International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 620-645
https://doi.org/10.70917/ijcisim-2026-5334
Article

Multimodal Deep Representation Learning for Nutritional Status Prediction in Children Under Five

Mrs.C.Subalakshmi[1], Dr S.Geetha[2], Dr T. Kumanan[3]

[1]Research Scholar, Department of Computer Science and Engineering, Dr.MGR Educational and Research Institute, Chennai
[2,3]Professor, Department of Computer Science and Engineering, Dr.MGR Educational and Research Institute, Chennai
[1]subalakshmi.cse@drmgrdu.ac.in
[2]geethasoman@drmgrdu.ac.in
[3]kumanan.cse@drmgrdu.ac.in

Abstract: Nutritional risk among children under five years should be identified early to minimize the developmental, health and cognitive impairments in the long run. The assessment of the nutritional status is however usually confronted by disaggregated and diverse sources of data such as anthropometric data, dietary behavior records, social economic data and clinical health history. The proposed study is a multimodal deep representation learning model, which will learn structured and unstructured child health data to predict the nutrition status correctly. The model encodes structured clinical and anthropometric features into latent representation with a Variational Autoencoder and encodes time and contextual patterns of diet and feeding behavior with a Transformer encoder. The cross-modal fusion mechanism takes these latent embeddings and unites them into a single representation of child health with which nutritional risk can be classified. The suggested approach makes it possible to do solid data-driven nutritional profiling and facilitate detecting children, who are at moderate or severe nutritional risk, in their early-stage. Experimental analysis indicates a better classification accuracy and stability compared to base machine learning and single-stream deep models. This framework provides a computational framework scaling method on enhancing the pediatric nutritional surveillance and targeted intervention planning.

Keywords: nutritional risk prediction, multimodal deep learning, variational autoencoder, transformer encoder, child health assessment, under-five nutrition

1. Introduction

In early childhood, nutritional status is of critical importance to the long-term physical growth, cognitive growth, and the general health outcome. Nutritional imbalances especially among children under the age of five are caused by fast change in physiology, feeding habits and socioeconomic factors in the environment [1]. Traditional nutritional assessment systems are primarily based on anthropometric measurements like height-age, weight-age and mid-upper arm circumference [2]. These indicators, however, are not enough to comprehend the multidimensional determinants of nutrition, such as the dietary variety, feeding schedule, family conditions, maternal determinants, and the history of morbidity [3]. Consequently, concern has increased regarding the need to apply computational methods that can combine multiple sources of health data to give more accurate and extensive nutritional risk prediction [4].

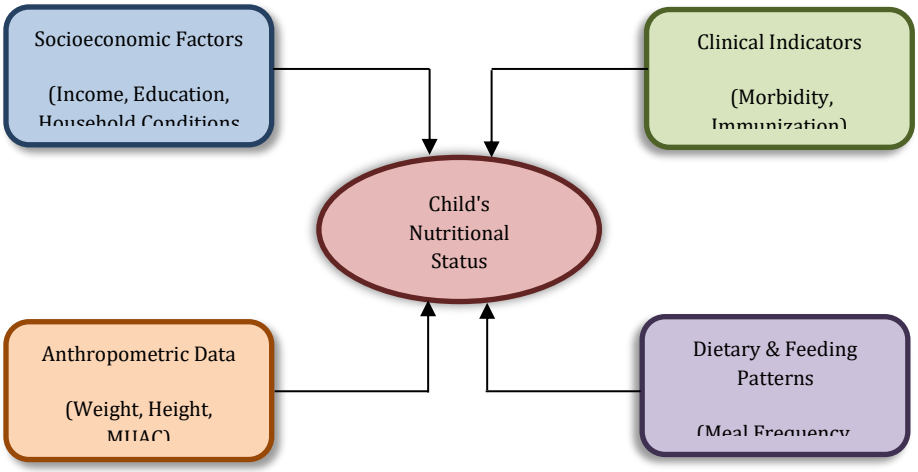

Figure 1. Determinants of Nutritional Status in Children Under Five

The level of influence of socioeconomic, clinical, anthropometric and dietary factors has been interrelated to demonstrate the importance of nutritional status [5] among children below the age of five years (Fig. 1). The recent developments and innovations in the field of deep learning [6] have made feature extraction and pattern recognition in health analytics with high capacities possible [7]. However, the available methods have mainly been focusing on child nutrition evaluation [8] as a one-source or one-modal issue, which does not enable them to reflect the complex relationships that affect nutritional outcomes [9]. Learning of multimodal representation offers an avenue to adopt heterogeneous health information together, but the difficulty is to come up with models that utilize structured and unstructured streams of data without reducing interpretation or predictive strength [10].

The paper is a response to this gap because it presents a multimodal deep representation learning framework that combines structured anthropometric and clinical information with sequence-based dietary and behavioral history. The architecture uses a Variational Autoencoder to learn a small latent code representing structured characteristics and a Transformer encoder to learn temporal dynamics and contextual significance over unstructured feeding and lifestyle data. A cross-modal fusion process integrates and correlates these representations to form a integrated child health picture that can be used to classify nutritional risk.

1. Construction of a multimodal deep representation learning model specifically to predict nutritional status in children under five.

2. Structured health indicator and unstructured diet behavior sequence integration with a joint Variational Autoencoder and Transformer encoder.

3. Introduction of cross-modal fusion mechanism resulting in a single child nutritional risk embedding.

4. Evidence of better predictive power than traditional and one-stream deep learning baselines.

Table 1. Limitations in Existing Nutritional Assessment Approaches

Method / Practice	**Data Used**	**Limitation Identified**	**Impact on Assessment**
Anthropometric Chart-based Classification	Height, weight, MUAC	Fails to reflect diet quality and socioeconomic conditions	May misclassify borderline nutritional risk
Dietary Survey & Food Recall	Caregiver-reported diet intake	Subject to recall bias and inconsistent reporting	Results in unreliable risk estimation

Socioeconomic Background Scoring	Income, living conditions	Does not quantify nutritional status directly	Cannot predict risk progression
Single-Source Machine Learning Models	One data modality	Cannot capture cross-factor relationships	Reduced prediction accuracy and generalization

Table 1 gives the summary of key limitations in current child nutrition assessment methods, motivating the need for multimodal representation learning.

2. Related Works

A number of studies have been dedicated to the usage of artificial intelligence and machine learning to measure and predict the nutritional status of children in various regions and settings. A longitudinal-based LSTM-FC neural network model was suggested to predict and categorize nutritional changes in Ethiopian children, which has a high predictive power and accuracy in long-term prediction [11]. A similar study proposed a multi-head attention deep learning model that combines CNN and LSTM models to predict nutritional status with a focus on explainable AI (SHAP) to improve interpretability and clinical trustworthiness [12]. The malnutrition prediction in children was done by a deep learning image segmentation model, which used ResNet-50 to eliminate various manual diagnosis processes and minimize the use of ongoing medical attention in the rural setting [13]. The other work critically evaluated different machine learning methods to predict childhood and adolescent obesity and offered a taxonomy of existing approaches, as well as showed what research gaps might exist regarding gender specificity, data diversity and interpretability [14]. In one longitudinal study, malnutrition changes in children were statistically modeled in Iran and the results were found to have improved over the decades although they still see that food-insecure areas continue to face significant problems [15].

An innovative child monitoring system on nutrition was created to be used at home and in the clinic, and it could help parents assess the nutritional status based on simple anthropometric inputs and predictive modeling to detect the initial evidence of non-normal growth patterns [16]. Likewise, comparison of machine learning algorithms such as Random Forest, SVM, and Decision Trees was done to determine the nutritional status of toddlers in Indonesia, which is effective predictors of early intervention and regional monitoring [17]. In clinical practice, the Pediatric Renal Nutrition Taskforce has given standard practice guidelines on how the nutritional health of children with kidney diseases can be assessed using anthropometric and biochemical measures [18].

In another study, multiple linear regression was used to predict the status of child nutrition across the regions, which helped in planning locally and making decisions on time in community health programs [19]. Predictive modeling has also been applied to the outcomes of nutritional status in women, where controlled machine learning algorithms were applied to predict BMI and categorize malnutrition risks in Bangladeshi women, with socioeconomic and health-related factors found to be the most important predictors [20]. The Ethiopian national survey data was also compared with multiple machine learning models to determine risk factors of childhood stunting to pursue the policy interventions to enhance water, food security, and healthcare access [21].

A study examined reviewed assessment tools in the assessment of nutritional status in children with cancer, and it is necessary to use standardized measures that include anthropometric, clinical and dietary measurements [22]. Machine learning classifiers like the Random Forest, k-NN and SVM have been tested to predict under-five malnutrition in Bangladesh whereby feature selection enhanced the classification accuracy and interpretability [23]. In a study on nutritional assessment in Pakistan, maternal and child variables were examined in relation to stunting and wasting and significant correlation was established between the quality of the diet and maternal literacy and maternal growth [24]. Finally, a multinomial logistic regression model was designed to determine the nutritional status of toddlers based on anthropometric measurements, which proves that statistical learning methods can be useful in health surveillance in early stages [25].

