AI-based malnutrition assessment remains underexplored due to limited public datasets and challenges in adapting vision models to low-resource clinical settings (Khan et al. 2023, 2022). Early work relied on small-scale data: ARAN (Mohammed Khan et al. 2025) includes only 512 Kurdish children from a single region, while AnthroVision (Khan et al. 2024) contains 2,141 multi-view samples of Indian children, but is geographically restricted to Jodhpur, Rajasthan. A recent multitask model (DomainAdapt) (Khan et al. 2024) trained on AnthroVision showed improved classification but suffered from low malnourishment recall ($\sim$67%). These limitations highlight the need for more diverse datasets and tailored methods for robust screening.

Foundation models like CLIP (Radford et al. 2021), trained on large image–text data, exhibit strong generalization across visual tasks and transfer well to medical domains. MedCLIP (Wang et al. 2022), adapted using unpaired radiology image–report pairs, outperforms prior methods in zero-shot prediction, classification, and retrieval. Such models offer rich semantic features and robustness to domain shifts, making them well-suited for anthropometric prediction across diverse populations.

Graph neural networks (GNNs) and multi-view learning provide structured approaches to model body morphology from images. GNNs represent body parts or landmarks as graph nodes, capturing relational constraints to improve shape inference. For instance, (Li et al. 2020) models multiscale joint relations for realistic pose prediction, while (Kolotouros, Pavlakos, and Daniilidis 2019) uses graph convolutions on body meshes to regress 3D shape. Multi-view methods further enhance anthropometric estimates by aggregating information across poses. Liu et al. (Liu, Sowmya, and Khamis 2018) show that even linear models achieve accurate height and MUAC prediction with multi-angle inputs. Together, GNNs and multi-view consistency improve robustness in image-based anthropometry.

Retrieval-augmented learning addresses class imbalance by incorporating external exemplars during inference. RAC (Long et al. 2022; Liu et al. 2025) introduces a FAISS-based memory index to retrieve nearest neighbors for improved classification. COBRA (Das et al. 2025) further optimizes retrieval via mutual information to balance similarity and diversity, enhancing recall with minimal overhead.

Methodology

The architecture of our framework, *NutriScreener*, is shown in Figure 2, which consists of four core components: (1) a CLIP image encoder that extracts semantic features from each viewpoint, (2) a GAT that models inter-pose relationships to produce consistent multi-view predictions, (3) a retrieval module that queries a curated KB using global embeddings to provide representative support, and (4) a context-aware fusion mechanism that adaptively combines GAT and retrieval predictions based on confidence and local embedding density. This design enables generalization under pose variability, class imbalance, and domain shifts.

Problem Statement: Given a set of multi-pose RGB images $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,P_i}\}$ for a subject i, where each $x_{i,j}$ corresponds to the j^{th} available viewpoint (e.g., frontal, lateral, back, selfie), and a scalar age value a_i, the task is to jointly predict: (i) a binary nutritional status label $y_i^{\text{class}} \in \{0, 1\}$, and (ii) a 4-dimensional vector of anthropometric measurements $y_i^{\text{reg}} \in \mathbb{R}^4$, including Height (Ht.), Weight (Wt.), MUAC, and Head Circumference (HC).

For each subject i, we obtain P pose images. Each is embedded via a pretrained CLIP encoder into a 1024D vector. These are concatenated with age to form node features $v_{i,j} \in \mathbb{R}^{1025}$, which are used as nodes in a fully connected

