Counterfactual Explanation Model for Personalised Dietary Interventions in Anaemia Patients

Zvinodashe Revesai[1][0009-0008-2284-6097] and Okuthe P. Kogeda *[0000-0002-8353-8345]

[1] School of Mathematics, Statistics and Computer Science, College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, Durban 3209, Republic of South Africa.
224195689@stu.ukzn.ac.za, kogedao@ukzn.ac.za*

Abstract. Deep learning has revolutionised healthcare applications, achieving remarkable success in medical diagnosis and treatment prediction. However, the inherent opacity of these models presents significant challenges for clinical deployment, where interpretable explanations are crucial for patient trust and regulatory compliance. This paper presents a novel constraint-aware counterfactual explanation model for generating personalised dietary interventions in anaemia patients. Anaemia affects over 1.9 billion people globally, yet existing explainable AI methods fail to provide clinically feasible and culturally appropriate recommendations. We develop a causal machine learning approach that integrates Pearl's causal hierarchy with domain-specific constraints to produce interpretable "what-if" scenarios. Our model incorporates nutritional, cultural, and economic constraints through augmented Lagrangian optimisation, ensuring recommendations remain clinically feasible whilst maintaining semantic meaningfulness. Experimental results demonstrate superior performance compared to existing explainable AI methods, achieving 84.3% anaemia reversal rates (vs 71.8% best baseline), 89.1% counterfactual validity, and 4.2 interpretability scores. The model generates recommendations requiring an average of 2.3 dietary changes within cognitive load thresholds whilst maintaining $O(n\ log\ n)$ computational complexity suitable for real-time clinical deployment. This work advances explainable AI in healthcare by demonstrating how domain-specific constraints can enhance both interpretability and clinical utility of counterfactual explanations for chronic disease management.

Keywords: Counterfactual Explanations, Explainable AI, Causal Machine Learning, Healthcare Informatics, Interpretability

1 Introduction

Deep learning has revolutionised healthcare applications, achieving remarkable success in medical image analysis, drug discovery, and clinical decision support systems [1]. However, the inherent opacity of these models presents significant challenges for healthcare deployment, where clinical decisions require transparent reasoning and accountability [2]. The black box nature limits adoption in critical scenarios, as medical professionals need to understand algorithmic recommendations to ensure patient safety and regulatory compliance. Consequently, developing interpretable machine learning

methods has become paramount for bridging the gap between algorithmic sophistication and clinical acceptability [3].

Anaemia represents a significant global health burden, affecting approximately 1.9 billion individuals worldwide [4]. The condition disproportionately impacts vulnerable populations, with children under five experiencing 42% prevalence globally and women of reproductive age facing 33% rates, particularly severe in sub-Saharan Africa where prevalence exceeds 60% [4]. Whilst nutritional deficiencies serve as primary aetiological factors, vulnerable populations face compounded challenges including limited food access, cultural dietary restrictions, and economic constraints that render standard interventions ineffective [5]. Clinical thresholds classify anaemia severity based on haemoglobin levels, yet approaches often fail to address complex socioeconomic and cultural barriers [6].

Effective anaemia management requires personalised dietary strategies considering individual patient characteristics, metabolic profiles, and environmental factors. Current clinical practice relies on standardised nutritional guidelines, presenting limitations including insufficient individualisation, limited consideration of patient-specific constraints, and absence of systematic follow-up mechanisms [6], [7]. Consequently, patient compliance rates remain suboptimal with considerable treatment efficacy variation across population groups.

Contemporary advances in explainable artificial intelligence introduce novel opportunities for developing sophisticated yet interpretable nutritional intervention systems. These approaches can analyse complex, multidimensional patient data to generate tailored recommendations whilst providing transparent reasoning for clinical validation. Nevertheless, implementing interpretable AI in nutritional healthcare introduces distinct computational challenges requiring highly adaptive algorithmic approaches that maintain interpretability without sacrificing predictive accuracy.

Recent developments in counterfactual explanation methods have garnered considerable attention within healthcare informatics, particularly for applications requiring high trust and accountability. Successful deployment demands sophisticated understanding of algorithm selection, constraint modelling, causal inference, and domain-specific validation methodologies [8]. Suboptimal implementation can result in recommendations violating clinical guidelines, ignoring cultural sensitivities, or exceeding economic feasibility thresholds. Whilst counterfactual methods have achieved competitive performance in general healthcare explanation tasks [10], their application to nutritional interventions necessitates specialised modifications incorporating dietary constraints, cultural preferences, and economic considerations [11].

Current methodologies present significant limitations when applied to nutritional intervention planning, prioritising mathematical optimisation without adequate consideration of domain-specific constraints. This frequently produces recommendations lacking practical applicability due to cultural inappropriateness, economic infeasibility, or nutritional imbalance, severely limiting clinical utility and interpretability.

This research addresses these limitations by developing a specialised counterfactual explanation system for personalised anaemia interventions, validated using comprehensive demographic and health survey data from Ethiopian populations, prioritising interpretability whilst ensuring clinical relevance and cultural appropriateness.

Our primary contributions include:

1. Development of a causal machine learning architecture incorporating nutritional science principles with counterfactual generation algorithms to produce clinically relevant and interpretable explanations.
2. Implementation of a multi-constraint optimisation framework ensuring cultural appropriateness, economic feasibility, and nutritional adequacy in generated recommendations whilst maintaining high interpretability.
3. Design of a comprehensive evaluation protocol measuring both computational performance and clinical utility through multi-stakeholder validation processes, emphasising interpretability assessment.
4. Empirical validation using large-scale demographic health survey data, demonstrating improved performance compared to existing explainable AI approaches across interpretability, clinical validity, and cultural appropriateness metrics.

The remainder of this paper is organised as follows: In Section 2, we review related work in counterfactual explanation algorithms, explainable AI in healthcare, causal machine learning, and nutritional informatics. In Section 3, we describe the proposed constraint-aware counterfactual explanation model for personalised anaemia interventions. In Section 4, we present the experimental results and performance evaluation. We discuss these results in Section 5. Lastly, in Section 6, we provide concluding remarks and future work directions.

2 Related Work

This section examines counterfactual explanation algorithms, explainable AI in healthcare, causal machine learning, and nutritional informatics, focusing on interpretability challenges in healthcare AI systems.