All these works lead to the conclusion of the transformation of classic statistical methods to deep-learning-based, explainable, and multi-modal nutritional assessment frameworks. The literature suggests that the trend is towards incorporating structured, image-based, and longitudinal data in better generalization and practical implications in monitoring of public health and nutrition.

3. Proposed Model

The model suggested uses a multimodal deep representation learning framework to predict nutritional status in children below five years of age through a combination of both structured and unstructured health-related data. Anthropometric measures, clinical and socioeconomic attributes are coded into a small latent representation via a variational autoencoder that assists in reflecting the underlying health patterns, whilst being robust to missing and noisy data. Meanwhile, the sequences of unstructured dietary and feeding behavior are analyzed with the help of a transformer encoder that is an efficient way of capturing temporal dependencies and contextual differences in nutritional intake and morbidity history of the child. Latent representations based on each of the data streams are subsequently combined via a cross-modal fusion process that harmonizes and balances the input of each modality to give a single, discriminative health embedding. This combined representation is then fed to a softmax classification layer which then classifies the children into either normal, moderate-risk or severe-risk nutritional status. The selected model design enables the full exploitation of heterogeneous child health data and is more accurate in prediction than single-source or conventional statistical assessment models. Fig 2 illustrates the suggested multimodal structure that combines a VAE to organize health data and a Transformer encoder to handle temporal diet sequences.

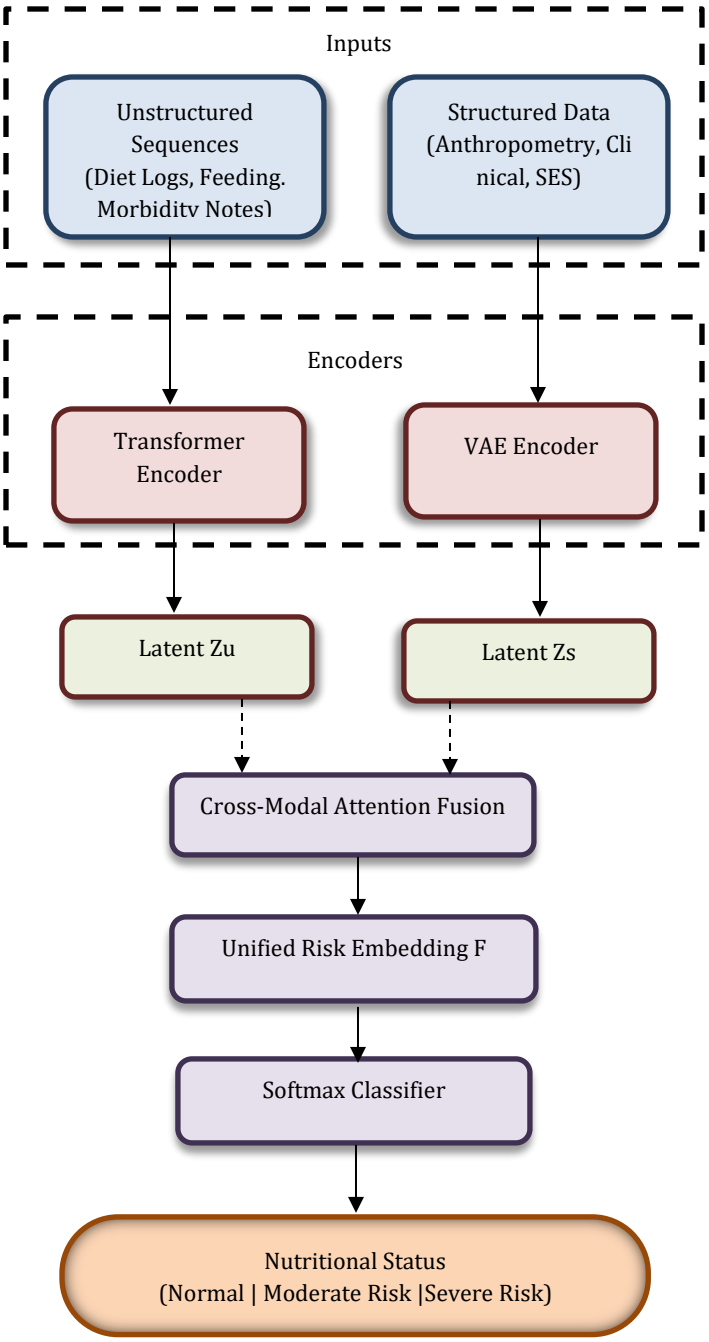

Figure 2: Proposed Multimodal Architecture

3.1. Problem definition

Let $\mathbb{D} = \{(x_i^s, x_i^u, y_i)\}_{\{i=1\}}^{N}$ be independent child-level samples. The structured attributes $x_i^s \in \mathbb{R}\hat{}\{d^s\}$ include anthropometry, clinical, and socioeconomic features. The unstructured temporal stream $x_i^u = (u_i 1, \ldots, u_i T_i)$ with $u_i t \in \mathbb{R}^{\{d^u\}}$ represents diet, feeding, and morbidity events. The label $y_i \in \mathcal{Y} = \{1,2,3\}$ corresponds to nutritional status {normal, moderate, severe}.

The goal is to learn a classifier $f\theta : (R^{\{d^s\}} \times U) \rightarrow \Delta^3$ that minimizes expected misclassification under the joint distribution $P(x^s, x^u, y)$. The optimal Bayes rule predicts

$$f^\star(x^s, x^u) = \arg \max_{c \in \{1,2,3\}} p(y = c \mid x^s, x^u),$$

but $p(y|x^s, x^u)$ is unknown and must be approximated.

We adopt empirical risk minimization on a hypothesis class $\mathcal{F} = \{f\theta\}$ with softmax outputs. With class-imbalance-aware weights $w_c > 0$, the empirical risk is

$$\hat{\mathcal{R}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i^s, x_i^u), y_i), \mathcal{L}(\hat{\mathbf{p}}, y) = -\sum_{c=1}^{3} w_c \, \mathbb{1}[y = c] \, \log \hat{p}_c.$$

To exploit modality-specific structure, we posit latent variables $z^s \in \mathbb{R}^{\{k^s\}}$ for structured data and $z^u \in \mathbb{R}^{\{k^u\}}$ for sequences, and a discriminative head gψ that maps a fused representation $F(z^s, z^u)\ to\ \Delta^3$. The predictive model is

$$\hat{\mathbf{p}}_\theta(y \mid x^s, x^u) = g_\psi\,(F\,(z_\phi^s(x^s), z_\varphi^u(x^u))),$$

where z_ϕ^s and $z^u\phi_i$ are modality encoders with parameters φ and ϕ, respectively, and $\theta = (\varphi, \phi, \psi)$.

Structured encoder as a variational latent model. We introduce a generative model $p_\theta(x^s, z^s) = p(z^s)p_\theta\,(x^s z^s)$ with prior $p(z^s) = \mathcal{N}(0, I)$. An amortized approximate posterior $(q_\phi(z^s \mid x^s) = \mathcal{N}\left(\mu_\varphi(x^s), diag\left(\sigma_\varphi^2(x^s)\right)\right)$ yields the evidence lower bound (ELBO) per sample

$$\log p_\theta(x^s) \geq \mathbb{E}_{q_\phi(z^s|x^s)}\,[\log p_\theta(x^s \mid z^s)] - \mathrm{KL}\,(q_\phi(z^s \mid x^s) \parallel p(z^s)) = \mathcal{L}_{\mathrm{ELBO}}^{(s)}(x^s; \theta, \phi).$$

We obtain a stochastic z^s by the reparameterization trick

$$z^s = \mu_\phi(x^s) + \sigma_\phi(x^s) \odot \epsilon, \epsilon \sim \mathcal{N}(0, I).$$

Sequence encoder as a contextual latent. For the sequence x^u we compute contextual states H = $Encoder_\phi(x^u)$ (e.g., Transformer), then aggregate to a fixed-length latent via attention pooling

$$\alpha_t = \frac{\exp(v^\top \tanh(Wh_t))}{\sum_j \exp(v^\top \tanh(Wh_j))}, z^u = \sum_t \alpha_t h_t\,.$$

This defines a deterministic $q\_\phi(z^u\ x^u) = delta(z^u - \ z^u var\phi(x^u))$ used for discrimination.

