A Decentralized Federated Learning Framework for Privacy-Preserving Malnutrition Risk Prediction in Children

Shruthi S,
CSIR 4PI,
Bangalore,
Karnataka

M Dinesh Gokul Das,
CSIR 4PI,
Bangalore,
Karnataka

GopalKrishna Patra,
CSIR 4PI,
Bangalore,
Karnataka

Abstract—Malnutrition among children under five remains a critical public health challenge in low-income countries, including Ethiopia. The implementation of centralised machine learning methods faces difficulties because they violate patient privacy laws and existing health system limitations in resource-scarce environments. This paper presents a decentralised federated learning (DFL) framework using the ProxyFL gossip architecture, enabling collaborative malnutrition risk prediction across distributed nodes without any raw data sharing. Each node independently trains a Random Forest (RF) or XGBoost (XGB) classifier on a local partition of an Ethiopian pediatric dataset, exchanging only Fernet-encrypted feature importance vectors as proxy representations. The research examines the three-client federated system, which distributes its dataset among three separate health nodes that each contain approximately 534 training samples after removing the outliers. The study evaluates RF and XGBoost performance across all three client partitions by conducting a direct side-by-side assessment. The three clients show RF test AUC-ROC results of 0.721, 0.718 and 0.656, while XGBoost delivers results of 0.689, 0.674 and 0.539. The ProxyFL gossip protocol is validated as stable and algorithm-agnostic across all 20 communication rounds. SHAP analysis confirms BMI as the universally dominant predictor, with clinically coherent secondary features across all client partitions.

Keywords—Malnutrition, Federated Learning, ProxyFL, Gossip Learning, Privacy-Preserving ML, SHAP, XGBoost, Random Forest, Ethiopian Pediatric Data

I. Introduction

According to WHO anthropometric Z-scores, childhood malnutrition is linked to more than half of all under-five deaths worldwide, with compounding long-term effects on cognitive development and adult-onset disease susceptibility. Ethiopia routinely ranks among the countries in sub-Saharan Africa with the highest rates of childhood stunting and wasting. Although anthropometric and clinical feature machine learning models have shown excellent predictive performance, almost all current methods necessitate centralised data pooling, a paradigm incompatible with patient privacy laws, institutional autonomy, and the disjointed health infrastructure of low-resource settings.

Federated Learning (FL) makes it possible to train models while maintaining local raw data. However, Standard Federated Averaging (FedAvg) reintroduces centralisation at the gradient aggregation layer, posing a risk of single-point failure and governance vulnerabilities. Using Random Forest and XGBoost classification techniques, the study employs ProxyFL, an entirely decentralised DFL framework that uses gossiping to predict malnutrition. Fernet-encrypted feature importance vectors, which are abstract proxy representations devoid of any raw data or model weights, are used by the framework to achieve advanced structural protection. The most clinically realistic deployment scenario for a small network of district-level health facilities is the 3-client configuration, where each node contains roughly 534 training samples. Four primary contributions are presented in this paper are: (i) a comprehensive DFL framework for pediatric malnutrition prediction that can be deployed in low-resource, infrastructure-minimal environments; (ii) a direct per-client comparison of RF and XGBoost under 3-client federated partitioning; (iii) a detailed figure-level analysis of metrics dashboards, confusion matrices, SHAP explainability, and confidence calibration for each algorithm and client; and (iv) a quantitative head-to-head summary of algorithm robustness under federated conditions with limited data.

II. Background and Clinical Domain Motivation

A. Related Literature Studies

Federated Learning (FL) allows training models across many devices while keeping data on each device. McMahan et al. [1] created the FedAvg algorithm, which combines device updates through a central server and maintains performance with heterogeneous device data. The central system introduces difficulties that affect system expansion and operational dependability. According to Li et al. [2], the three primary challenges of FL networks involve excessive communication requirements, device-specific differences, and privacy threats that emerge from gradient sharing. Earlier, Shokri and Shmatikov [3] suggested sharing only some parameters to protect privacy in collaborative learning, but their method still reveals some model details. Lundberg et al. [4] use explainable AI techniques like SHAP for understanding the model behaviour by identifying the most impactful features for prediction.