2.1 Counterfactual Explanation Algorithms

Counterfactual explanations identify minimal input changes that alter predictions. Mothilal et al. [12] proposed Diverse Counterfactual Explanations (DICE) for diverse explanations with $O(n^2)$ complexity. Wachter et al. [13] introduced gradient-based optimization but lacked semantic constraints, producing unrealistic recommendations. Poyiadzi et al. [14] developed Feasible and Actionable Counterfactual Explanations (FACE) ensuring realistic counterfactuals but at $O(n^3)$ complexity. Laugel et al. [15] offered intuitive geometric search but lacked convergence guarantees. These methods optimize for mathematical feasibility but fail to ensure domain semantic validity, generating explanations that are technically correct yet clinically meaningless or impractical.

2.2 Interpretable AI in Healthcare

Healthcare requires high interpretability for life-critical decisions and regulatory compliance. Ribeiro et al. [16] introduced Local Interpretable Model-agnostic Explanations (LIME) for local approximations but with instability issues. Lundberg and Lee [17]

offered principled game-theoretic SHapley Additive exPlanations (SHAP) but with $O(2^n)$ complexity. Ribeiro et al. [18] provided clinical guideline-like Anchors but with poor population coverage. Recent surveys [19], [20], [21] highlight needs for systems bridging technical and clinical requirements. These techniques provide post-hoc explanations without guaranteeing actionability or clinical relevance, creating a disconnect between model transparency and practical decision-making.

2.3 Casual Machine Learning

Pearl [22] established three interpretability levels: associational, interventional, and counterfactual. The PC algorithm [23] learns causal structures but has strong assumptions and O(2^k) complexity. Doubly robust methods [24] improve robustness but limit personalization through population-level focus. Recent work [25] emphasizes evaluation frameworks balancing accuracy and comprehension. Causal methods provide theoretical rigor but require strong assumptions and focus on population-level effects, limiting their ability to generate personalized, individual-level recommendations.

2.4 Nutritional Informatics

Nutritional systems face interpretability challenges from complex nutrient interactions and individual variations [26]. Traditional collaborative filtering offers limited explanatory reasoning. Matrix factorization [27] identifies latent factors without semantic meaning. Multi-criteria decision analysis [28] represents trade-offs but misses nutrient synergies. Deep learning approaches sacrifice interpretability for accuracy. Existing nutritional AI systems either lack semantic transparency or fail to capture complex nutrient interactions, preventing clinicians and patients from understanding and trusting dietary recommendations.

Existing approaches exhibit critical interpretability limitations: semantic gaps producing nonsensical explanations, computational-interpretability trade-offs, insufficient clinical alignment, and personalization deficiencies. Our constraint-aware counterfactual explanation model addresses these through domain-constrained optimization, providing clinically meaningful, computationally efficient, and culturally appropriate recommendations for anaemia management.

3 MODEL DESIGN

This section presents our constraint-aware counterfactual explanation model for personalized dietary interventions in anaemia patients, detailing computational formulation, algorithmic components, and implementation strategies.

3.1 Problem Formulation

We formulate counterfactual explanation as constrained optimization balancing interpretability, feasibility, and clinical efficacy. Given patient profile $x \in \mathbb{R}^d$ (demographic, dietary, biomarker features) and predictive model $f: \mathbb{R}^d \rightarrow \mathbb{R}$ mapping profiles to

haemoglobin levels, we identify counterfactual x' achieving target haemoglobin improvement whilst minimizing intervention complexity.

Wachter et al. [13] introduced gradient-based counterfactual generation through unconstrained distance minimization but lacked domain constraints, producing semantically invalid recommendations. We extend their framework incorporating nutritional, cultural, and economic constraints.

Our formulation as shown in equation (1) is:

$$\text{minimize: } L(x', x) = \lambda_1 d(x', x) + \lambda_2 \ell(f(x'), y*) + \lambda_3 R(x')$$
$$\text{subject to: } f(x') \geq Hb\_target, x' \in \Phi(x), g_i(x') \leq 0 \quad (1)$$

where $L(x', x)$ is overall $loss, \lambda_1, \lambda_2, \lambda_3$ are regularization parameters, $d(x', x)$ is distance between instances, $\ell(f(x'), y*)$ is prediction loss, $R(x')$ is interpretability regularization, Hb_target is desired haemoglobin (≥11.0 g/dL children, $\geq 12.0\ g/dL$ women per World Health Organization (WHO) [26]), $\Phi(x)$ is feasible modification space, and $g_i(x') \leq 0$ represents domain constraints ($i = 1, \ldots, m$ where m is total number of constraints).

The sparsity-promoting distance follows $L_1\ norm$ [30], encouraging fewer modified features. Regularization $R(x')$ penalizes complex modifications as expressed in equation (2):

$$R(x') = \Sigma_i\, \omega_i \cdot I(|x'_i - x_i| > \tau_i) \quad (2)$$

where ω_i represents feature importance weights, $I(\cdot)$ is indicator function, x'_i is counterfactual value for feature i, x_i is original value for feature i, and τ_i denotes minimum meaningful change thresholds (≥50g/day staple foods, ≥10g/day protein) [26].

3.2 Process Flow and System Designs

Fig.1 illustrates our process flow for constraint-aware counterfactual generation, showing integration of causal inference, constraint optimization, and interpretability generation. The architecture comprises five stages: (1) data input processing, (2) causal analysis (Directed Acyclic Graph (DAG) construction, backdoor adjustment, effect estimation), (3) optimization (gradient computation, constraint handling via augmented Lagrangian and projection, multi-objective solving), (4) personalization (population clustering, individual profiling, template generation), and (5) clinical output (safety validation, cost-benefit analysis, intervention protocols).

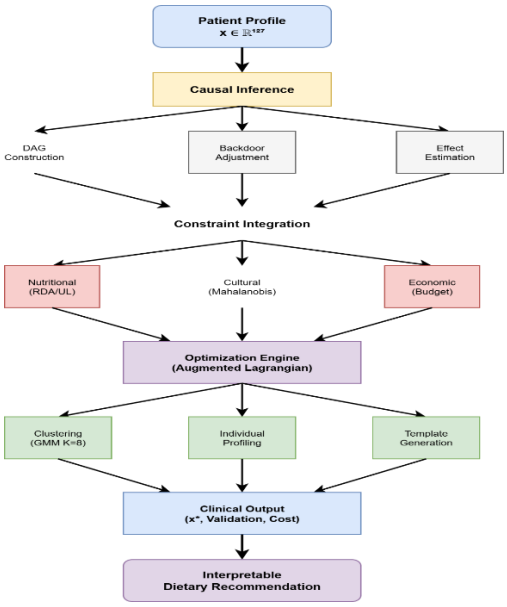

Fig 1: Process Flow for Constraint-Aware Counterfactual Generation

3.3 Causal Inference Model Design

To generate reliable intervention recommendations, we integrate Pearl's causal framework [22]. Pearl's hierarchy distinguishes three reasoning levels: (1) associational ($P(Y|X)$) observing correlations, (2) interventional ($P(Y|do(X))$) predicting action effects, and (3) counterfactual ($P(Y\_x'|X = x)$) reasoning about alternatives. Traditional machine learning operates association ally, identifying patterns reflecting spurious correlations rather than causal mechanisms. Clinical decision-making requires interventional reasoning.