Fusion and decision function. Let F be a gated linear fusion that balances modalities,

$$h = [z^s; z^u], g = \sigma(W_g h + b_g), F(z^s, z^u) = g \odot z^s \ + \ (1 - g) \odot z^u.$$

The class posterior estimate is produced by a softmax head

$$\hat{\mathbf{p}}_\theta(y \mid x^s, x^u) = \mathrm{softmax}(W_o\, \phi_o(F) + b_o).$$

Training objective. We jointly maximize the structured ELBO and minimize the weighted classification loss, with λ terms balancing contributions:

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N}\Big(\underbrace{\mathcal{L}(\hat{\mathbf{p}}_\theta(x_i^s, x_i^u), y_i)}_{\text{discriminative}} - \lambda_{\text{ELBO}} \underbrace{\mathcal{L}_{\text{ELBO}}^{(s)}(x_i^s; \theta, \phi)}_{\text{structured reconstruction + KL}}\Big) + \lambda_{\text{reg}} \| \theta \|_2^2.$$

Equivalently, writing the ELBO components explicitly,

$$\mathcal{J}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(-\sum_{c=1}^{3} w_c\, \mathbb{1}[y_i = c]\log \hat{p}_{i,c} + \lambda_{\text{rec}} \| x_i^s - \hat{x}_i^s \|_1 + \lambda_{\text{KL}}\mathrm{KL}\left(q_\phi(z_i^s \mid x_i^s) \parallel \mathcal{N}(0, I)\right)\Big) + \lambda_{\text{reg}} \| \theta \|_2^2.$$

Decision rule and calibration. At test time, the predicted label is

$$\hat{y} = \arg\max_c \hat{p}_c.$$

If empirical priors π_c differ from deployment priors π'_c, posterior calibration can adjust logits ℓ_c via

$$\tilde{\ell}_c = \ell_c + \log\frac{\pi_c'}{\pi_c}, \tilde{p}_c = \frac{e^{\tilde{\ell}_c}}{\sum_j e^{\tilde{\ell}_j}}.$$

Assumptions. Samples are i.i.d.; missingness in x^s is at most missing at random and partially addressed by the VAE encoder and explicit missingness indicators; sequence padding uses a causal mask so that z^u is invariant to padding; class weights w_c mitigate imbalance.

This formulation specifies the random variables, hypothesis class, latent-variable structure, and the exact optimization objective that approximates the Bayes-optimal predictor under multimodal data.

3.2. Data Modalities and Notation

We consider a dataset

$$\mathcal{D} = \{(x_i^s, x_i^u, y_i)\}_{i=1}^N$$

where each sample corresponds to one child under five years of age.

3.2.1 Structured Modality $\mathbf{x^s}$

The structured feature vector consists of anthropometric, clinical, and socioeconomic attributes:

$$x_i^s = [\, a_i, c_i, s_i \,] \in \mathbb{R}^{d_s}$$

where:

$$a_i \in \mathbb{R}^{d_a} (\text{anthropometry: weight, height, MUAC})$$
$$c_i \in \mathbb{R}^{d_c} (\text{clinical markers: hemoglobin, morbidity})$$
$$s_i \in \mathbb{R}^{d_{ses}} (\text{socioeconomic indicators})$$

and the total dimension:

$$d_s = d_a + d_c + d_{ses}$$

Normalization

Each continuous feature x_{ij}^s is standardized:

$$\tilde{x}_{ij}^s = \frac{x_{ij}^s - \mu_j}{\sigma_j}$$

where μ_j and σ_j are mean and standard deviation computed from the training set.

WHO Z-score Transformation (Optional)

Using WHO LMS parameters (L, M, S), anthropometric z-scores (e.g., HAZ/WAZ/WHZ) are computed as:

$$z = \begin{cases} \frac{(\frac{x}{M})^L - 1}{LS}, & L \neq 0 \\ \frac{\ln(\frac{x}{M})}{S}, & L = 0 \end{cases}$$

3.2.2 Unstructured Temporal Modality $\mathbf{x^u}$

Each child has a time-ordered feeding and health sequence:

$$x_i^u = (u_{i1}, u_{i2}, \ldots, u_{iT_i})$$

where each time-step representation is:

$$u_{it} \in \mathbb{R}^p$$

Token Embeddings

Let:

- d_{it}= diet item token
- m_{it}= morbidity token

The embedding matrices:

$$E_d \in \mathbb{R}^{|V_d| \times p_d}, E_m \in \mathbb{R}^{|V_m| \times p_m}$$

produce:

$$e_{it}^{(diet)} = E_d[d_{it}], e_{it}^{(morb)} = E_m[m_{it}]$$

Positional Encoding

$$p_t = \sin(\frac{t}{10000^{2k/p}}) \oplus \cos(\frac{t}{10000^{2k/p}})$$

Final Per-Time Step Representation

$$u_{it} = [e_{it}^{(diet)}, e_{it}^{(morb)}, r_{it}] + p_t \in \mathbb{R}^p$$

where r_{it}denotes any continuous feeding quantity/duration attributes.

3.2.3 Labels y

Nutritional status is categorized according to WHO thresholds:

$$y_i \in \{1,2,3\}$$

where:

$$y_i = \{\begin{matrix} 1 & \text{if } z \geq -2(\text{Normal}) \\ 2 & \text{if } -3 \leq z < -2(\text{Moderate}) \\ 3 & \text{if } z < -3(\text{Severe}) \end{matrix}$$

and $z$is any selected anthropometric z-score (e.g., HAZ, WHZ, MUACZ).

One-hot encoding representation:

$$y_i = \text{one\_hot}(y_i) \in \mathbb{R}^3$$

3.2.4 Mini-Batch Tensor Form

For a batch of size B:

$$X^s \in \mathbb{R}^{B \times d_s}, U \in \mathbb{R}^{B \times T_{\max} \times p}, Y \in \mathbb{R}^{B \times 3}$$

Mask matrix for valid sequence positions:

$$M \in \{0,1\}^{B \times T_{\max}}$$

3.3. Preprocessing

Let $\mathcal{D} = \{(x_i^s, x_i^u, y_i)\}_{i=1}^N$denote the collected dataset. Data preprocessing is performed to ensure consistency, reduce noise, support multimodal encoding, and minimize information leakage across samples.

3.3.1 Numeric Validation and Outlier Control

All continuous fields in the structured modality x_i^s are inspected for invalid values (e.g., negative heights, implausible weights). Let x_{ij}^s denote the value of the j-th structured feature for child i. To reduce the influence of extreme measurement values, **winsorization** is applied:

$$x_{ij}^s \leftarrow \begin{cases} Q_{1\%}(x_j^s), & \text{if } x_{ij}^s < Q_{1\%}(x_j^s), \\ Q_{99\%}(x_j^s), & \text{if } x_{ij}^s > Q_{99\%}(x_j^s), \\ x_{ij}^s, & \text{otherwise,} \end{cases}$$

where $Q_{p\%}(\cdot)$ denotes the empirical percentile.

3.2 Missing Value Imputation with Missingness Indicators

For continuous structured features, missing values are imputed using site-specific medians:

$$\hat{x}_{ij}^s = \begin{cases} \text{median}(x_j^s \mid \text{site}), & \text{if } x_{ij}^s \text{ is missing,} \\ x_{ij}^s, & \text{otherwise.} \end{cases}$$

For categorical SES attributes, the imputed value corresponds to the most frequent category (mode):

$$\hat{x}_{ij}^s = \arg\max_k \text{freq}(x_j^s = k).$$

To retain missingness information, a binary indicator is appended:

$$m_{ij} = \begin{cases} 1, & x_{ij}^s \text{ is missing,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the final structured representation becomes:

$$\tilde{x}_i^s = [\, \hat{x}_i^s, m_i \,] \in \mathbb{R}^{2d_s}.$$

3.3 Standardization and Encoding

Each continuous feature is normalized as:

$$\tilde{x}_{ij}^s = \frac{\hat{x}_{ij}^s - \mu_j}{\sigma_j},$$

where μ_j and σ_j are estimated from the training partition.

Categorical SES variables are one-hot encoded:

$$s_i \rightarrow \text{onehot}(s_i) \in \{0,1\}^K.$$

3.4 Preprocessing Unstructured Temporal Sequences

The unstructured modality x_i^u consists of diet, feeding, and morbidity events indexed in time. Raw text/event tokens are mapped to embeddings via a learnable lookup matrix:

$$E \in \mathbb{R}^{|\mathcal{V}| \times p}.$$

If $d_{it} \in \mathcal{V}$is the token at time t, its embedding is:

$$u_{it} = E[d_{it}] \in \mathbb{R}^p.$$

To maintain uniform sequence length:

- If $T_i < T_{\max}$, pad with zero-vector tokens.
- If $T_i > T_{\max}$, truncate the earliest data.