Current research conducted between 2022 and 2025 demonstrates that gradient-based FL systems are increasingly encountering additional technical difficulties. Kairouz et al. [5] identify two existing problems, privacy risks and difficulties with scaling solutions. Karimireddy et al. [6] present SCAFFOLD, which enhances learning capabilities through data changes while maintaining shared gradient information. Differential privacy techniques help to protect data from unauthorised access, but they decrease the effectiveness of machine learning models. The new gossip-based FL techniques developed by Lalitha [7] and his team require no central servers for operation, but they still need to exchange model parameters. Nguyen et al. [8] developed a system that enables different models to operate together by sharing the output while the communication-saving methods developed by Li et al. [9] and his team in 2024 decrease bandwidth requirements, yet they do not protect against privacy threats.

Overall, current methods cannot handle decentralisation, privacy, and communication efficiency simultaneously. The new ProxyFL framework solves these problems by swapping gradient sharing for encrypted feature-based proxies. The important research gaps identified are:1. The majority of federated learning (FL) systems depend on central servers to operate, while FedAvg functions as their main server, which creates problems for their ability to scale up and use their services in healthcare environments that use distributed systems. 2. Privacy Leakage via Gradients: Current methods require sharing gradients or parameters, which allow attackers to guess or rebuild private data. 3. Lack of Model Behaviour Analysis under Data Fragmentation: Previous research dedicated its efforts to studying optimisation efficiency while providing only a narrow view of how machine learning models function in environments with multiple data sources and limited available data.

Unlike conventional federated learning systems that rely on iterative parameter aggregation, the proposed ProxyFL framework focuses on privacy-preserving representation exchange through encrypted feature-importance proxies. The objective of this study is not to optimise a shared global model through gradient fusion, but rather to investigate whether decentralised peer-aware learning can maintain clinically meaningful predictive performance without exposing model parameters, gradients, or raw patient data

B. WHO Standards

The WHO and UNICEF recommend three primary anthropometric indices [10] for child nutritional status assessment. The Weight-for-Age Z-score (WAZ) measurement system identifies two types of undernutrition, which include both short-term and long-term conditions. The Height-for-Age Z-score (HAZ) measurement system assesses permanent nutritional deprivation, which results in stunting. The Weight-for-Height Z-score (WHZ) measurement system detects acute weight reduction, which results in wasting. A Z-score calculation uses the formula $Z = (X - M) / SD$ where X represents the observed measurement and M describes the WHO reference median, and SD functions as the reference standard deviation. The threshold of $Z \leq -2$ establishes a universal standard for malnutrition detection, while $Z \leq -3$ establishes a requirement for immediate medical intervention because of severe health conditions.

C. Feature Design and Clinical Alignment

The Ethiopian dataset includes age, weight and height measurements together with information about medically significant comorbidities, which include anaemia, diarrhoea, malaria and tuberculosis and the biological pathways that lead to malnutrition. The condition of anaemia demonstrates both micronutrient deficiency and continuous undernutrition. Infectious diseases cause the body to reduce nutrient absorption while they increase metabolic requirements and interfere with immune system processes. The BMI feature which the system developed through weight/height² calculation, generates a single strong combined predictive model that unites the weight-height relationship that supports WHZ-based wasting assessment. Tree-based ensemble methods function effectively with these domain-aligned features because they can capture non-linear relationships without needing Z-score calculations during inference.

III. Dataset and Problem Formulation

Dataset

The study uses a publicly available Ethiopian pediatric malnutrition dataset [11], which contains 4,098 records and 14 features that were created through BMI engineering. The study used age in months and gender, height and weight, BMI, stunting condition, underweight status, overweight status, anaemia, malaria, diarrhoea, tuberculosis and nutritional status category and a binary malnutrition risk target label. The preprocessing process of the data includes two steps: first, which involves label encoding of categorical variables; second, which uses RobustScaler for normalisation to reduce outlier effects found in pediatric anthropometric data.