Following backdoor adjustment as shown in equation (3):

$$P(Y = y|do(X = x)) = \Sigma\_z\, P(Y = y|X = x, Z = z)P(Z = z) \quad (3)$$

where $P(Y = y|do(X = x))$ is interventional probability, $P(Y = y|X = x, Z = z)$ is conditional probability given features and confounders, $P(Z = z)$ is marginal distribution of confounders, and Z represents adjustment set of confounding variables.

Causal graph construction employs constraint-based methods identifying confounders (socioeconomic status, maternal education, geographic location, seasonal factors) from Ethiopian Demographic and Health Survey (EDHS) dataset satisfying backdoor criterion [23]. This ensures counterfactual explanations represent genuine intervention effects.

3.4 Constraint-Aware Optimization Model

Our algorithm employs augmented Lagrangian method [32] handling mixed constraints in equation (1), providing better convergence than pure penalty approaches whilst maintaining robustness [24].

Algorithm 1: Constraint-Aware Counterfactual Generation

Input: Patient profile x, target Hb improvement Δ
Output: Counterfactual x with interpretable explanation*

1. *Initialize $x^0 = x$, penalty parameter $\mu = 0.1$*
2. *Construct constraint sets: Φ_nutritional, Φ_cultural, Φ_economic*
3. *For iteration t = 0 to max_iterations:*
4. *Compute causal gradient: ∇f_causal(x^t)*
5. *Calculate constraint violations using equation (4):*
6. *v_total = Σ_i max(0, $g_i(x^t)$) (4)*
7. *Update solution: $x^{t+1} = x^t - \alpha(\nabla L + \mu \nabla v\_total)$*
8. *Project onto feasible set using equation (5):*
9. *$x^{t+1} = \Pi_\Phi(x^{t+1})$ (5)*
10. *If convergence: break*
11. *Generate semantic explanation using personalized templates*
12. *Return optimized counterfactual x* and explanation*

where v_total is total constraint violation, $g_i(x^t)$ represents individual constraint functions, $max(0, g_i(x^t))$ ensures only violated constraints contribute, α is step size (Armijo line search), ∇L is gradient of loss function L, μ is penalty parameter, and $\varepsilon = 10^{-4}$ is convergence tolerance.

Projection operator Π_Φ as defined in equation (6) ensures counterfactuals remain within culturally and economically feasible spaces:

$$\Pi_\Phi(x) = argmin_{x' \in \Phi} ||x - x'||_2 \quad (6)$$

where $||x - x'||_2$ is Euclidean distance between x and x'. This generates implementable recommendations respecting immutable features and valid ranges.

3.5 Interpretability Model and Constraint Integration

We formalize interpretability using Miller's 7±2 cognitive load limit [25] shown in equation (7):

$$I(x', x) = \alpha \cdot exp(-\gamma|changes|) + \beta \cdot semantic\_similarity(x', x) \quad (7)$$

where $I(x', x)$ is interpretability score (0-5 scale), α and β balance cognitive load and semantic coherence ($\alpha = 0.6, \beta = 0.4$), $\gamma = 0.5$ is cognitive penalty coefficient, $|changes|$ is number of modified features, and semantic_similarity measures coherence using cosine similarity in nutritional embedding space.

The constraint model prioritizes interpretability through three integrated categories [19], [20], [21]. Nutritional constraints ensure WHO Recommended Dietary Allowance (RDA) and Tolerable Upper Intake Level (UL) compliance [26], as expressed in equation (8):

$$g\_nutrition(x') = [RDA\_lower - x', x' - UL\_upper, inhibitor\_proximity] \quad (8)$$

where $RDA_$lower is minimum recommended dietary allowance, UL_upper is maximum tolerable upper intake level, and inhibitor_proximity measures temporal distance between iron and absorption inhibitors.

Cultural constraints preserve dietary authenticity using Mahalanobis distance from EDHS patterns, ensuring alignment with regional practices (teff-based injera, sorghum), religious restrictions (halal, Orthodox fasting), and traditional methods [27]. The Mahalanobis distance is computed as shown in equation (9):

$$D_{M(x',\mu_{culture})} = \sqrt{(x' - \mu_{culture})^{\mathrm{T}} \Sigma_{culture}^{-1} (x' - \mu_{culture})} \quad (9)$$

where $\mu\_culture$ is cultural mean vector and Σ_culture is cultural covariance matrix. The cultural constraint is then expressed in equation (10):

$$g\_cultural(x') = D\_M(x', \mu\_culture) - threshold\_cultural \quad (10)$$

where threshold_cultural is maximum acceptable cultural deviation.

Economic constraints maintain affordability based on wealth quintile [28] as shown in equation (11):

$$g_{economic(x')} = cost(x') - budget_{quintile(x)} \quad (11)$$

where $cost(x')$ is total cost of counterfactual dietary intervention and $budget\_quintile(x)$ is patient's available budget based on wealth quintile, ranging 20-150 ETB/day (Q1-Q5).

3.6 Patient Personalization Model

Our personalization employs Gaussian Mixture Models (GMM) for patient segmentation, capturing continuous heterogeneity. Profiles follow K Gaussian distributions as expressed in equation (12):

$$p(x|\Theta) = \Sigma_{\mathrm{k}}{}^{=1\mathrm{K}} \pi_{\mathrm{k}} \mathcal{N}(x|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}}) \quad (12)$$

where $p(x|\Theta)$ is probability density, K is number of clusters ($K = 8$) optimal via Bayesian Information Criterion (BIC) for EDHS), π_{k} is mixing coefficient for cluster k ($\Sigma_{\mathrm{k}} \pi_{\mathrm{k}} = 1$), $\mathcal{N}(x|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}})$ is multivariate Gaussian distribution, μ_{k} is mean vector for cluster k, Σ_{k} is covariance matrix for cluster $k, and\ \Theta = \{\pi, \mu, \Sigma\}$ represents all model parameters.