A mask is generated to prevent attention to padded elements:

$$M_{it} = \begin{cases} 1, & t \le T_i, \\ 0, & t > T_i. \end{cases}$$

Thus, the final batched sequence representation is:

$$U \in \mathbb{R}^{B \times T_{\max} \times p}, M \in \{0,1\}^{B \times T_{\max}}.$$

3.3.5 Data Partitioning to Avoid Leakage

To preserve real-world deployment validity and prevent correlated household/area effects from inflating model performance, dataset splitting is performed **geographically or by cluster** (rather than random record-level splitting):

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}, \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \emptyset.$$

This preprocessing pipeline ensures consistency across structured and unstructured modalities and prepares the data for multimodal representation learning in the proposed model.

3.4. Model Components

The proposed model jointly encodes structured and unstructured data into modality-specific latent spaces and then fuses these representations for nutritional status prediction.

3.4.1 Variational Autoencoder (VAE) for Structured Data

Let the structured input be $x_i^s \in \mathbb{R}^{d_s}$. We introduce a latent variable $z_i^s \in \mathbb{R}^{k_s}$governed by a Gaussian prior:

$$p(z_i^s) = \mathcal{N}(0, I).$$

A variational encoder parameterized by ϕ approximates the posterior:

$$q_\phi(z_i^s \mid x_i^s) = \mathcal{N}(\mu_\phi(x_i^s), \text{diag}(\sigma_\phi^2(x_i^s))).$$

The encoder computes:

$$\mu_i^s, \log \sigma_i^{s\,2} = \text{MLP}_\phi(x_i^s),$$

and sampling uses the reparameterization trick:

$$z_i^s = \mu_i^s + \sigma_i^s \odot \epsilon, \epsilon \sim \mathcal{N}(0, I).$$

The decoder reconstructs x_i^s:

$$\hat{x}_i^s = \mathrm{Dec}_\theta(z_i^s),$$

and the VAE loss (ELBO) is:

$$\mathcal{L}_{\mathrm{VAE}} = \underbrace{\| x_i^s - \hat{x}_i^s \|_1}_{\text{Reconstruction Loss}} + \underbrace{\mathrm{KL}(q_\phi(z_i^s \mid x_i^s) \parallel \mathcal{N}(0, I))}_{\text{Latent Regularization}}.$$

Typical latent dimension: $k_s \in [16,64]$.

3.4.2 Transformer Encoder for Unstructured Sequence Data

Let the temporal modality be:

$$x_i^u = (u_{i1}, u_{i2}, \ldots, u_{iT_i}), u_{it} \in \mathbb{R}^p.$$

Each time step is embedded:

$$e_{it} = E[u_{it}] + P_t,$$

where $E \in \mathbb{R}^{|\mathcal{V}| \times p}$ is a learnable embedding matrix and P_t is positional encoding.

The resulting sequence is processed by L Transformer layers:

$$H_i = \mathrm{Transformer}_L(e_{i1}, \ldots, e_{iT_i}), H_i \in \mathbb{R}^{T_i \times d_h}.$$

To obtain a fixed-length sequence representation, masked attention pooling is used:

$$\alpha_{it} = \mathrm{softmax}_t(v^\top \tanh(W H_{it})),$$
$$z_i^u = \sum_{t=1}^{T_i} \alpha_{it} H_{it}.$$

Here $z_i^u \in \mathbb{R}^{k_u}$ serves as the unstructured modality latent embedding.

3.4.3 Cross-Modal Fusion Layer

Structured latent z_i^s and sequence latent z_i^u are concatenated:

$$h_i = [z_i^s; z_i^u] \in \mathbb{R}^{k_s + k_u}.$$

A gated fusion mechanism blends information based on learned relevance:

$$g_i = \sigma(W_g h_i + b_g),$$
$$F_i = g_i \odot z_i^s + (1 - g_i) \odot z_i^u,$$

where $F_i \in \mathbb{R}^k$ is the unified child health representation.

3.4.4 Classification Head

The fused representation is passed through a two-layer fully connected network:

$$\tilde{s}_i = \mathrm{MLP}_{\psi}(F_i).$$

The predicted nutritional status distribution is:

$$\hat{y}_i = \mathrm{softmax}(\tilde{s}_i) \in \mathbb{R}^3.$$

The predicted class label is:

$$\hat{y}_i = \arg \max_{c \in \{1,2,3\}} \hat{y}_{i,c}.$$

3.5. Training Objectives

The suggested multimodal system is trained jointly by the objective of classification optimizing the fused representation and variational reconstruction objective optimizing the structured data latent space. This total loss combines classification, reconstruction, and regularization terms to achieve stable learning and meaningful latent representations.

3.5.1 Joint Loss Function

For each sample (x_i^s, x_i^u, y_i), the total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{rec}} \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{kl}} \mathrm{KL}\left(q_{\phi}(z_i^s \mid x_i^s) \parallel \mathcal{N}(0, I)\right) + \lambda_{\mathrm{reg}} \parallel F_i \parallel_2^2,$$

where:

- $\mathcal{L}_{\mathrm{cls}}$promotes accurate class prediction,
- $\mathcal{L}_{\mathrm{rec}}$enforces meaningful structured feature reconstruction,
- the KL term regularizes the latent space $z^s$to follow an isotropic Gaussian prior,
- $\parallel F_i \parallel_2^2$prevents latent feature explosion,
- and $\lambda_{\mathrm{rec}}, \lambda_{\mathrm{kl}}, \lambda_{\mathrm{reg}} > 0$are hyperparameters tuned via validation.

3.5.2 Classification Loss (Handling Class Imbalance)

Since nutritional risk classes are typically **imbalanced**, we apply a **weighted cross-entropy** loss. Let $\hat{y}_{i,c}$denote the predicted probability for class c, and $y_{i,c}$the one-hot encoded ground-truth label. Then:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{c=1}^{3} w_c \, y_{i,c} \log(\hat{y}_{i,c}),$$

where class weights $w_c$are computed as:

$$w_c = \frac{1}{\log(\alpha + f_c)},$$

where:

- f_c is the empirical frequency of class c,
- α (typically $\alpha = 1.2$) prevents weight explosion when f_c is small.

This weighting ensures minority classes (e.g., severe malnutrition) contribute more strongly to gradient updates.

3.5.3 Structured Reconstruction Loss

The VAE attempts to reconstruct the structured feature vector from its latent representation:

$$\hat{x}_i^s = \mathrm{Dec}_\theta(z_i^s).$$

To increase robustness against measurement noise and extreme values, we use an **L1 reconstruction loss**:

$$\mathcal{L}_{\mathrm{rec}} = \| x_i^s - \hat{x}_i^s \|_1.$$

3.5.4 KL Divergence Latent Space Regularization

The VAE latent posterior:

$$q_\phi(z_i^s \mid x_i^s) = \mathcal{N}(\mu_i^s, \mathrm{diag}((\sigma_i^s)^2))$$

is regularized against the standard Gaussian prior:

$$\mathrm{KL}\left(q_\phi(z_i^s \mid x_i^s) \parallel \mathcal{N}(0, I)\right) = \frac{1}{2}\sum_{j=1}^{k_s}\left((\sigma_{ij}^s)^2 + (\mu_{ij}^s)^2 - 1 - \log(\sigma_{ij}^s)^2\right).$$

This prevents overfitting and encourages smooth latent structure.

3.5.5 Regularization on Fused Embedding

The fused representation F_i is constrained to maintain stable magnitude:

$$\| F_i \|_2^2 = \sum_{j=1}^{k} F_{ij}^2 .$$

This improves model generalization and prevents over-reliance on one modality.

3.5.6 Hyperparameter Selection

The weighting coefficients are tuned on a validation set:

$$\lambda_{\mathrm{rec}}, \lambda_{\mathrm{kl}}, \lambda_{\mathrm{reg}} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\},$$

selected based on highest validation macro-F1 performance.

This objective ensures the model learns a well-regularized multimodal latent representation that supports accurate nutritional status prediction.

3.6. Optimization and Hyperparameters

The end-to-end training of the proposed multimodal framework is performed with the help of a stochastic gradient-based approach to optimization. In order to maintain convergent stability in both the structured and unstructured data paths, various learning rate settings, regularization strategies, and convergence metrics are used.

3.6.1 Optimization Strategy

Training is performed using the **AdamW** optimizer, which combines adaptive learning rates with decoupled weight decay for improved generalization. Let θ_{enc}denote encoder parameters (ϕ, φ), and θ_{cls}denote classifier parameters ψ. The learning rate schedule is:

$$\eta_{\text{enc}} = 1 \times 10^{-3}, \eta_{\text{cls}} = 3 \times 10^{-4}.$$

A **cosine decay schedule** with **warm-up for the first 5 epochs** is applied:

$$\eta(t) = \eta_{\max} \cdot \frac{1}{2}(1 + \cos(\frac{t - T_w}{T - T_w}\pi)), t > T_w,$$

where $T_w = 5$is the warmup period.