Based on model prediction behaviour, an outlier filtering procedure was carried out as a first step to enhance dataset quality. Stratified cross-validation was used to train a Support Vector Machine (SVM) classifier as part of an initial training setup for the dataset. This stage's confusion matrix results were examined to find samples that produced false positive (FP) and false negative (FN) predictions. Measurement variability, overlapping feature distributions, or inconsistent recorded clinical attributes can all lead to ambiguous or noisy data points in which the feature representation does not clearly match the ground-truth label.

The training pipeline extracted the indices of instances that were consistently classified correctly as true positives (TP) and true negatives (TN) across the cross-validation folds to retain only trustworthy samples. Stable prediction results were achieved through hyperparameter tuning which used grid search and 3-fold stratified cross-validation before the data extraction process. A refined dataset with only

observations that were confidently classified was produced by eliminating samples that produced FP or FN predictions. By eliminating confusing areas in the feature space and lessening the impact of noisy or contradictory samples, this filtering step makes it possible for the learning algorithms used in the federated learning experiments to make decisions with greater precision. Although Random Forest [12] and XGBoost [13] demonstrate the best predictive performance, SVM was chosen for the preprocessing stage because of its effectiveness in learning clear margin-based decision boundaries, which helps in identifying ambiguous samples responsible for false positive and false negative predictions (Table I).

Federated Partitioning of Data for 3-Client Configuration

The 3-client federated simulation requires a preprocessed dataset that is divided among three client nodes using the round-robin method, ensuring equal class distribution across nodes. About 428 training, validation, and test samples are sent to each client. The system ensures complete operational separation between all nodes during every processing stage. The study investigates how inter-client performance differences emerge from remaining variations in local class-conditional feature distributions, which exist in every real-world sample distribution that reaches this size, although the round-robin method achieves a near-equivalent distribution.

Problem Definition

The task is binary malnutrition risk classification: given a child's demographic, anthropometric, and clinical profile, predict whether the child meets the WHO Z-score-based criterion for malnutrition risk. The federated constraint prohibits any raw data exchange between nodes; each client trains and evaluates exclusively on its local partition throughout the experiment.

IV. Methodology

A. Decentralised Federated Learning Formulation

The complete dataset is represented by the union of all client health node data partitions D, which consists of private data held by each client k. Each node implements malnutrition risk prediction through a local empirical risk minimisation framework, which uses the equation $\theta_k^* = \arg\min \ell(f(\cdot; \theta_k), D_k)$, where f is the classifier used to determine optimal model parameters. The clients conduct their individual local optimisation processes while maintaining complete confidentiality of their data, model parameters and learning process gradients.

B. ProxyFL: Gossip-Based Communication

ProxyFL uses its gossip communication protocol to eliminate the need for centralised parameter aggregation, which requires all connections to use centralised aggregation. Each client provides a FastAPI REST endpoint that delivers its current feature importance vector as the proxy representation that shows learned nutritional risk patterns while hiding both original data and internal model details. The system maintains complete confidentiality because it never sends model weights, decision tree structures or gradient data throughout the entire process. A client at each of 20 communication rounds chooses one peer from its available neighbourhood and uses HTTP GET to acquire that peer's feature importance vector which it compares against its own vector through cosine similarity calculation to check distributional convergence. The system uses the gossip mechanism to monitor convergence progress.

The existing system currently forbids direct parameter fusion between different nodes, but still allows feature-level learning information to be shared through proxy representations, which maintain complete separation of local model parameters. The system design decision establishes a clear primary goal which focuses on protecting user privacy while making it easy to implement security measures in healthcare environments with limited technological resources.

C. Algorithm and Implementations

Random Forest is implemented using scikit-learn's RandomForestClassifier with 500 trees, random feature subspace selection at each split. Ensemble averaging (bagging) across 500 independently grown trees provides a strong variance-reduction mechanism that is structurally robust under limited per-client data. RF training loss remains near-zero and stable throughout all 20 gossip epochs.