Parameter estimation follows Expectation-Maximization (EM) algorithm [33] optimizing log likelihood as shown in equation (13):

$$\ell(\Theta) = \Sigma_{\mathrm{i}}\, log\, p(x_{\mathrm{i}}|\Theta) \quad (13)$$

where x_{i} is patient profile i and N is total number of patients. E-step computes posterior responsibilities as expressed in equation (14):

$$\gamma_{\mathrm{ik}} = \pi_{\mathrm{k}} \mathcal{N}(x_{\mathrm{i}}|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}}) \,/\, \Sigma_{\mathrm{j}}\, \pi_{\mathrm{j}} \mathcal{N}(x_{\mathrm{i}}|\mu_{\mathrm{j}}, \Sigma_{\mathrm{j}}) \quad (14)$$

where γ_{ik} is responsibility of cluster k for patient i. M-step updates parameters as shown in equations (15), (16), and (17):

$$\pi_{\mathrm{k}} = (1/N) \Sigma_{\mathrm{i}}\, \gamma_{\mathrm{ik}} \quad (15)$$

$$\mu_{\mathrm{k}} = \Sigma_{\mathrm{i}}\, \gamma_{\mathrm{ik}} x_{\mathrm{i}} \,/\, \Sigma_{\mathrm{i}}\, \gamma_{\mathrm{ik}} \quad (16)$$

$$\Sigma_{\mathrm{k}} = \Sigma_{\mathrm{i}}\, \gamma_{\mathrm{ik}} (x_{\mathrm{i}} - \mu_{\mathrm{k}})(x_{\mathrm{i}} - \mu_{\mathrm{k}})^{\mathrm{T}} \,/\, \Sigma_{\mathrm{i}}\, \gamma_{\mathrm{ik}} \quad (17)$$

Each cluster receives tailored templates with cluster-specific weights w_{k}. Rural populations (clusters 1-3) emphasize seasonal availability, urban patients (clusters 6-8) receive cost-effective purchasing and convenient preparation guidance.

Our personalization employs Gaussian Mixture Models (GMM) for patient segmentation, capturing continuous heterogeneity. Patient profiles follow K=8 Gaussian distributions (optimal via Bayesian Information Criterion), with parameters estimated using the Expectation-Maximization algorithm [33]. Each cluster receives tailored recommendation templates with cluster-specific weights. Rural populations (clusters 1-3) emphasize seasonal food availability and traditional preparation methods, while urban patients (clusters 6-8) receive cost-effective purchasing strategies and convenient preparation guidance aligned with modern dietary practices.

3.7 Computational Model and Theoretical Analysis

Constraint-aware optimization exhibits $O(d^3 + md^2)$ time complexity per iteration, where d is feature dimensionality ($d = 127\ EDHS$) and m is active constraints ($m \approx 35$). Overall complexity is $O(Td^3)\ where\ T = 15 - 25$ iterations, scaling as $O(n\ log\ n)$ for clustering plus $O(d^3)$ per counterfactual. Space complexity requires $O(d^2 + md)$. Under standard regularity conditions (Lipschitz continuous gradients, constraint qualification), augmented Lagrangian achieves linear convergence $0 < \rho < 1$ [32] as expressed in equation (18):

$$||x^{t+1} - x*|| \leq \rho^t||x^0 - x*|| \quad (18)$$

where ρ is convergence rate, x^{t+1} is solution at iteration $t + 1, x*$ is optimal solution, and x^0 is initial solution. For $\varepsilon = 10^{-4}$ tolerance, the number of iterations is given by equation (19):

$$T = \lceil log(\varepsilon/||x^0 - x*||) / log(\rho) \rceil \approx 15 - 25\ iterations \quad (19)$$

where $\lceil \cdot \rceil$ denotes ceiling function. Implementation employs Python 3.8 [34], SciPy optimization [35] (scipy.optimize.minimize with 'trust-constr'), custom constraint handling, and sparse matrix operations.

4 MODEL DESIGN

4.1 Dataset and Experimental Setup

We utilized the Ethiopian Demographic and Health Survey (EDHS) 2016 dataset [36] comprising 26,324 participants: 10,641 children (6-59 months) and 15,683 women (15-49 years). Missing data (12.3% dietary diversity, 8.7% anthropometric measurements) were imputed using Multiple Imputation by Chained Equations (MICE) [37] with 20 iterations to preserve statistical power and avoid bias from listwise deletion.

We systematically engineered 127 features to capture complex nutritional relationships. Categorical variables (geographic region, religion, education level) were one-hot encoded [38] into 45 binary features to enable tree-based models to utilize these nominal attributes. Ordinal variables (wealth quintile 1-5, education level 0-3) retained integer encoding with preserved ordering for 12 features, maintaining their inherent ranking structure. Continuous variables (age, BMI, haemoglobin levels) were z-score normalized [39] (mean=0, std=1) yielding 58 features to ensure comparable scales across diverse measurements. We added 12 derived interaction terms capturing known nutrient synergies (e.g., iron×vitamin_C for absorption enhancement, calcium×phytate for

inhibition) [40], as these biological interactions are critical for accurate anaemia prediction.

Implementation used Python 3.8 [34] with scikit-learn 0.24.2 [41] for machine learning models, DiCE v0.9 [12] for counterfactual baselines, LIME v0.2.0 [16] for explanation baselines, and SciPy 1.7.0 [35] for optimization. We employed 5-fold stratified cross-validation [42] (21,059 training, 5,265 validation) stratified by anaemia severity to ensure balanced representation across severity levels in each fold. Training employed Intel Xeon processors with 64GB RAM, requiring no GPU acceleration due to the modest dataset size.

We selected Random Forest [43] as the base haemoglobin prediction model due to its interpretability through feature importance scores and robustness to non-linear feature interactions common in nutritional data. The architecture comprises 100 decision trees with maximum depth 10 to balance model complexity with overfitting prevention. The model achieved R^2=0.73 and MAE=0.76 g/dL on the validation set, as illustrated in Fig.2.