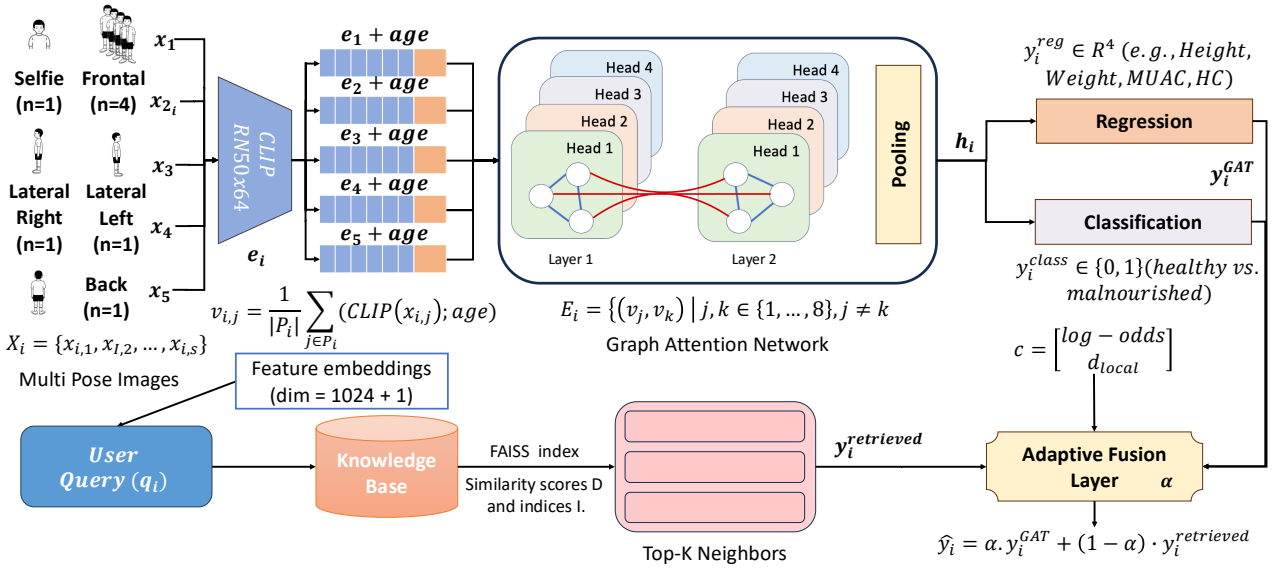


Figure 2: Architectural overview of *NutriScreener*. The system takes a subject's multi-pose images and age as input. It extracts pose-wise visual embeddings using a CLIP-based encoder and aggregates them, then passes the aggregated embeddings through a graph attention network to produce image-based predictions for both classification and regression tasks. In parallel, the same aggregated embeddings are used to query a FAISS-indexed knowledge base, yielding retrieval-based predictions. An adaptive fusion layer then combines the graph-based and retrieval-based outputs.

undirected graph. This design allows GATs to model inter-pose dependencies while retaining pose-specific signals.

To enhance sensitivity to malnourished cases and support cross-population generalization, we introduce a retrieval-augmented module. A global embedding is computed by averaging the pose embeddings. Appending age yields a 1025D RAC query, q_i, which retrieves top-k similar samples from a FAISS-based KB. Final predictions combine GAT and retrieval outputs using a learned weight. This adaptive fusion supports robustness to class imbalance, pose variability, and domain shift. These modules are described below.

Multi-Pose Embedding Extraction: Each image $x_{i,j}$ is passed through a frozen CLIP encoder (ResNet-50×64 variant) (Radford et al. 2021), producing a 1024-dimensional embedding $e_{i,j} \in \mathbb{R}^{1024}$. To incorporate subject metadata, we append the scalar age a_i to each embedding, forming enriched pose-level feature vectors: $v_{i,j} = [e_{i,j}; a_i] \in \mathbb{R}^{1025}$. The multi-pose design (top-left block in Figure 2) offers two advantages: (i) it mitigates the limitations of any single pose by aggregating redundant cues across views, and (ii) it enables generalization across variable capture conditions (e.g., occlusion or missing poses), critical for deployment in uncontrolled field settings.

Graph Construction and GAT Inference: The enriched pose-level embeddings $\{v_{i,1}, \dots, v_{i,P_i}\}$ are treated as nodes V_i of a fully connected undirected graph $G_i = (V_i, E_i)$. Each edge in E_i links a pair of poses $(v_{i,j}, v_{i,k})$ with $j \neq k$. This representation enables pairwise message exchange between all poses, allowing the model to reason over inter-pose correlations such as asymmetrical fat loss or posture-related distortions, both of which are often indicative of malnutrition. The resulting graph is processed using a two-layer GAT, where each layer applies multi-head self-attention to selectively aggregate information from neighboring nodes. The final node representations are then globally pooled to obtain a subject-level embedding h_i, which is subsequently passed to task-specific heads for regression and classification. By leveraging cross-pose attention, the GAT module improves robustness to pose imbalance and occlusion while also enhancing interpretability through attention weights computed across views.

Knowledge Base Construction

To support retrieval-based inference, we construct an in-house **Malnutrition Knowledge Base (MalKB)** of pose-level embeddings and ground-truth labels from clinically collected data. The KB consists of 1984 multi-view RGB images captured from 248 pediatric subjects.

Image Capture Protocol: Images were acquired using a standard consumer-grade smartphone (OnePlus Nord) equipped with a Sony IMX586 sensor (48MP, f/1.75 aperture, OIS/EIS). Capture guidelines prescribed an approximate subject distance of 165 cm and a camera height of 50 inches, but minor variability was allowed to improve model robustness. Each subject was photographed across eight views: four frontal, one lateral-left, one lateral-right, one posterior, and one frontal selfie.

Ground-Truth Labels: Trained healthcare workers recorded anthropometric measurements, including Height (H), Weight (W), Middle-Upper Arm Circumference (MUAC), and Head Circumference (HC), using standardized clinical protocols. All data entries were logged via a GUI-assisted

tool to ensure consistency. The resulting dataset forms the retrieval corpus for nearest-neighbor lookups during inference. Each subject's global pose-averaged embeddings and labels are indexed using FAISS.