XGBoost is implemented using the XGBoost library with L1/L2 regularisation and sequential gradient boosting. Each successive tree corrects the residual errors of all predecessors, enabling strong asymptotic performance on large datasets. However, at approximately 910 samples per client, this mechanism progressively specialises to training-set noise: training loss collapses to near-zero within 2 to 3 epochs while validation and test losses rise monotonically across all 20 gossip epochs.

The proposed framework was implemented in Python using a modular architecture to simulate decentralised client nodes and their communication. Machine learning models were implemented using the scikit-learn and XGBoost libraries, while FastAPI was used to expose REST endpoints that enable peer-to-peer communication between clients. Each node independently performs local model training on its private dataset partition and periodically exchanges feature importance vectors with neighbouring nodes through HTTP requests as part of the ProxyFL gossip protocol. The system implements resource monitoring, model evaluation, and dataset preprocessing procedures, which allow the experimental setup to simulate actual decentralised learning environments while maintaining complete data separation between clients. The system establishes user connections through feature importance proxies, which enable ProxyFL to address three major challenges that face healthcare applications: the system eliminates requirements for consistent central server connections, it secures data protection by eliminating central data repositories, and it prevents attackers from accessing training data through gradient attacks. The framework demonstrates compatibility with all algorithms according to test results, which show that both RF and XGBoost function with the gossip protocol without requiring any modifications.

D. Client Configurations

The decentralised learning environment is modelled with several client nodes, which serve as training participants who complete their training separately. This study employs 1-client, 2-client, and 3-client configurations to assess the distributed learning effect on model performance. The dataset distribution across clients uses a round-robin distribution strategy, which guarantees

that all samples are distributed equally among all participating nodes. The method produces an even distribution of data to every client while keeping their local datasets completely distinct from one another. The distribution is assumed to be approximately IID, as each client receives a representative subset of the overall dataset. Researchers selected the approximately IID partitioning strategy because they wanted to study the impacts of decentralised communication, together with restricted sample sizes that each participant could access. More complex non-IID partitioning scenarios representing heterogeneous clinical populations can be investigated in future work.

Each client trains an identical Random Forest model using only its locally available data. The training process starts at every node and continues until model updates get shared between clients through the decentralised communication system. This system allows multiple parties to learn together while they maintain their original data confidentiality.

The study investigates how model performance decreases when client data reduction occurs through different client number experiments. The experimental setup also helps in evaluating the trade-off between data distribution and generalisation in a decentralised federated learning environment.

Each client in the Random Forest model setup creates a local ensemble which includes 500 decision trees that reach a maximum depth of 20 while using Gini impurity as their default split criterion. The training process operates through 20 local epochs which occur during each communication round and the decentralised learning process takes place across 15 communication rounds. Clients share their locally developed model parameters during each round to support their shared learning process. The chosen hyperparameters establish an equilibrium between model complexity and computational efficiency which enables consistent system performance throughout the various distributed nodes.

The XGBoost model requires each client to build a gradient boosting ensemble through 300 boosting rounds at a learning rate of 0.1 and a maximum tree depth of 6. The default values of hyperparameters become active because the system requires protection against overfitting. The system executes local training for 20 epochs during each communication round while conducting decentralised training across 15 communication rounds. The model updates become available to all clients after each round, which enables the model to improve by acquiring knowledge from other nodes. The hyperparameters create a trade-off between two different aspects of distributed learning, which affects the system's ability to generalise.

The proposed framework prioritises structural privacy protection and decentralised system strength over achieving fast worldwide model convergence. The proxy representations, which have clinical importance, enable healthcare facilities with restricted resources to transfer data while safeguarding their gradient information and model weight data. The system functions as an effective replacement for existing federated optimisation methods.

E. Experimental Setup

The experiments evaluated a decentralised federated learning framework across varying numbers of clients. The researchers processed training data through preprocessing steps before they created train and test sets which they distributed to clients using a round-robin partitioning method that produced approximately equal and independent data shares for each node.