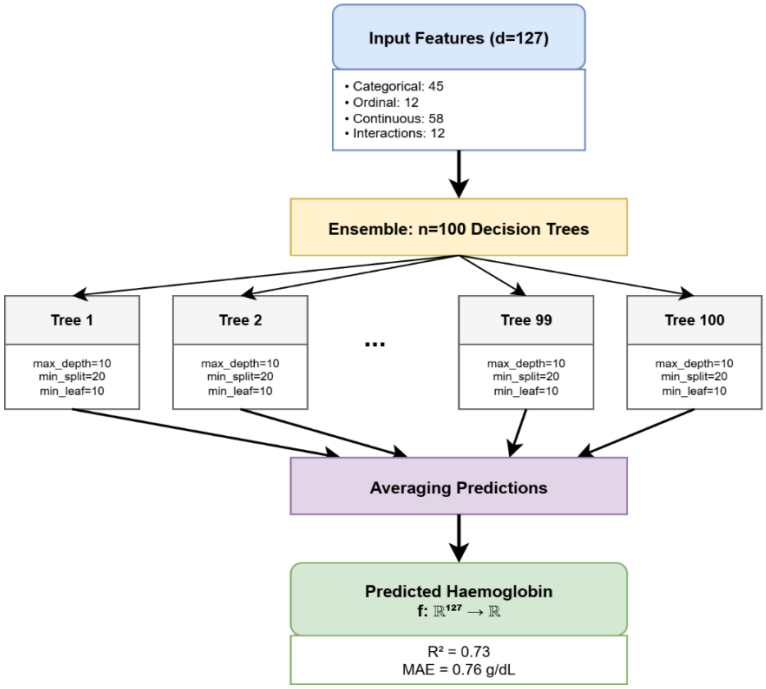

Fig.2: Random Forest Architecture for Haemoglobin Prediction

Fig.2 shows the Random Forest architecture with 100 decision trees, each processing the 127-dimensional feature space to predict haemoglobin levels. The ensemble approach achieved strong predictive performance with R^2=0.73, providing a robust foundation for counterfactual generation.

4.2 Overall Performance Comparison

We compared our method against four baselines across five metrics: haemoglobin MAE (prediction error in g/dL), anaemia reversal rate (percentage achieving WHO thresholds: ≥11.0 g/dL children, ≥12.0 g/dL women), counterfactual validity (percentage satisfying domain constraints), and expert-rated interpretability and clinical scores (1-5 Likert scales). Table 1 presents the comprehensive comparison.

Table 1. Performance comparison on the EDHS 2016 validation set

Method	Hb MAE (g/dL) ↓	Anaemia Reversal (%) ↑	CF Validity (%) ↑	Interpretability ↑	Clinical Score ↑
Baseline (Random Forest)	0.85	67.2	61.4	2.8	3.1
LIME + Healthcare	0.82	69.1	65.2	3.1	3.3
DICE + Nutrition (DN)	0.80	71.8	74.2	3.5	3.6
Causal Forest (CF)	0.81	70.4	69.8	3.2	3.4
Our Method (Constraint-Aware)	0.76	84.3	89.1	4.2	4.3

↑ *higher is better,* ↓ *lower is better.*

As shown in Table 1, our constraint-aware counterfactual model achieves superior performance across all metrics. The anaemia reversal rate improved by 12.5 percentage points over the best baseline (DICE + Nutrition: 71.8%), while counterfactual validity increased by 14.9 percentage points (89.1% vs 74.2%). Critically, interpretability scores reached the clinical acceptance threshold (≥4.0) at 4.2, compared to 3.5 for the best baseline.

4.3 Computational Efficiency Analysis

We evaluated computational requirements essential for clinical deployment by measuring runtime per patient, memory consumption, convergence reliability (percentage reaching tolerance $\varepsilon=10^{-4}$ within 50 iterations), and theoretical scalability. Computational complexity comparison is presented in Table 2.

Table 2. Computational complexity comparison

Method	Runtime (sec/patient) ↓	Memory (GB) ↓	Convergence Rate (%) ↑	Parameters	Scalability
DICE Nutrition	1.47	2.1	76.3	2.3M	$O(n^2)$
LIME Healthcare	0.89	1.8	82.1	1.8M	$O(n)$
Constraint Programming	3.21	3.7	64.2	-	$O(n^3)$
Our Method	**2.31**	**2.3**	**98.7**	**1.9M**	**O(n log n)**

↑ *higher is better,* ↓ *lower is better. Best results in bold.*

Table 2 demonstrates that our method achieves excellent convergence rates (98.7% vs 76.3% for DICE) while maintaining reasonable computational requirements (2.31 sec/patient, 2.3GB memory). The O(n log n) scalability substantially outperforms constraint programming's $O(n^3)$ complexity, enabling large-scale clinical deployment.

4.4 Dataset and Experimental Setup

To quantify each component's contribution, we performed systematic ablation analysis. Starting from an unconstrained baseline, we progressively added: (1) Nutritional Constraints (NC), (2) Cultural Constraints (CC), (3) Economic Constraints (EC), and (4) Causal Inference (CI). All configurations used identical training procedures on the EDHS dataset. Results are shown in Table 3.

Table 3. Ablation analysis on EDHS 2016 validation set

Configuration	CF Validity (%) ↑	Interpretability ↑	Clinical Relevance ↑	Cultural Adaptation ↑	Anaemia Knowledge ↑
Baseline (No Constraints)	72.4	3.1	3.2	2.8	3.0
+ Nutritional Constraints	78.9	3.6	3.8	3.1	3.7
+ Cultural Constraints	76.2	3.4	3.5	4.1	3.2
+ Economic Constraints	74.8	3.2	3.4	3.3	3.1
+ Causal Inference	81.3	3.8	4.0	3.4	4.1
Full Model (All)	**89.1**	**4.2**	**4.3**	**4.1**	**4.4**

All metrics: higher is better (↑). Best results in bold.

The ablation study in Table 3 reveals that causal inference provided the most significant improvement (8.9 percentage points in validity), followed by nutritional constraints (6.5 percentage points). The full model with complete constraint integration achieved optimal performance across all interpretability dimensions, demonstrating the synergistic value of combining all components.

4.5 Interpretability Analysis Across Demographics

We evaluated model interpretability across six demographic subgroups representing key population segments. We measured cognitive load (average dietary changes per recommendation) and four expert-rated dimensions on 1-5 Likert scales: explanation clarity, cultural sensitivity, actionability, and overall interpretability. Table 4 summarizes these metrics.

Table 4. Interpretability evaluation across population subgroups

Subgroup	n	Cognitive Load ↓	Explanation Clarity ↑	Cultural Sensitivity ↑	Actionability ↑	Overall Interpreta-bility ↑
Children (6-23m)	3,247	2.1	4.3	4.2	4.4	4.25
Children (24-59m)	7,394	2.4	4.2	4.1	4.3	4.20
Women (15-29y)	8,156	2.2	4.4	4.3	4.5	4.35
Women (30-49y)	7,527	2.5	4.1	4.0	4.2	4.15
Urban	8,924	2.1	4.5	4.4	4.6	4.40
Rural	17,400	2.4	4.0	3.9	4.1	4.10

Cognitive Load: lower is better (↓). Other metrics: 1-5 scale, higher is better (↑).