Retrieval-Augmented Classification

To improve sensitivity to the underrepresented malnourished class, we augment the GAT classifier with a FAISS-based retrieval head. For each subject i, we first compute a global query embedding $q_i = \frac{1}{P_i}\sum_{j=1}^{P_i} v_{i,j} \in \mathbb{R}^{1025}$. Then retrieve its top-$k$ nearest neighbors from the KB, obtaining cosine distance $\{d_j\}_{j=1}^{k}$ and corresponding binary labels $\{y_j^{\text{kb}}\}_{j=1}^{k}$. We normalize the cosine distances with a temperature-scaled softmax,

$$\tilde{w}_j \;=\; \frac{\exp(-d_j/\tau)}{\sum_{m=1}^{k}\exp(-d_m/\tau)},$$

Here, d_j is the cosine distance to the j^{th} retrieved neighbor, k is the number of retrieved samples, τ controls the sharpness of the softmax, and $\tilde{w}_j$ denotes the normalized retrieval weight. We then apply a class-specific boost

$$\omega_j \;=\; \begin{cases}\gamma, & y_j^{\text{kb}} = 1 \text{ (malnourished)},\\ 1, & y_j^{\text{kb}} = 0 \text{ (healthy)},\end{cases}$$

Here, ω_j is a multiplicative class-specific factor that upweights malnourished neighbors by γ (when $y_j^{kb} = 1$) and leaves healthy neighbors unchanged (when $y_j^{kb} = 0$).

The adjusted weights are re-normalized across all neighbors m, $w_j = \tilde{w}_j\omega_j / \sum_m \tilde{w}_m\omega_m$, and the retrieval-based prediction is

$$y_i^{\text{retrieved}} = \sum_{j=1}^{k} w_j\, y_j^{\text{kb}}.$$

Next, we compute an auxiliary context vector

$$c_i = \left[\log\tfrac{p_i}{1-p_i},\ \bar{d}\right],$$

where $p_i = \sigma(y_i^{\text{GAT}})$ ($\sigma(\cdot)$ = sigmoid) and $\bar{d} = \frac{1}{k}\sum_j d_j$. A small MLP then predicts a fusion coefficient $\alpha \in [0,1]$ from $[\,y_i^{\text{GAT}},\ y_i^{\text{retrieved}},\ c_i\,]$, and we form the final logit

$$\hat{y}_i^{\text{CLS}} = \alpha^{\text{CLS}}\, y_i^{\text{GAT}} + (1-\alpha^{\text{CLS}})\, y_i^{\text{retrieved}}$$

This adaptive fusion automatically shifts weight toward retrieval when the KB is dense around q_i, and relies on the GAT when neighbors are sparse.

Retrieval-Augmented Regression

Similar to classification, we compute the global query embedding q_i for subject i by averaging their pose-level embeddings and use FAISS to retrieve the top-k most similar subjects from the KB. Each retrieved sample contributes a ground-truth vector $y_j^{\text{reg}} \in \mathbb{R}^4$, containing four measurements: Ht, Wt, MUAC, and HC.

The cosine distances $\{d_1, \ldots, d_k\}$ are passed through a softmax to produce normalized weights $\{w_1, \ldots, w_k\}$, for retrieval-based regression:

$$y_i^{\text{retrieved}} = \sum_{j=1}^{k} w_j \cdot y_j^{\text{reg}}.$$

This estimate is fused with the GAT-based regression output y_i^{GAT} using a learnable scalar fusion coefficient $\alpha \in [0,1]$, yielding the final prediction:

$$\hat{y}_i^{\text{reg}} = \alpha^{\text{reg}} \cdot y_i^{\text{GAT}} + (1-\alpha^{\text{reg}}) \cdot y_i^{\text{retrieved}}$$

All retrieval hyperparameters (k, τ_{class}, γ, τ_{reg}) were tuned on the AnthroVision validation set. Here, k is the number of retrieved neighbors, τ_{class} and τ_{reg} control similarity weighting for classification and regression, and γ upweights malnourished exemplars. Fusion coefficients α^{CLS} and α^{reg} are learned during training.

Experiments

Datasets: We evaluate on three cross-continental multi-pose datasets comprising child and adult subjects. Each sample includes images from various anatomical views (e.g., frontal, lateral, back) with corresponding demographic and anthropometric data.

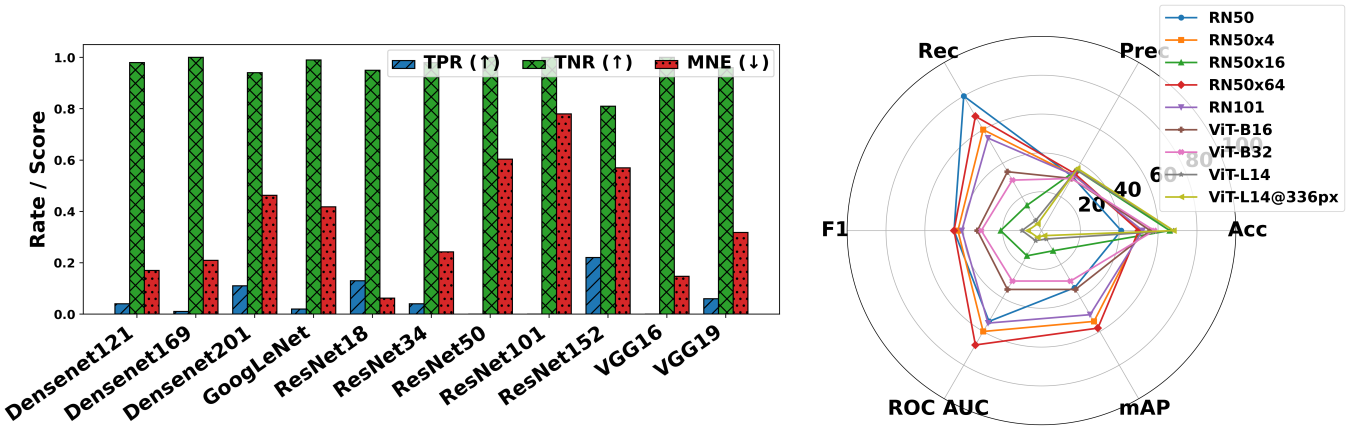
- AnthroVision (Khan et al. 2024): Our primary dataset with $2,141$ children from clinical and community settings, containing multi-pose images and detailed labels (gender, Ht (cm), Wt (kg), HC (cm), waistline (cm), MUAC (cm), and age (months)). Binary malnutrition labels are derived from WHO z-score thresholds (World Health Organization 2006). It serves as the main dataset for training, validation, and ablation, and is the largest available with malnutrition and anthropometric labels.
- The Age-Restricted Anonymized (ARAN) (Mohammed Khan et al. 2025): Contains 512 children (16–98 months) from two hospitals (H1: 404, H2: 108), with four face-anonymized views and measurements for Ht, Wt, waist, and HC.
- CampusPose: An in-house college-aged dataset (80 subjects) with multi-pose images and measurements for Ht, Wt, MUAC, HC, WC, and age. Malnutrition labels are based on BMI thresholds.