V. RESULTS

The evaluation of the decentralised federated learning framework's performance test uses a configuration with three clients. The system operates with three nodes, which maintain separate datasets, while each client trains a local model using ProxyFL gossip communication. The experiments compare the performance of Random Forest (RF) and XGBoost (XGB) classifiers trained on identical partitions of the processed dataset. The evaluation of the model uses AUC-ROC, F1 score, Balanced Accuracy, and AUPRC, which together provide an all-inclusive evaluation of classification results in situations where classes are not equally distributed. Using confusion matrices, confidence calibration plots, and SHAP-based feature analysis across several clients, the study investigates prediction behavior and model interpretability.

TABLE I. PERFORMANCE METRICS COMPARISON OF DIFFERENT ALGORITHMS

Algorithm	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.5857	0.5375	0.5931	0.5639	0.6203
AdaBoost	0.6293	0.6000	0.5379	0.5673	0.6772
LightGBM	0.6667	0.6234	0.6621	0.6421	0.7391
CatBoost	0.6791	0.6312	0.6966	0.6623	0.7538
XGBoost	0.7040	0.6736	0.6690	0.6713	0.7577
SVM	0.7321	0.6954	0.7241	0.7095	**0.7939**
Random Forest	**0.7508**	**0.7481**	0.6759	**0.7101**	0.7751

[a] Bold values indicate the best performance for each metric.

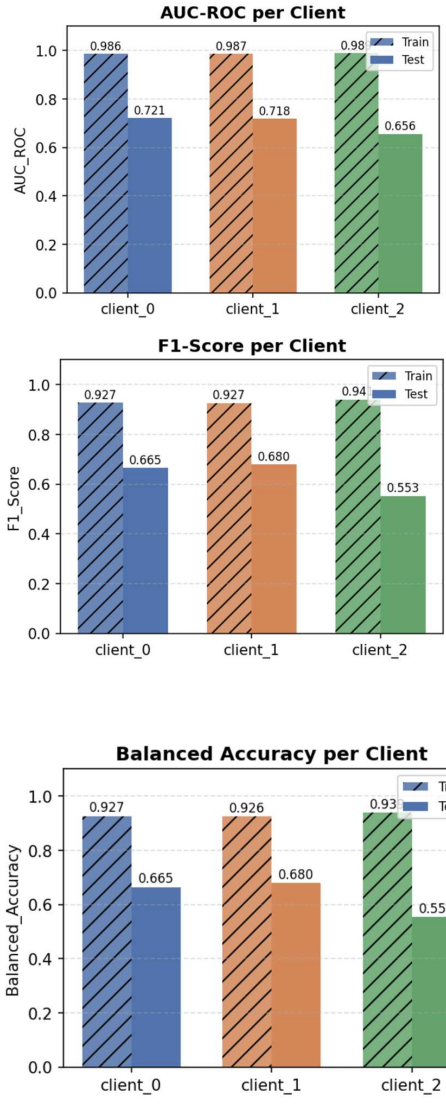
Table I displays the results, which show how seven machine learning algorithms performed when they were tested on the malnutrition prediction dataset. The two highest performing models across different assessment criteria are Random Forest and XGBoost. Random Forest displays its best results through its capacity to detect both positive and negative outcomes, which leads to its highest accuracy of 0.7508, precision of 0.7481 and F1 score of 0.7101. XGBoost provides strong predictive ability because it achieves 0.7040 accuracy, 0.6713 F1 score and 0.7577 ROC AUC value, which exceeds the performance of all boosting techniques, including LightGBM and traditional models like AdaBoost and Logistic Regression. Random Forest and XGBoost provide dependable results across all assessment criteria because they successfully handle complex feature relationships in the dataset, while

SVM achieves slightly better ROC AUC results. The results from this research support the selection of Random Forest and XGBoost as the main algorithms that will be used in future federated learning research. The baseline experiments demonstrate that Random Forest and XGBoost stand as the two most effective tree-based ensemble methods, which will be tested in decentralised environments. This testing ensures that any observed reductions in federated performance result from distribution learning constraints instead of having subpar base classifiers.

Random Forest — 3-Client Results

The RF 3-client configuration trains an independent Random Forest classifier on approximately 428 local training samples at each of three nodes, with ProxyFL gossip operating across all 20 communication epochs.