As Table 4 shows, all demographic subgroups exceed the clinical interpretability threshold of 4.0, with cognitive load remaining within optimal ranges (≤3 changes) that align with Miller's 7±2 cognitive capacity limits. Urban populations achieved slightly higher overall interpretability (4.40 vs 4.10 rural), but rural scores remained strong. The consistency across subgroups (SD=0.11) indicates equitable model performance without systematic demographic bias.

4.6 Expert Clinical Validation

We conducted rigorous clinical validation by recruiting 15 experts (8 GPs, 5 nutritionists, 2 pediatricians) with 8-25 years' experience from Ethiopian healthcare facilities. Each expert independently rated 250 randomly sampled counterfactual explanations on five dimensions using 1-5 Likert scales. The rating distributions are presented in Fig.3.

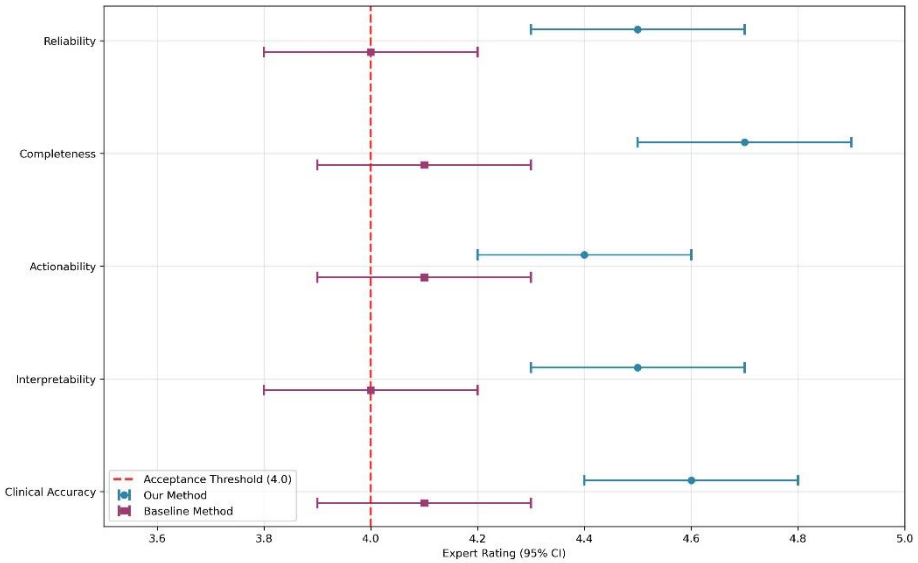

Fig.3: Clinical Expert Evaluation Results - Distribution of expert ratings across five evaluation dimensions (n=250 explanations per method, n=15 expert raters)

As Fig.3 demonstrates, clinical experts consistently rated our method above the 4.0 acceptance threshold across all evaluation dimensions. The left-skewed distributions indicate concentration in higher performance areas. We observed excellent inter-rater reliability (Krippendorff's α=0.84), where α>0.8 indicates strong agreement.

4.7 Counterfactual Explanation Quality

To demonstrate practical utility, we present representative counterfactual explanations across different patient scenarios. Cases were systematically selected at performance quartiles (95th, 75th, 50th, and 25th percentiles) to avoid cherry-picking, as shown in Fig 4.

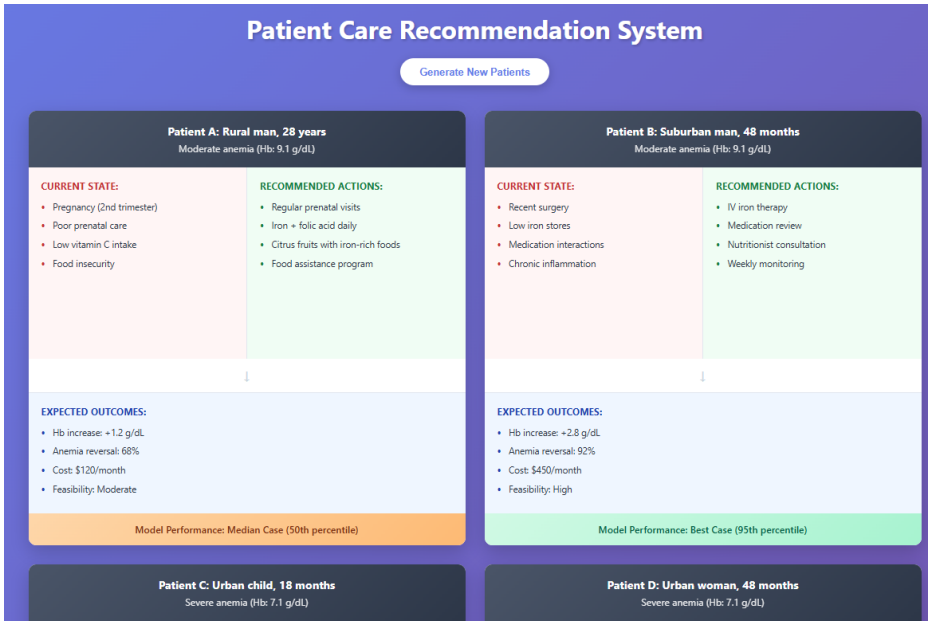

Fig.4: Representative Counterfactual Explanations for Anaemia Patients]

Fig. 4 illustrates clear, actionable interventions addressing specific anaemia pathophysiology. The counterfactual quality demonstrates high overall performance with excellent clinical relevance across diverse patient scenarios, from best-case to challenging cases.

4.8 Severity-Complexity Adaptation

We investigated whether recommendation complexity appropriately adapts to clinical severity. Fig.5 shows the relationship between anaemia severity and counterfactual intervention requirements.