Ethical clearance for this data collection was obtained from the Institutional Ethics Committee (IEC) IIT-AIIMS Jodhpur, India.

Implementation Details

Training Objectives: We train the network end-to-end with a joint loss $\mathcal{L} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{reg}}$ (binary cross-entropy with logits loss (BCElogitloss) and mean squared error). The implementations are in PyTorch and PyTorch Geometric with 4-fold cross-validation with random seed (42). Experiments are performed on the workstation with Ubuntu 20.04 LTS along with four NVIDIA A100 GPUs (40GB each), 128GB RAM, Python 3.10, PyTorch 2.1.0, CUDA 11.8, and cuDNN 8.4. The model begins with a frozen CLIP (RN50×64) backbone and processes 1025-dimensional pose embeddings via a 2-layer GAT (8 heads, dropout 0.1). Training uses a batch size of 8 for 50 epochs (early stop) per fold with the Adam optimizer (1×10^{-3} learning rate).

Evaluation Metrics: Classification metrics include accuracy, precision, recall, F1 score, ROC AUC, and mAP, while regression is evaluated with RMSE and MAE. We also tried



(a) CNN classification metrics and mean normalized regression error

(c) CLIP encoder comparison.

Figure 3: Baseline model performance across (a) classification errors, (b) regression RMSE, and (c) encoder variants.

maximizing F1 and balanced accuracy; Youden's index gave the best trade-off between sensitivity and specificity.

Component Benchmarking: We benchmark different components of *NutriScreener* in the following manner:

1. CNN-Based Predictors: We evaluate a range of standard convolutional backbones for joint classification and regression. Architectures include ResNet (18/34/50/101/152), DenseNet (121/169/201), GoogLeNet, and VGG (16/19), all pretrained on ImageNet. Each pose image is processed independently.

2. CLIP Encoder Variants: We benchmark multiple CLIP variants (RN50, RN50×4, RN50×16, RN50×64, RN101, ViT-B16, ViT-B32, ViT-L14, ViT-L14@336px) for pose-wise classification. Each pose image is encoded as an individual input for CLIP Zero Shot Evaluation.

3. Pretrained vs. Fine-Tuned CLIP in End-to-End Pipeline: We test whether fine-tuning a selected CLIP (RN50×64) encoder on our dataset improves performance. Both image and text encoders were fine-tuned using prompts like "{*pose*} *photo of a* {*malnourished/healthy* child}," with the rest of the NutriScreener pipeline (GAT, retriever, fusion) unchanged. We then compared the fine-tuned and frozen variants within the full framework.

Baseline Comparisons: We benchmark *NutriScreener* against existing methods and ablations under three settings:

1. Baseline Models for Malnutrition Detection: We compare different versions of *NutriScreener* with DomainAdapt (Khan et al. 2024). Other versions include:

- **DomainAdapt** (Khan et al. 2024): Prior benchmark multitask model that combines classification and regression.
- ***NutriScreener* (BCE)**: Uses weighted binary cross-entropy loss on the GAT output to up-weight malnourished samples.
- ***NutriScreener* (Focal)**: Replaces BCE with focal loss to emphasize hard-to-classify minority-class samples.
- ***NutriScreener* (Context)**: Adds calibrated log-odds and local density estimates as auxiliary inputs to the fusion module. Trained with weighted BCE.
- ***NutriScreener* (Weighted)**: Applies temperature-scaled retrieval weighting with minority-class boosting, combined with context and weighted BCE.

2. Cohort Generalization: To assess robustness across settings, we split AnthroVision into clinical and community cohorts and report AUC and mAP for both DomainAdapt and our strongest variant from previous baselining.

3. Cross Domain Anthropometric Comparison: We compare against SOTA estimation networks from related domains such as Altinigne et al. (Altinigne, Thanou, and Achanta 2020) (IMDB, full-body) and Dantcheva et al. (Dantcheva, Bremond, and Bilinski 2018) (VIP, face).