Performance Metrics



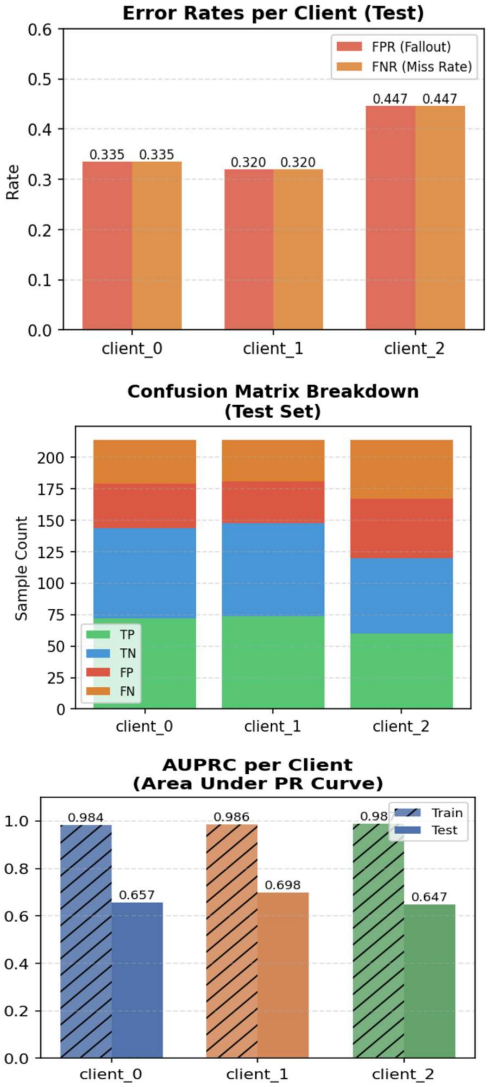


Figure 1: Comprehensive performance evaluation of a federated learning setup using a Random Forest model across three clients, illustrating training vs. testing metrics (AUC-ROC, F1-score, Balanced Accuracy, and AUPRC), overfitting gaps, error rates, and confusion matrix distributions, highlighting variability in generalisation and model performance across decentralised data sources.

For all three clients' training, AUC was observed above 0.986, confirming that the RF ensemble fits its local 428-sample data partition effectively across all nodes. The divergence occurs entirely at the test level. Client 0 achieves a test AUC of 0.721, an F1 score of 0.665, and an AUPRC of 0.657. Client 1 achieves a test AUC 0.718 with the best F1 across all three clients at 0.680 and AUPRC 0.698—a slightly better precision-recall tradeoff than Client 0 despite marginally lower AUC, indicating that Client 1's partition composition favours higher F1 at its operating threshold. Client 2's performance metrics: test AUC 0.656, F1 0.553, AUPRC 0.647, and FPR=FNR=0.447.

Confusion Matrices

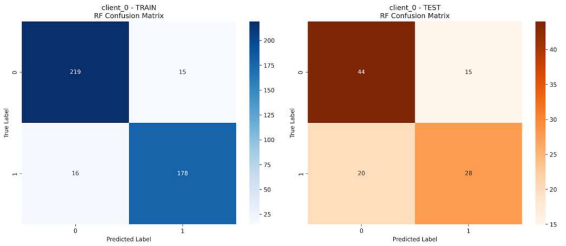


Fig. 2a: Client 0 RF CM. Train: 219 TN, 178 TP, 15 FP, 16 FN. Test: 44 TN, 28 TP, 15 FP, 20 FN.

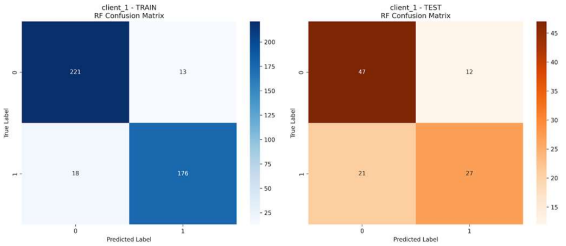


Fig. 2b: Client 1 RF CM. Train: 221 TN, 176 TP, 13 FP, 18 FN. Test: 47 TN, 27 TP, 12 FP, 21 FN.