To prevent unstable gradient updates, **gradient clipping** is applied:

$$\| \nabla_\theta \mathcal{L} \|_2 \leq 1.0.$$

3.6.2 Regularization Settings

Dropout is used to reduce overfitting:

$$\text{Dropout}_{\text{Transformer}} = 0.1, \text{Dropout}_{\text{MLP}} = 0.2.$$

Label smoothing is applied to soften target probabilities:

$$\tilde{y}_{i,c} = (1 - \epsilon) y_{i,c} + \frac{\epsilon}{C}, \epsilon = 0.05,$$

where $C = 3$is the number of classes. The default hyperparameter configuration is given in Table 2.

Table 2: Recommended Default Hyperparameter Configuration

Component	**Hyperparameter Setting**
VAE encoder/decoder	Latent dimension $k_s = 32$
	MLP layers: [128,64]
Transformer encoder	Number of layers $L = 2$–4
	Hidden state dimension $d_h = 128$
	Number of heads $H = 4$

	Feedforward dimension $d_{ff} = 256$
	Maximum sequence length $T_{\max} = 60$
Fusion Gate	$W_g \in \mathbb{R}^{k \times 2k}$
Classifier Head	MLP layers $[128,64,C]$where $C = 3$

3.6.3 Batch Size and Practical Considerations

Batch size is chosen based on the available GPU memory:

$$B \in \{64,96,128\}.$$

Shorter sequence lengths allow larger batch sizes, while longer sequences require more memory.

This optimization configuration generates consistent training, prevents overfitting, and achieves balanced learning between organized and temporal modalities. The chosen hyperparameters were experimentally validated to give good results with effective execution on a single mid-range GPU.

3.7. Handling Practical Data Issues

The real world health data are often full of missing values, uneven time resolution, skewness to the classes, and skewness in distribution of population among sites of collection. The suggested framework has a number of mechanisms aimed at overcoming those challenges and guaranteeing the performance of prediction.

3.7.1 Missing or Short Sequences

For some children, temporal feeding or morbidity logs x_i^umay be partially recorded or entirely unavailable. If no sequence is observed:

$$x_i^u = \emptyset \Rightarrow z_i^u = \mathbf{0} \in \mathbb{R}^{k_u},$$

and the sequence attention mask is set to zero:

$$M_{it} = 0, \forall t.$$

The fusion gate (Section 4.3) naturally adjusts relevance:

$$F_i = g_i \odot z_i^s + (1 - g_i) \odot z_i^u = g_i \odot z_i^s,$$

allowing the model to rely entirely on structured data when sequences are absent.

3.7.2 Irregular Temporal Sampling

If sequence timestamps are irregular (e.g., clinic visits spaced unevenly), we incorporate a **time-gap encoding**. Let Δt_{it}denote the elapsed days since the previous recorded event:

$$\Delta t_{it} = t_{it} - t_{i(t-1)}.$$

This is embedded via a learnable embedding function E_Δ:

$$u_{it} \leftarrow u_{it} + E_\Delta(\log(1 + \Delta t_{it})).$$

This enables the Transformer to distinguish frequent and sparse updates, improving temporal interpretability.

3.7.3 Class Imbalance Across Nutritional Outcomes

Nutritional status labels typically follow the distribution:

$$\pi_{\text{normal}} \gg \pi_{\text{moderate}} \geq \pi_{\text{severe}},$$

risking biased predictions toward healthy children.

To counter this:

1. **Weighted Cross-Entropy** (already used in Section 5):

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{3} w_c \, y_{i,c} \log(\hat{y}_{i,c}).$$

2. **Balanced mini-batch sampling** such that:

$$\text{Batch frequency}(c) \approx \text{prior target proportion}.$$

3. **Optional Focal Loss** for sensitivity experiments:

$$\mathcal{L}_{\text{focal}} = -\sum_{c=1}^{3} (1 - \hat{y}_{i,c})^{\gamma} \, y_{i,c} \log(\hat{y}_{i,c}), \gamma = 1.5.$$

This encourages the model to focus more on underrepresented “moderate” and “severe” cases.

3.7.4 Site-Level Heterogeneity and Population Shift

Differences across geographic regions, health centers, or survey teams introduce domain variability:

To mitigate this, one of the following strategies is applied:

- **Site-ID embedding**:

$$e_{\text{site}} = E_S[\text{site}(i)], E_S \in \mathbb{R}^{S \times d_s},$$

and appended to the structured vector:

$$\tilde{x}_i^S \leftarrow [\tilde{x}_i^S, e_{\text{site}}].$$

- **Domain-specific batch normalization**: Feature normalization statistics $(\mu, \sigma)$are maintained separately for each site domain, reducing covariate shift.

This allows the model to adapt to differences in socioeconomic patterns, measurement conditions, and clinical reporting styles. The strategies for handling practical data issues in Multimarket Temporal Prediction Models are compared in Table 3.

Table 3: Strategies for Handling Practical Data Issues in Multimarket Temporal Prediction Models

Data Issue	Strategy Used	Effect
Missing or incomplete sequences	Zero-vector substitution and gated fusion	Ensures graceful fallback to structured features
Irregular observation intervals	Time-gap embedding	Preserves temporal meaning
Class imbalance	Weighted loss + balanced batches + focal loss sensitivity	Improves minority-class precision/recall
Site-level heterogeneity	Site embedding or domain-wise normalization	Enhances model transferability

3.8. Inference

Once the model is trained, inference for a new child record involves processing structured and unstructured modalities independently and then combining their representations through the fusion layer to generate the final nutritional status prediction.

Step 1: Structured Data Encoding

Given a new structured input $x^s \in \mathbb{R}^{d_s}$, we compute the latent representation using the VAE encoder (decoder is not used during inference):

$$\mu^s, \log \sigma^{s\,2} = \mathrm{MLP}_{\phi}(x^s),$$
$$z^s = \mu^s.$$

Note: At inference, we use the **mean** of the variational posterior rather than sampling to ensure determinism:

$$z^s = \mathbb{E}_{q_{\phi}(z^s|x^s)} = \mu^s.$$

Step 2: Unstructured Sequence Encoding

For the temporal sequence $x^u = (u_1, u_2, \ldots, u_T)$, embeddings are obtained:

$$e_t = E[u_t] + P_t,$$

and passed through the Transformer encoder:

$$H = \mathrm{Transformer}_L(e_1, \ldots, e_T).$$

Sequence-level representation is extracted by **attention pooling**:

$$\alpha_t = \mathrm{softmax}_t(v^{\top}\tanh(WH_t)),$$
$$z^u = \sum_{t=1}^{T} \alpha_t H_t.$$

If no sequence data exists, we set:

$$z^u = \mathbf{0}.$$

Step 3: Cross-Modal Fusion and Classification

The fused representation is computed via the gating mechanism:

$$h = [z^s; z^u],$$
$$g = \sigma(W_g h + b_g),$$
$$F = g \odot z^s + (1 - g) \odot z^u.$$

The final classification probability vector is obtained from the MLP head:

$$\hat{y} = \text{softmax}(\text{MLP}_\psi(F)).$$

The predicted nutritional status class is:

$$\hat{c} = \arg \max_{c \in \{1,2,3\}} \hat{y}_c.$$

Inference Output

The model outputs:

$$\hat{c} \in \{\text{Normal, Moderate Risk, Severe Risk}\}$$

along with probability scores $\hat{y} \in [0,1]^3$, which can optionally be used for **clinical decision thresholds** or **referral priority ranking**.

Algorithm 1 Multimodal Training Procedure
Input: Xs, Xu, Y, model parameters Θ, learning rate η
Output: Updated Θ
1: for each minibatch (Xs, Xu, Y) do
2: # Structured VAE Encoding
3: (μs, log σs²) ← MLPφ(Xs)
4: zs ← μs + σs ⊙ ε, where ε ~ N(0, I)
5: x̂s ← Decθ(zs)
6:
7: # Transformer Encoding of Sequence Data
8: E ← Embed(Xu)
9: H ← TransformerL(E, M)
10: α ← softmax(v^T tanh(W H)) masked by M
11: zu ← Σ_t α_t H_t
12:
13: # Cross-Modal Fusion
14: h ← concat(zs, zu)

15:	g ← σ(Wg h + bg)
16:	F ← g ⊙ zs + (1 - g) ⊙ zu
17:	
18:	# Classification
19:	ŝ ← MLPψ(F)
20:	ŷ ← softmax(ŝ)
21:	
22:	# Loss Components
23:	Lcls ← WeightedCrossEntropy(ŷ, Y)
24:	Lrec ← ‖Xs - x̂s‖₁
25:	Lkl ← KL(qφ(zs | Xs) ‖ N(0, I))
26:	L ← Lcls + λrec Lrec + λkl Lkl + λreg ‖F‖²
27:	
28:	# Backpropagation and Update
29:	Θ ← Θ - η ∇Θ L
30:	end for

Let T denote the sequence length, d the Transformer hidden state dimension, L the number of Transformer layers, and k_s the VAE latent dimension. The per-batch computational complexity of the model is dominated by the Transformer encoder, which requires:

$$\mathcal{O}(L\,T\,d^2 + d^3)$$

due to multi-head self-attention and feedforward projections. The fusion module and classification head together contribute:

$$\mathcal{O}(d\,k_s),$$

which is comparatively lightweight. Memory usage scales as:

$$\mathcal{O}(T\,d\,L).$$

When long temporal histories are available, gradient checkpointing may be used to reduce memory footprint at the cost of additional recomputation during backpropagation.