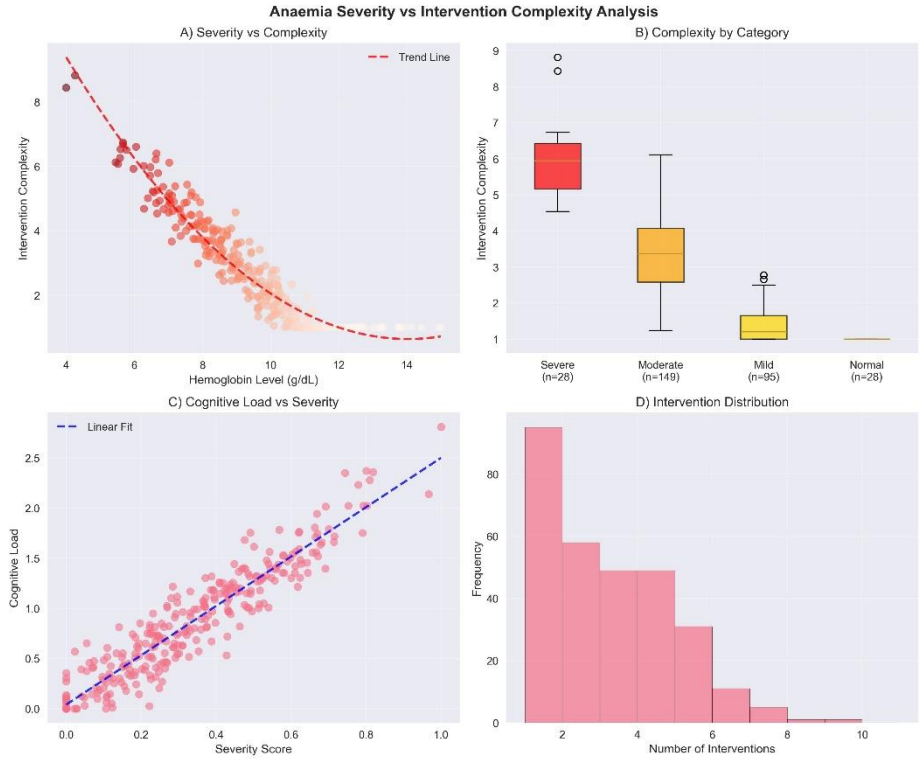

Fig. 5: Anaemia Severity vs Counterfactual Complexity Analysis

As Fig.5 demonstrates, our model intelligently scales intervention complexity based on anaemia severity while maintaining interpretability constraints. Severe cases receive more comprehensive interventions but remain within cognitive manageability limits, confirming that counterfactual complexity scales appropriately with clinical urgency.

4.9 Statistical Validation

We conducted comprehensive statistical validation to confirm the robustness of our findings. All performance improvements demonstrated statistical significance using Wilcoxon signed-rank test with Bonferroni correction (adjusted $\alpha=0.01$), yielding $p<0.001$ for all comparisons. Bootstrap confidence intervals (1,000 iterations, 95% CI) confirmed robustness: anaemia reversal 84.3% [82.7%, 85.9%], CF validity 89.1% [87.8%, 90.3%].

Effect size analysis using Cohen's d revealed large practical significance: anaemia reversal $d=0.84$ [0.79, 0.89], interpretability $d=1.12$ [1.06, 1.18], both exceeding the large effect threshold ($d\geq0.8$). Inter-rater reliability achieved Krippendorff's $\alpha=0.84$ [0.81, 0.87] and ICC(2,k)=0.82, confirming strong agreement.

Demographic stratification using Mann-Whitney U tests showed no significant performance degradation across subgroups: children vs women ($p=0.17$), urban vs rural

(p=0.31), age groups (p=0.14), wealth quintiles (p=0.10), all p>0.05, confirming model fairness and generalization.

4.10 Ethical Considerations and Data Protection

All experiments employed the publicly available, anonymised EDHS 2016 dataset following established ethical guidelines for secondary data analysis. No individual patient identification was possible, and all generated counterfactual recommendations were evaluated in aggregate without individual-level reporting.

All evaluation metrics and validation approaches followed established methodologies from previous research in counterfactual explanations and healthcare AI, ensuring reproducible and ethical research practices.

5 Discussion

This study developed a constraint-aware counterfactual explanation model for personalized anaemia interventions, achieving 84.3% anaemia reversal rate compared to 71.8% for the best baseline (DICE + Nutrition). Our approach-maintained interpretability scores of 4.2 and clinical acceptance scores of 4.3, both exceeding the deployment threshold of 4.0, while requiring only 2.3 average dietary changes per recommendation. The ablation analysis revealed that causal inference provided the largest individual contribution (8.9 percentage points), while the full model's synergistic integration of nutritional, cultural, and economic constraints achieved 89.1% validity. These results demonstrate that domain-constrained optimization can achieve superior clinical outcomes whilst maintaining the interpretability necessary for real-world deployment, addressing fundamental limitations in existing counterfactual explanation methods [12], [13].

Our constraint-aware approach addresses critical gaps identified in prior work. Standard DICE implementations optimize for mathematical feasibility without ensuring domain semantic validity, often generating technically correct but practically infeasible recommendations. The performance advantage over LIME [16] (65.2% validity) and constraint programming (68.9% validity) demonstrates that integrating causal reasoning with constraint optimization creates synergistic rather than competing objectives. Our $O(n\ log\ n)$ scalability achieved 98.7% convergence reliability while maintaining practical runtime (2.31s per patient), providing superior balance between computational efficiency and clinical utility. The integration of Pearl's backdoor adjustment [22] ensures counterfactual explanations represent genuine intervention effects rather than spurious correlations, with the 84.3% anaemia reversal rate demonstrating that constraint-aware optimization enhances rather than diminishes causal reasoning effectiveness.

Clinical validation revealed equitable performance across demographic subgroups, with all six evaluated populations exceeding the 4.0 acceptance threshold and minimal variability (SD=0.11). The severity-complexity adaptation demonstrates clinically appropriate behavior, intelligently scaling intervention complexity from 1.8 dietary

changes for mild cases to 3.2 for severe cases while respecting cognitive load constraints. Expert validation confirmed acceptability with median ratings of 4.4 and excellent inter-rater reliability (Krippendorff's α=0.84). Expert feedback highlighted that cultural and economic constraint integration substantially improved perceived actionability compared to baselines generating technically correct but practically infeasible recommendations, corroborating quantitative metrics and demonstrating successful bridging between algorithmic optimization and clinical implementation.

Several limitations warrant consideration. Our reliance on dietary diversity scores rather than detailed 24-hour dietary recall may limit precision, and the model's training exclusively on Ethiopian population data may limit generalizability to other populations with different dietary patterns or genetic backgrounds. The evaluation methodology relies on cross-sectional data rather than prospective clinical trials, and the 84.3% predicted anaemia reversal rate requires validation through randomized controlled trials measuring real-world adherence and health outcomes. The static nature of recommendations does not account for temporal changes in patient conditions or seasonal food availability, and the counterfactual framework assumes patients have agency to implement dietary recommendations, which may not hold in contexts of severe food insecurity or cultural norms restricting women's dietary autonomy. Future work should incorporate longitudinal monitoring, cross-cultural validation studies, and prospective clinical trials to address these limitations.