Cross-Dataset Analysis: We evaluate the generalization ability of *NutriScreener* by varying the retrieval KB while keeping the training set fixed to AnthroVision with the largest sample size. Different KBs, such as MalKB, ARAN subsets (H1, H2), and CampusPose dataset, are used at inference to assess their impact across test cohorts.

Overall Results

CNN and VLM Baselines: Figures 3 (a) shows that standard CNN-based multitask models (ResNet, DenseNet, VGG, GoogLeNet) struggle with malnutrition classification, exhibiting high Mean Normalized Errors for anthropometric regression and polarising True Negative and True Positive rates due to class imbalance and lack of pose-aware structure. To explore stronger backbones, we evaluated vision-language model (VLM) encoders, comparing multiple CLIP variants on per-pose image classification (Figure 3 (c)). Among them, RN50×64 achieved the highest ROC AUC (68%) and mAP (58%), with balanced recall and precision, outperforming smaller variants (e.g., RN50 overpredicts with recall: 80%, precision: 32%) and larger ones (e.g., RN50×16 yields low recall: 15%). This supports RN50×64 as the most robust and balanced encoder under class imbalance.

Impact of Fine-Tuning: To test whether domain-specific adaptation improves generalization, we fine-tuned both the

Model	Acc ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	mAP ↑	H ↓	W ↓	MUAC ↓	HC ↓
DA	0.68 ± 0.01	0.63 ± 0.01	0.67 ± 0.04	0.64 ± 0.03	0.55 ± 0.02	0.35 ± 0.02	22.00 ± 1.07	12.40 ± 0.32	3.55 ± 0.24	5.05 ± 0.18
C+GNN	**0.76 ± 0.03**	**0.66 ± 0.04**	0.54 ± 0.07	0.59 ± 0.04	**0.82 ± 0.03**	**0.66 ± 0.03**	7.37 ± 0.55	5.82 ± 0.30	3.80 ± 0.28	5.23 ± 0.22
Ret-only	0.53 ± 0.04	0.36 ± 0.06	0.66 ± 0.05	0.45 ± 0.03	0.61 ± 0.07	0.38 ± 0.02	9.48 ± 0.37	7.89 ± 0.63	3.12 ± 0.47	2.76 ± 0.28
NS-W	0.74 ± 0.04	0.56 ± 0.01	**0.79 ± 0.02**	**0.66 ± 0.01**	**0.82 ± 0.01**	0.65 ± 0.02	**6.38 ± 0.46**	**5.32 ± 0.56**	**2.80 ± 1.47**	**2.97 ± 1.54**

Table 1: Performance comparison of baseline models and the proposed NutriScreener (Weighted) model on the AnthroVision dataset. DA = DomainAdapt, C+GNN = CLIP+GNN, Ret-only = Retrieval-only, NS-W = NutriScreener (Weighted). Regression metric is RMSE in "*cm*" (H, MUAC, HC) and "*kg*" (W). Values are reported as mean ± standard deviation.

Variant	Acc ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	mAP ↑	H ↓	W ↓	MUAC ↓	HC ↓
BCE	0.66 ± 0.07	0.47 ± 0.04	**0.81 ± 0.12**	0.59 ± 0.04	0.78 ± 0.04	0.60 ± 0.03	10.93 ± 0.48	8.33 ± 0.26	3.88 ± 0.22	3.15 ± 0.19
Focal	0.62 ± 0.07	0.42 ± 0.02	0.73 ± 0.07	0.53 ± 0.01	0.73 ± 0.04	0.52 ± 0.03	10.82 ± 0.45	8.46 ± 0.28	4.28 ± 0.24	3.33 ± 0.20
Context	0.73 ± 0.04	0.54 ± 0.02	0.65 ± 0.01	0.59 ± 0.01	0.78 ± 0.04	0.58 ± 0.03	11.10 ± 0.30	8.65 ± 0.30	3.63 ± 0.16	5.38 ± 0.19
NS-W	**0.74 ± 0.04**	**0.56 ± 0.01**	0.79 ± 0.02	**0.66 ± 0.01**	**0.82 ± 0.01**	**0.65 ± 0.02**	**6.38 ± 0.46**	**5.32 ± 0.56**	**2.80 ± 1.47**	**2.97 ± 1.54**

Table 2: Ablation study of NutriScreener variants. Regression metric is RMSE in "*cm*" (H, MUAC, HC) and "*kg*" (W).

	AUC		mAP	
Model	Comm	Clin	Comm	Clin
DomainAdapt	0.64	0.51	0.40	0.35
NutriScreener (Weighted)	0.78	0.74	0.53	0.58

Table 3: Comparison of community and clinical performance with baseline.