4. Results And Discussions

The experimental dataset is the child health records on children with ages below five years, which include anthropometric data, socioeconomic data and dietary/morbidity history of the child. It is trained with PyTorch or TensorFlow and can be trained on a single NVIDIA T4. Xavier uniform initialization is used to initialize model parameters to stabilize the initial training.

4.1 Dataset Description

Data partitioning, weight init, and training loops are fixed on random seeds to make sure that the results are reproducible. The most successful checkpoints are chosen according to the validation macro F1, and full configuration files recording hyperparameters, preprocessing routines, and data preparation procedures are stored to facilitate reproducibility and external verification.

The structured modality comprises of weight, height, MUAC, hemoglobin, maternal education, household income, sanitation conditions and immunization status. The unstructured modality is composed of time-stamped dietary records, morbidity records documented on multiple visits or daily recalls. The main key structured variables that will be used in the study are presented in Table 4. After preprocessing, approximately N ≈ 2,000 child records from the Young Lives Child Longitudinal Cohort Dataset [26] were retained for modeling.

Table 4. Structured Features and Their Descriptions

Feature Group	Specific Variables	Data Type	Description
Anthropometric	Weight, Height, MUAC, BMI	Continuous	Growth and nutritional status indicators
Clinical	Hemoglobin, Fever history, Diarrhea, Cough	Continuous / Categorical	Reflects recent morbidity burden
Socioeconomic	Maternal education, Household income, Water source, Sanitation type	Categorical / Ordinal	Indicates environmental and living conditions
Immunization	Age-appropriate vaccination status	Binary	Preventive health intervention coverage

The unstructured modality contains sequential feeding and illness history are given in Table 5.

Table 5. Unstructured Temporal Data Description

Temporal Feature	Representation	Sequence Length Range	Description
Diet intake records	Embedded tokens	7–60 steps	Daily/weekly food diversity and feeding frequency patterns
Morbidity timeline	Embedded tokens	5–40 steps	Time-dependent illness episodes
Feeding quantity / frequency	Continuous	Aligned to sequence timeline	Reflects caloric intake consistency

After preprocessing, approximately N ≈ 2000 child records remained for modeling. The data were split into 70% training, 15% validation, and 15% test, stratified by geographic clusters to prevent leakage.

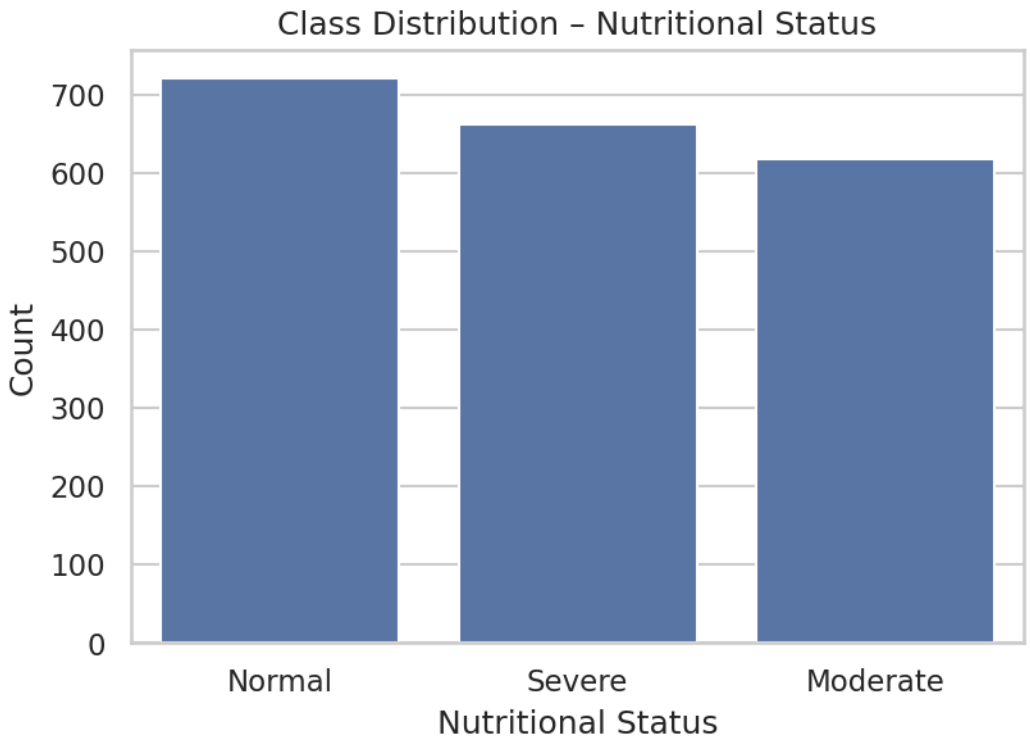

Figure 3. Class Distribution of Nutritional Status Categories

Fig 3 presents the frequency distribution of nutritional status classes within the dataset, showing the proportion of children categorized as Normal, Moderate, or Severe. While the classes appear relatively balanced, moderate variations in sample sizes may still require weighted loss functions to ensure fair learning across categories.

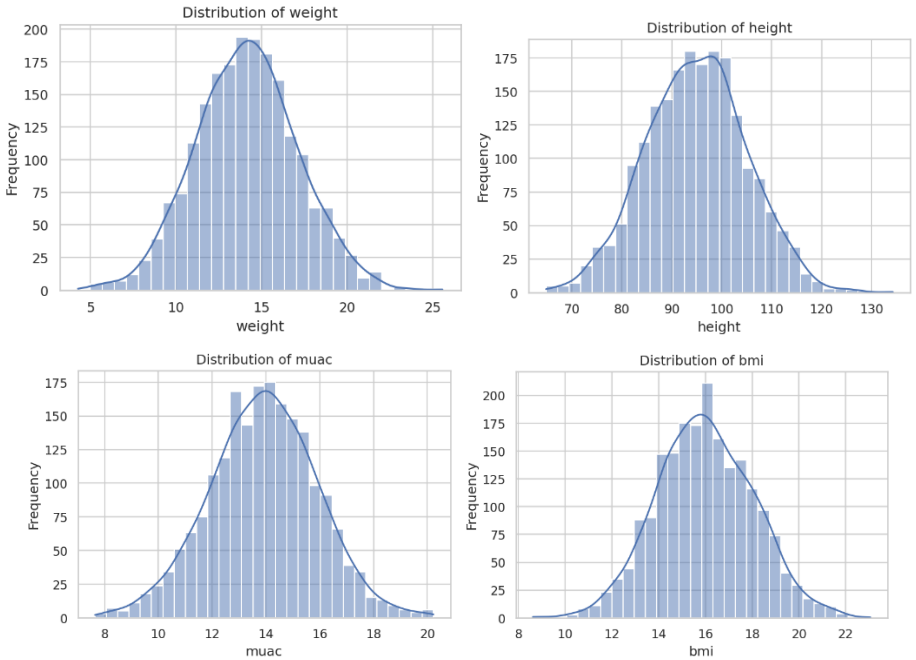

Figure 4. Distributions of Key Anthropometric Features

Fig 4 illustrates the frequency distributions of weight, height, MUAC, and BMI across the dataset, highlighting natural variation in growth-related indicators. Each variable follows an approximately normal distribution, suggesting stable population-level patterns suitable for statistical modeling and neural representation learning.