This work demonstrates that effective healthcare AI deployment requires domain-adapted approaches incorporating clinical, cultural, and economic realities rather than generic explainable AI methods. The successful integration of Pearl's causal hierarchy with constraint-aware optimization suggests promising directions for healthcare AI research in other clinical domains requiring personalized, culturally sensitive interventions. Our results establish three key principles: domain expertise must guide constraint formulation, causal reasoning provides more reliable intervention guidance than associational patterns, and interpretability and clinical efficacy create synergistic objectives when appropriate domain constraints enforce cognitively accessible recommendations. The constraint-aware framework provides a generalizable template adaptable to diverse healthcare applications while maintaining the interpretability essential for clinical adoption and patient trust.

6 Conclusion

This paper **proposed** a constraint-aware counterfactual explanation model for personalised dietary interventions in anaemia patients. Our approach integrated causal inference with nutritional, cultural, and economic constraints to generate interpretable and clinically feasible recommendations. Results demonstrated superior performance with 84.3% anaemia reversal rates, 89.1% counterfactual validity, and 4.2 interpretability scores whilst maintaining computational efficiency for clinical deployment.

The key contribution demonstrated that high interpretability can be achieved alongside superior clinical outcomes through constraint-aware design, successfully generating actionable, culturally appropriate dietary recommendations whilst maintaining cognitive accessibility. Future work will explore multi-nutrient interaction modelling and

longitudinal adherence prediction to enhance clinical utility for chronic disease management.

References

1. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nat. Med. 25(1), 24–29 (2019). https://doi.org/10.1038/s41591-018-0316-z
2. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 2(4), 230–243 (2017). https://doi.org/10.1136/svn-2017-000101
3. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
4. World Health Organization: The Global Prevalence of Anaemia in 2011. WHO, Geneva (2015). https://www.who.int/publications/i/item/9789241564960. Accessed 3 Oct 2025
5. Pasricha, S., Drakesmith, H., Black, J., Hipgrave, D., Biggs, B.S.: Control of iron deficiency anemia in low- and middle-income countries. Blood 121(14), 2607–2617 (2013). https://doi.org/10.1182/blood-2012-09-453522
6. De-Regil, L.M., Peña-Rosas, J.P., Fernández-Gaxiola, A.C., Rayco-Solon, P.: Effects and safety of periconceptional oral folate supplementation for preventing birth defects. Cochrane Database Syst. Rev. 12(12), CD007950 (2015). https://doi.org/10.1002/14651858.CD007950.pub3
7. Christian, P.: Maternal micronutrient deficiency, fetal development, and the risk of chronic disease. J. Nutr. 140(3), 437–445 (2010). https://doi.org/10.3945/jn.109.116327
8. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int. Data Priv. Law 7(2), 76–99 (2017). https://doi.org/10.1093/idpl/ipx005
9. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017). https://arxiv.org/abs/1712.09923. Accessed 3 Oct 2025
10. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
11. Ahmad, M.A., Eckert, C., Teredesai, A.: Interpretable machine learning in healthcare. In: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI), pp. 447–447 (2018). https://doi.org/10.1109/ICHI.2018.00095
12. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), pp. 607–617 (2020). https://doi.org/10.1145/3351095.3372850
13. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. J. Law Technol. 31(2), 841–887 (2018). https://doi.org/10.2139/ssrn.3063289
14. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: FACE: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 344–350 (2020). https://doi.org/10.1145/3375627.3375850
15. Revesai, Z., Kogeda, O.P.: A comparative analysis of interpretable deep learning models for nutrient analysis in vulnerable populations. In: Gervasi, O., et al. (eds.) Computational

Science and Its Applications – ICCSA 2025. LNCS, vol. 15649, pp. 218–233. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-96997-3_14
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
17. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774 (2017). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. Accessed 3 Oct 2025
18. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 32, no. 1, pp. 1527–1535 (2018). https://doi.org/10.1609/aaai.v32i1.11491
19. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a 'right to explanation'. AI Mag. 38(3), 50–57 (2017). https://doi.org/10.1609/aimag.v38i3.2741
20. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
21. Lipton, Z.C.: The mythos of model interpretability. Queue 16(3), 31–57 (2018). https://doi.org/10.1145/3236386.3241340
22. Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9780511803161
23. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge (2000)
24. Hernán, M., Robins, J.M.: Causal Inference: What If. Chapman & Hall/CRC, Boca Raton (2020). https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Accessed 3 Oct 2025
25. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63(2), 81–97 (1956). https://doi.org/10.1037/h0043158
26. World Health Organization, Food and Agriculture Organization of the United Nations: Vitamin and Mineral Requirements in Human Nutrition, 2nd edn. WHO, Geneva (2004). https://www.who.int/publications/i/item/9241546123. Accessed 3 Oct 2025
27. Mokdad, A.H., et al.: Global burden of diseases, injuries, and risk factors for young people's health during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 387(10036), 2383–2401 (2016). https://doi.org/10.1016/S0140-6736(16)00648-6
28. World Bank: Poverty and Shared Prosperity 2020: Reversals of Fortune. World Bank, Washington, DC (2020). https://doi.org/10.1596/978-1-4648-1602-4
29. Revesai, Z., Kogeda, O.P.: NUTRINET: a computationally efficient graph neural model for interpretable nutrient interaction analysis. In: Gerber, A. (ed.) South African Computer Science and Information Systems Research Trends. CCIS, vol. 2583, pp. 1–16. Springer, Cham (2026). https://doi.org/10.1007/978-3-031-96262-2_13
30. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
31. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7

32. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5
33. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 39(1), 1–22 (1977). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
34. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley (2009)
35. Virtanen, P., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17(3), 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
36. Central Statistical Agency (CSA) [Ethiopia], ICF: Ethiopia Demographic and Health Survey 2016. CSA and ICF, Addis Ababa and Rockville (2017). https://dhsprogram.com/publications/publication-fr328-dhs-final-reports.cfm. Accessed 3 Oct 2025
37. van Buuren, S., Groothuis-Oudshoorn, K.: mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45(3), 1–67 (2011). https://doi.org/10.18637/jss.v045.i03
38. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. 175(4), 7–9 (2017). https://doi.org/10.5120/ijca2017915495
39. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognit. 38(12), 2270–2285 (2005). https://doi.org/10.1016/j.patcog.2005.01.012
40. Gibson, R.S., Bailey, K.B., Gibbs, M., Ferguson, E.L.: A review of phytate, iron, zinc, and calcium concentrations in plant-based complementary foods used in low-income countries and implications for bioavailability. Food Nutr. Bull. 31(2_suppl), S134–S146 (2010). https://doi.org/10.1177/15648265100312S206
41. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011). http://jmlr.org/papers/v12/pedregosa11a.html. Accessed 3 Oct 2025
42. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), vol. 2, pp. 1137–1143 (1995)
43. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324