Dataset	Image Type	MAE (H)	MAE (W)
IMDB Dataset	Full-Body	6.13 cm	9.80 kg
VIP Attribute Dataset	Face	8.20 cm	8.51 kg
AnthroVision	Multi-Pose	4.69 cm	3.78 kg

Table 4: Existing work reporting mean absolute error (MAE) for Ht and Wt estimation from images.

image and text encoders of RN50×64 using prompts of the form: {*pose*} *photo of a* {*malnourished/healthy*} *child*. This encoder was plugged into the downstream NutriScreener pipeline (GNN, retrieval, fusion). The frozen pretrained encoder outperformed the fine-tuned version across all metrics, malnutrition recall (79% vs. 38%), Ht RMSE (6.38 vs. 8.87 cm), ROC-AUC (0.82 vs. 0.72), and mAP (0.65 vs. 0.54). These findings align with prior work cautioning against fine-tuning foundation models in low-resource settings due to representational collapse (Kumar et al. 2022; Gong et al. 2023). They also suggest that the pretrained CLIP features, learned from large-scale image-text pairs, encode semantically relevant anthropometric cues. Freezing the encoder preserves these representations and avoids overfitting to dataset-specific correlations, leading to better generalization under data scarcity. More analysis on these baselines is added in the *extended paper*.

Baseline Comparison with SOTA and Variant Progression: Tables 1 and 2 report performance across existing baselines and successive NutriScreener variants. 95% confidence intervals, Statistical significance analyses (Friedman and Wilcoxon tests) are provided in the extended version of the paper. The baseline, *DomainAdapt* (Khan et al. 2024) yields moderate classification performance (Recall: 0.67, F1: 0.64, AUC: 0.55) but exhibits poor regression accuracy (Ht RMSE: 22.0 cm), indicating limited generalizability under distributional shift and minority imbalance. A stronger non-retrieval baseline, *CLIP + GNN*, leveraging multi-pose CLIP (RN50x64) embeddings fused with age metadata via a GAT, improves AUC (0.82) and reduces regression error (Ht: 7.37 cm, Wt: 5.82 kg), although recall remains low (0.54), highlighting insufficient sensitivity to the minority class. A *Retrieval-only* setup, using similarity-weighted averaging of k nearest labelled neighbours, shows retrieval alone is insufficient, with unstable precision–recall balance and inconsistent regression performance.

Building on these observations, the introduction of retrieval-augmented training improves performance progressively across NutriScreener variants. *NutriScreener (BCE)* fuses FAISS-based neighbors with class-weighted BCE loss, increasing recall (0.81) but reducing precision (0.47) and worsening regression (Ht RMSE: 10.93 cm), suggesting sensitivity to noisy retrieved support. Replacing BCE with focal loss in *NutriScreener (Focal)* maintains high recall (0.73) but lowers F1 (0.53) and increases MUAC error (4.28 cm), likely due to over-penalization of easy samples. *NutriScreener (Context)* incorporates calibrated log-odds and local neighbor density as auxiliary fusion signals, yielding more balanced precision (0.54), stable AUC (0.78), and reduced error variance. Finally, *NutriScreener (Weighted)* applies temperature-scaled, class-aware weighting over retrieved neighbors and achieves the best overall performance: Recall (0.79), F1 (0.66), AUC (0.82), and lowest RMSEs across all anthropometric indicators (Ht: 6.38 cm, Wt: 5.32 kg, MUAC: 2.80 cm, HC: 2.97 cm), demonstrating that principled retrieval integration substantially enhances robustness and minority-class performance.

Cohort Generalization: Table 3 highlights the generalization strength of NutriScreener (Weighted) across both community and clinical settings, achieving AUC scores of 0.78 and 0.74, respectively, significantly outperforming DomainAdapt. The consistent mAP improvement confirms its robustness to population and context shifts, a critical re-

Test	Labels	KB	Rec ↑	F1 ↑	AUC ↑	mAP ↑	H ↓	W ↓	MUAC ↓	HC ↓
AV	Classification + Regression	NoRet	0.54±0.15	0.59±0.04	0.82±0.03	0.66±0.05	7.37±0.55	5.82±0.30	3.80±1.40	5.23±1.60
		H2	0.54±0.15	0.59±0.04	0.82±0.03	0.66±0.05	7.37±0.55	5.82±0.30	3.80±1.40	5.23±1.60
		CP	0.66±0.18	0.58±0.03	0.80±0.02	0.62±0.04	7.37±0.55	5.82±0.30	3.80±1.40	5.23±1.60
		MalKB+H2	0.73±0.14	0.60±0.04	0.80±0.02	0.62±0.03	6.53±0.50	5.96±0.26	3.32±1.44	2.93±1.56
		MalKB	0.79±0.02	0.66±0.01	0.82±0.01	0.65±0.02	6.38±0.46	5.32±0.56	2.80±1.47	2.97±1.54
H1	Regression (H, W, HC)	NoRet	-	-	-	-	7.87±0.30	5.91±0.47	-	4.59±1.98
		MalKB	-	-	-	-	7.87±0.30	5.91±0.47	-	4.59±1.98
		CP	-	-	-	-	7.87±0.30	5.91±0.47	-	4.59±1.98
		H2	-	-	-	-	7.21±0.02	5.67±0.38	-	4.20±0.18
CP	Classification + Regression	NoRet	0.75±0.01	0.63±0.04	0.64±0.04	0.69±0.02	23.98±0.27	17.36±0.16	9.46±0.28	3.98±0.31
		MalKB	0.75±0.01	0.63±0.04	0.64±0.04	0.69±0.02	23.98±0.27	17.36±0.16	9.46±0.28	3.98±0.31
		H2	0.72±0.04	0.65±0.01	0.65±0.02	0.69±0.02	23.98±0.27	17.36±0.16	9.46±0.28	3.98±0.31
		MalKB+H2	0.75±0.03	0.63±0.05	0.64±0.02	0.69±0.03	23.98±0.27	17.36±0.16	9.46±0.28	3.98±0.31