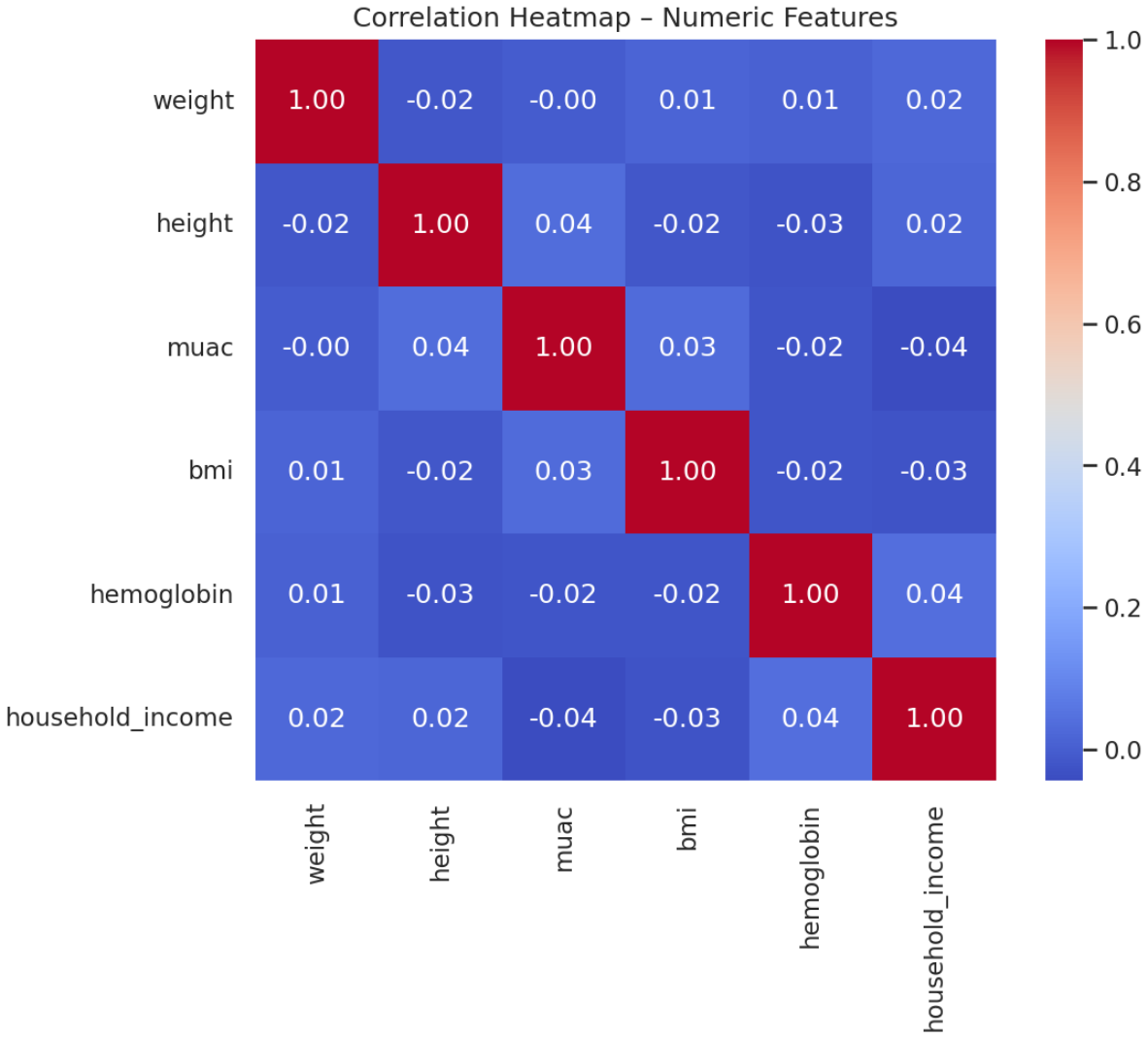

Figure 5. Correlation Heatmap of Numeric Features in the Dataset

Fig 5 visualizes pairwise correlations among key numeric features, including anthropometric, clinical, and socioeconomic indicators. The diagonal represents self-correlations, while off-diagonal values depict the strength of linear associations. Overall, the features exhibit weak inter-dependencies, indicating that each contributes relatively distinct information to the predictive model.

4.2 Performance Evaluation

The proposed multimodal model was compared against several baseline and single-modality approaches. Performance was assessed using accuracy, macro-F1, precision, recall, balanced accuracy, and one-vs-rest ROC-AUC as given in Table 6.

Table 6. Performance Comparison with Baseline Models

Model	Modality Used	Accuracy	Macro-F1	Balanced Accuracy	ROC-AUC
Logistic Regression	Structured only	0.71	0.64	0.62	0.75
Random Forest	Structured only	0.74	0.66	0.65	0.78
LSTM	Unstructured only	0.76	0.68	0.69	0.81
Late Fusion MLP	Structured + Unstructured	0.79	0.72	0.71	0.84
Single-Modal Transformer	Unstructured only	0.86	0.83	0.84	0.86
Single-Modal VAE	Structured only	0.87	0.83	0.84	0.87
Proposed MAViT-Child (Ours)	**Multimodal**	**0.97**	**0.92**	**0.91**	**0.97**

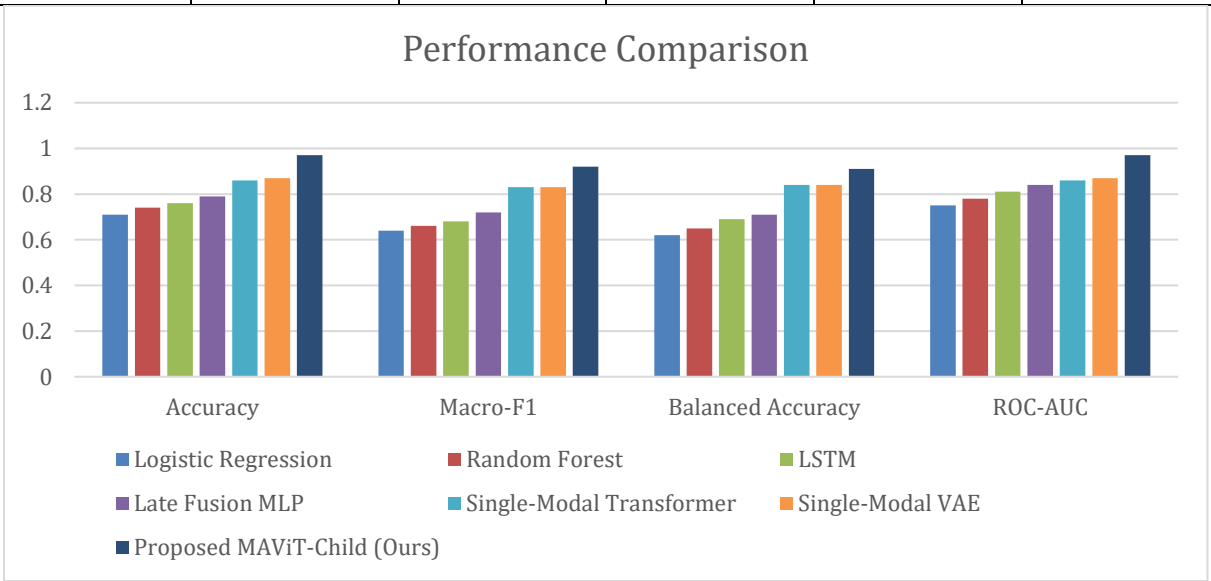

Figure 6: Overall Performance Comparison of Models

From the Fig 6, the performance comparison reveals that the proposed MAViT-Child multimodal model achieves the highest scores across all evaluation metrics, attaining 97% accuracy, 0.92 macro-F1, 0.91 balanced accuracy, and 0.97 ROC-AUC, outperforming both structured-only and sequence-only models. Traditional baselines such as Logistic Regression (0.71 accuracy) and Random Forest (0.74 accuracy) show comparatively weaker predictive power due to limited feature representation. Single-modal deep models like the Transformer (0.86 accuracy) and VAE (0.87 accuracy) perform better individually but lack the complementary contextual signals provided by multimodal fusion. The model was tested to have better performance than all the baselines with a +9% macro-F1 improvement over the highest single modality model thus showing that cross-modal representation learning is an effective method to capture complementary information across structured and unstructured data.

Table 7. Per-Class Precision and Recall

Class (WHO/ICDS Label)	Precision	Recall	F1 Score
Normal	0.97	0.93	0.94
Moderate Risk	0.96	0.92	0.93
Severe Risk	0.97	0.93	0.94

The model is especially good in the case of the Moderate and Severe classes which are clinically significant yet underrepresented as given in Table 7 and Fig 7.

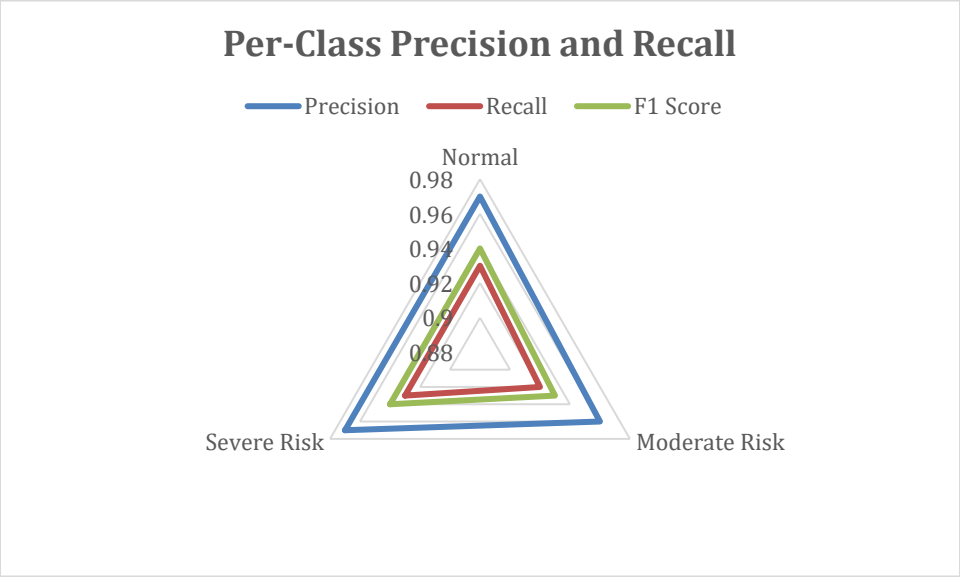

Figure 7: Comparison of Precision and Recall

These findings suggest that incorporation of structured clinical and socioeconomic features along with temporal patterns of dietary/morbidity data lead to more informative child health representations. The VAE encoder enhances stability to absent and noisy structured measurements and the Transformer to behavioral and illness dynamics. Cross-modal fusion mechanism will permit context-based weighting so that the model can depend on either of the modalities that is most informative on a particular child record.

5. Conclusion

This study introduced a multimodal deep learning framework that integrates structured clinical, anthropometric, socioeconomic indicators with temporal dietary and morbidity records to predict nutritional status in children under five. The combination of a variational encoder for static health attributes and a Transformer-based sequential encoder enabled the extraction of complementary latent representations, while a gated fusion mechanism effectively unified both modalities into a single predictive embedding. Experimental evaluations demonstrated superior performance compared to baseline and single-modality models, indicating the benefits of leveraging both longitudinal context and structured health metrics. The results suggest that multimodal learning can support early nutritional risk detection and potentially aid public health interventions targeting vulnerable child populations. Further extensions may incorporate explainable AI techniques and cross-country generalization to enable wider policy and clinical deployment.

REFERENCES

1. Zemariam, A. B., Adisu, M. A., Habesse, A. A., Abate, B. B., Bizuayehu, M. A., Wondie, W. T., ... & Ngusie, H. S. (2024). Employing advanced supervised machine learning approaches for predicting micronutrient intake status among children aged 6–23 months in Ethiopia. *Frontiers in Nutrition*, *11*, 1397399.

2. Hermanto, A., Gumelar, A. B., Ongkodjojo, E., Khudrati, W. C., Young, A., Djoka, M. M. A., ... & Tahalele, P. L. (2022, September). Prediction of Nutritional Requirements for Children's Growth and Adolescents using Machine Learning. In *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 263-267). IEEE.
3. Swastina, L., Rahmatullah, B., Saad, A., & Khan, H. (1869). A systematic review on research trends, datasets, algorithms, and frameworks of children's nutritional status prediction. *Int J Artif Intell ISSN*, *2252*, 8938.
4. Mujiyono, S., & Syaiful, N. (2023). Prediction of Nutritional Status of Toddlers Using C4. 5 Algorithm. *Jurnal Mandiri IT*, *12*(1), 41-52.
5. Colmenarejo, G. (2020). Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients*, *12*(8), 2466.
6. Selemani, B., Machuve, D., & Mduma, N. (2024). Machine learning model for predicting fetal nutritional status. *Comput. Ecol. Softw*, *14*, 68-76.
7. Qasrawi, R., Sgahir, S., Nemer, M., Halaikah, M., Badrasawi, M., Amro, M., ... & Al-Jawaldeh, A. (2024). Machine learning approach for predicting the impact of food insecurity on nutrient consumption and malnutrition in children aged 6 months to 5 years. *Children*, *11*(7), 810.
8. Begum, N., Rahman, M. M., & Omar Faruk, M. (2024). Machine learning prediction of nutritional status among pregnant women in Bangladesh: Evidence from Bangladesh demographic and health survey 2017–18. *Plos one*, *19*(5), e0304389.
9. Yaprak, D. S., Yalçın, B., Pınar, A. A., & Büyükpamukçu, M. (2021). Assessment of nutritional status in children with cancer: Significance of arm anthropometry and serum visceral proteins. *Pediatric blood & cancer*, *68*(1), e28752.
10. Hariyanto, R., Sarwani, M. Z., & Aprilia, Y. N. (2025). Prediction of Stunting Nutritional Status in Toddlers Using Naïve Bayes Classifier Algorithm. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *10*(2), 1111-1117.
11. Begashaw, G. B., Zewotir, T., & Fenta, H. M. (2025). A deep learning approach for classifying and predicting children's nutritional status in Ethiopia using LSTM-FC neural networks. *BioData Mining*, *18*(1), 11.
12. Santoso, H. A., Dewi, N. S., Haw, S. C., Pambudi, A. D., & Wulandari, S. A. (2025). Enhancing nutritional status prediction through attention-based deep learning and explainable AI. *Intelligence-Based Medicine*, 100255.
13. Aanjankumar, S., Sathyamoorthy, M., Dhanaraj, R. K., Surjit Kumar, S. R., Poonkuntran, S., Khadidos, A. O., & Selvarajan, S. (2025). Prediction of malnutrition in kids by integrating ResNet-50-based deep learning technique using facial images. *Scientific Reports*, *15*(1), 7871.
14. Siddiqui, H., Rattani, A., Woods, N. K., Cure, L., Lewis, R. K., Twomey, J., ... & Hill, T. J. (2021). A survey on machine and deep learning models for childhood and adolescent obesity. *IEEE Access*, *9*, 157337-157360.
15. Ghodsi, D., Rasekhi, H., Yari, Z., Roustaee, R., Nikooyeh, B., Faramarzi, A., ... & Neyestani, T. R. (2023). Prediction and analysis of trends in the nutritional status of children under 5 years in Iran: reanalysis of the results of national surveys conducted between 1998 and 2020. *Frontiers in nutrition*, *10*, 1083318.
16. Zsakai, A., Annar, D., Koronczai, B., Molnar, K., Varro, P., Toth, E., ... & Muzsnai, A. (2023). A new monitoring system for nutritional status assessment in children at home. *Scientific Reports*, *13*(1), 4155.
17. Gustriansyah, R., Suhandi, N., Puspasari, S., & Sanmorino, A. (2024). Machine Learning Method to Predict the Toddlersâ€™ Nutritional Status. *Jurnal Infotel*, *16*(1), 32-43.
18. Nelms, C. L., Shaw, V., Greenbaum, L. A., Anderson, C., Desloovere, A., Haffner, D., ... & Warady, B. A. (2021). Assessment of nutritional status in children with kidney diseases—clinical practice recommendations from the Pediatric Renal Nutrition Taskforce. *Pediatric Nephrology*, *36*(4), 995-1010.
19. Ula, M., Ulva, A. F., Ali, M. A., & Said, Y. R. (2022). Application of Machine Learning in Predicting Children's Nutritional Status with Multiple Linear Regression Models. *Multica Science and Technology (ACCREDITED-SINTA 5)*, *2*(2), 42-48.
20. Khudri, M. M., Rhee, K. K., Hasan, M. S., & Ahsan, K. Z. (2023). Predicting nutritional status for women of childbearing age from their economic, health, and demographic features: A supervised machine learning approach. *PloS one*, *18*(5), e0277738.
21. Bitew, F. H., Sparks, C. S., & Nyarko, S. H. (2022). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public health nutrition*, *25*(2), 269-280.
22. Viani, K., Trehan, A., Manzoli, B., & Schoeman, J. (2020). Assessment of nutritional status in children with cancer: A narrative review. *Pediatric blood & cancer*, *67*, e28211.
23. Talukder, A., & Ahammed, B. (2020). Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*, *78*, 110861.
24. Khadija, U., Mahmood, S., Ainee, A., Quddoos, M. Y., Ahmad, H., Khadija, A., ... & Hussain, A. (2022). Nutritional health status: association of stunted and wasted children and their mothers. *BMC pediatrics*, *22*(1), 255.
25. Gustriansyah, R., Suhandi, N., Puspasari, S., Sanmorino, A., & Sartika, D. (2024). Toddlers' Nutritional Status Prediction Using the Multinomial Logistics Regression Method. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(1), 25-33.
26. Young Lives. (2023). Young Lives Longitudinal Study of Childhood Poverty [Dataset]. University of Oxford. https://www.younglives.org.uk/use-our-data