Table 5: Impact of KB dataset choice on classification and regression performance when the model is trained solely on the AnthroVision dataset. Each test set is evaluated using different KBs. "-" indicates unavailability of ground truth. **Note:** Highly out-of-distribution KBs (CampusPose cohort) result in identical performance to No Retrieval pertaining to the fusion mechanism. CP: CampusPose. Regression metric is RMSE in "*cm*" (H, MUAC, HC) and "*kg*" (W).

quirement for real-world malnutrition screening.

Cross Domain Comparison: Table 4 presents results across different domains for the shared task of image-based Ht and Wt estimation. NutriScreener outperforms prior methods from adult subject and controlled imaging domains, demonstrating superior generalization and robustness in pediatric, real-world settings.

Cross-Dataset Analysis: Table 5 shows that the choice of retrieval KB significantly impacts classification and regression outcomes. *MalKB* refers to this in-house curated knowledge base used for retrieval-augmented inference. *NoRet* denotes the ablation setting where retrieval is disabled and inference relies solely on model predictions without any KB augmentation. On the AV test set, using NoRet, CampusPose, or H2 KBs results in lower recall (0.54–0.66) and moderate F1 (0.58–0.59), indicating poor malnutrition sensitivity. In contrast, MalKB achieves superior recall (0.79), F1 (0.66), AUC (0.82), and the lowest RMSEs across anthropometric measures (Ht: 6.38 cm, Wt: 5.32 kg, MUAC: 2.80 cm, HC: 2.97 cm). Even partial KB augmentation (MalKB+H2) improves recall (0.73), highlighting generalization from minor domain expansion. In the ARAN-H1 setting (regression-only), most KBs perform similarly (Ht RMSE: 7.87 cm), but H2 achieves better RMSEs (Ht: 7.21 cm, HC: 4.20 cm), indicating value in intra-cohort matching. For the CampusPose cohort, all KBs yield nearly identical results (recall: 0.72–0.75, F1: 0.60–0.63, AUC: $\sim$0.64, Ht RMSE: 23.98 cm), reflecting domain saturation and the limited utility of retrieval under large population shifts.

User Study: Clinician Feedback

To evaluate *NutriScreener*'s clinical utility, we conducted a user study with 12 medical professionals (mean experience: 9.5 years), including paediatricians and general practitioners. After a brief demo, clinicians received a standalone version of the toolkit, containing the trained model and knowledge base as embedded, non-reversible structures, and applied it to an average of 15 pediatric cases per doctor during regular clinical workflow. Feedback via 5 point Likert-scale indicated strong acceptance: clinical consistency (4.3/5), efficiency (4.6/5), trustworthiness (4.4/5), and deployment readiness (4.1/5). Participants especially valued the tool as an objective "second opinion," with one noting it successfully flagged a visually ambiguous malnourishment case. Open-ended responses suggested key improvements, including explicit uncertainty estimates and visual cues to highlight key visual areas. Clinicians found *NutriScreener* reliable, efficient, and suitable for real-world use, especially by community health workers in low-resource settings.

Conclusion and Future Work

Evaluated on cross-continent datasets (AnthroVision, ARAN, and CampusPose), *NutriScreener* establishes a new benchmark for child malnutrition screening, addressing both algorithmic and deployment-level challenges in low-resource contexts. By integrating class-boosted context-aware retrieval augmentation with multi-pose graph attention over CLIP features, *NutriScreener* achieves strong sensitivity (Recall: 0.79), generalization (AUC: 0.82), and low anthropometric prediction errors, outperforming CNN and domain-adaptive baselines. Cross-dataset results show up to 25% recall gain and up to 2.3 cm reduction in head circumference RMSE using demographically matched knowledge bases, highlighting the framework's adaptability to new populations. Clinician validation (trust: 4.4/5) confirms the system's accuracy and deployment readiness. Beyond technical gains, *NutriScreener* enables scalable, low-cost screening from routine images, reducing manual effort, and as an assistive tool, it supports early detection of at-risk children. It empowers frontline workers and facilitates timely care. Future work will expand knowledge diversity and add interpretability and reduce uncertainty.

Acknowledgments

M. Khan is partly supported through the Prime Minister's Research Fellowship, India. This research is also partially supported through Srijan: Center for Excellence in GenAI. The authors thank the clinicians and volunteers for their participation in data collection and field validation.

Ethical Statement

This study was conducted under institutional ethics approvals from IIT and AIIMS Jodhpur, India, and informed consent was obtained from all participants' parents or guardians with the option of data deletion upon request. The NutriScreener Toolkit follows privacy-by-design principles. It operates on non-reversible CLIP embeddings that cannot reconstruct original images, retains no personally identifiable information beyond age and anthropometrics, and stores all data on encrypted, access-controlled infrastructure. To prevent misuse, including unauthorized use for body-shaming, aesthetic assessment, or non-clinical surveillance, NutriScreener will be distributed under a restricted research license permitting only educational and research applications. Cross-population validation shows consistent performance with demographically aligned knowledge bases, supporting equitable and generalizable deployment.

References

Aanjankumar, S.; Sathyamoorthy, M.; Dhanaraj, R. K.; Surjit Kumar, S.; Poonkuntran, S.; Khadidos, A. O.; and Selvarajan, S. 2025. Prediction of malnutrition in kids by integrating ResNet-50-based deep learning technique using facial images. *Scientific Reports*, 15(1): 7871.

Altinigne, C. Y.; Thanou, D.; and Achanta, R. 2020. Height and weight estimation from unconstrained images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2298–2302. IEEE.

Dantcheva, A.; Bremond, F.; and Bilinski, P. 2018. Show me your face and I will tell you your height, weight and body mass index. In *Proceedings of the 24th International Conference on Pattern Recognition*, 3555–3560. IEEE.

Das, A. M.; Bhatt, G.; Kumari, L.; Verma, S.; and Bilmes, J. 2025. COBRA: COmBinatorial Retrieval Augmentation for Few-Shot Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazar'e, P.-E.; Lomeli, M.; Hosseini, L.; and J'egou, H. 2024. The Faiss library. *ArXiv*, abs/2401.08281.

Gong, C.; Chen, Z.; Wang, X.; Wang, W.; Zhang, Z.; and Yu, Z. 2023. ARF: Anchor-based Robust Finetuning for Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Janssen, S. M.; Bouzembrak, Y.; and Tekinerdogan, B. 2024. Artificial Intelligence in Malnutrition: A systematic literature review. *Advances in Nutrition*, 15(9): 100264.

Khan, M.; Agarwal, S.; Vatsa, M.; Singh, R.; and Singh, K. 2023. NutriAI: AI-powered child malnutrition assessment in low-resource environments. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6378–6385.

Khan, M.; Khurshid, M.; Vatsa, M.; Singh, R.; Duggal, M.; and Singh, K. 2022. On AI approaches for promoting maternal and neonatal health in low resource settings: a review. *Frontiers in Public Health*, 10: 880034.

Khan, M.; Singh, R.; Vatsa, M.; and Singh, K. 2024. DomainAdapt: Leveraging Multitask Learning and Domain Insights for Children's Nutritional Status Assessment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 606–616. Springer.

Khandelwal, Y.; Arvind, M.; Kumar, S.; Gupta, A.; Danisetty, S. K.; Bagad, P.; Madan, A.; Lunayach, M.; Annavajjala, A.; Maiti, A.; et al. 2024. Nurturenet: a multi-task video-based approach for newborn anthropometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 332–342.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4501–4510.

Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; and Liang, P. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *Proceedings of the International Conference on Learning Representations*.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. Dynamic Multiscale Graph Neural Networks for 3D Skeleton-Based Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 214–223.

Li, Y.; Guo, J.; Qi, L.; Li, W.; and Shi, Y. 2025. Text and Image Are Mutually Beneficial: Enhancing Training-Free Few-Shot Classification with CLIP. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5039–5047.

Liu, T.; Zhang, H.; Parashar, S.; and Kong, S. 2025. Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Y.; Sowmya, A.; and Khamis, H. 2018. Single Camera Multi-View Anthropometric Measurement of Human Height and Mid-Upper Arm Circumference Using Linear Regression. *PLOS ONE*, 13(4): e0195600.

Long, A.; Yin, W.; Ajanthan, T.; Nguyen, V.; Purkait, P.; Garg, R.; Blair, A.; Shen, C.; and van den Hengel, A. 2022. Retrieval Augmented Classification for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6959–6968.

Mohammed Khan, H. H.; van Wanrooij, C.; Postma, E. O.; Güven, C.; Balvert, M.; Raof Saeed, H.; and Ali Al Jaf, C. O. 2025. ARAN: Age-Restricted Anonymized Dataset of Children Images and Body Measurements. *Journal of Imaging*, 11(5): 142.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Tay, W.; Quek, R.; Kaur, B.; Lim, J.; Henry, C. J.; et al. 2022. Use of facial morphology to determine nutritional status in older adults: opportunities and challenges. *JMIR Public Health and Surveillance*, 8(7): e33478.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Wang, J.; He, C.; and Long, Z. 2023. Establishing a machine learning model for predicting nutritional risk through facial feature recognition. *Frontiers in Nutrition*, 10: 1219193.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3876–3889.

Welthungerhilfe. 2025. ChildGrowthMonitor. https://github.com/Welthungerhilfe/ChildGrowthMonitor. AI-powered mobile and web solution for monitoring child growth and malnutrition Accessed: Nov 11, 2025.

World Health Organization. 2006. WHO Child Growth Standards. Available at https://www.who.int/tools/child-growth-standards.Accessed:Nov11,2025.

World Health Organization. 2024. Malnutrition. https://www.who.int/news-room/fact-sheets/detail/malnutrition. Accessed: Nov 11, 2025.

Yu, X.; Wu, Z.; Zhang, L.; Zhang, J.; Lyu, Y.; and Zhu, D. 2024. Cp-clip: Core-periphery feature alignment clip for zero-shot medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 88–97. Springer.