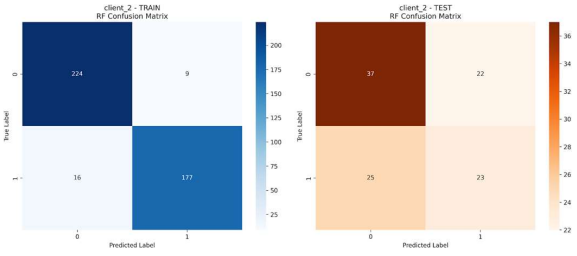


Fig. 2c: Client 2 RF CM. Train: 224 TN, 177 TP, 9 FP, 16 FN. Test: 37 TN, 23 TP, 22 FP, 25 FN.

The three clients show identical training performance with TN ranging 219–224 and TP ranging 176–178 and 9–15 FP and 16–18 FN, to show that the RF ensemble achieves identical performance across all local partitions, while the inter-client test divergence matches generalisation effects that produce different results from actual differences in local training performance. The test set results show that Clients 0 and 1 display the expected clinical pattern. Client 2 displays the most critical clinical pattern because test FP counts reach 22 while true negative counts reach 37, which results in approximately 37 per cent of healthy children being wrongly identified as at-risk, while false negative counts of 25 exceed true positive counts of 23, which results in more at-risk children being missed than those who have been correctly identified. The decline in discrimination performance across both class directions proves that 910 samples from Client 2's dedicated partition do not provide enough data to support reliable RF boundary learning in this feature space.

Confidence Calibration

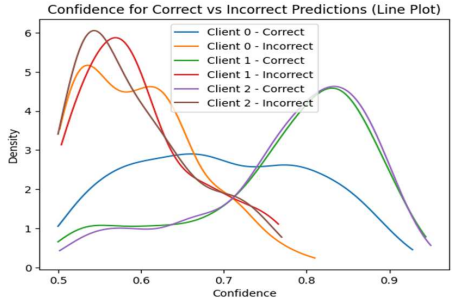


Fig. 3a: Confidence distribution comparison for correct vs. incorrect predictions across three federated clients, highlighting calibration behaviour and the separation between well-classified and misclassified samples.

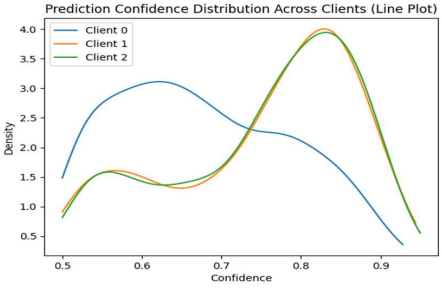


Fig. 3b: Overall prediction confidence distribution across clients, illustrating variations in model certainty and consistency in confidence estimation across decentralised data sources.

The study presents its most important clinical result through the confidence calibration analysis. Fig. 3a shows that correct and incorrect predictions largely occupy non-overlapping confidence bands across all three RF clients: correct predictions peak in the 0.75–0.95 range, while misclassifications concentrate between 0.50 and 0.65, with negligible overlap above approximately 0.70. Although the two groups' performance levels vary, their RF confidence results exhibit a similar pattern, with accurate predictions at the three client scales, so the distance between them remains constant across Clients 0, 1, and 2. Plotting the distribution (Fig. 3b) shows structural differences between clients: Client 0's wider bimodal profile reflects more uncertainty on a greater fraction of its test cases, consistent with its partition presenting more ambiguous feature compositions near the decision boundary, whereas Clients 1 and 2 assign high confidence sharply (peaks at ~0.83). This calibration feature allows organisations to implement multiple clinical deployment levels without requiring a separate calibration model for their system. The system uses predictions above 0.75 confidence to start nutritional intervention, while it sends 0.50 to 0.70 predictions for clinical Z-score assessment, and it monitors high-confidence negative predictions through standard procedures.

A. XGBoost — 3-Client Results

The XGBoost 3-client configuration trains an independent XGBoost classifier on each of three nodes, using the same ~428-sample partitions as in the RF experiments, under the same ProxyFL gossip protocol.

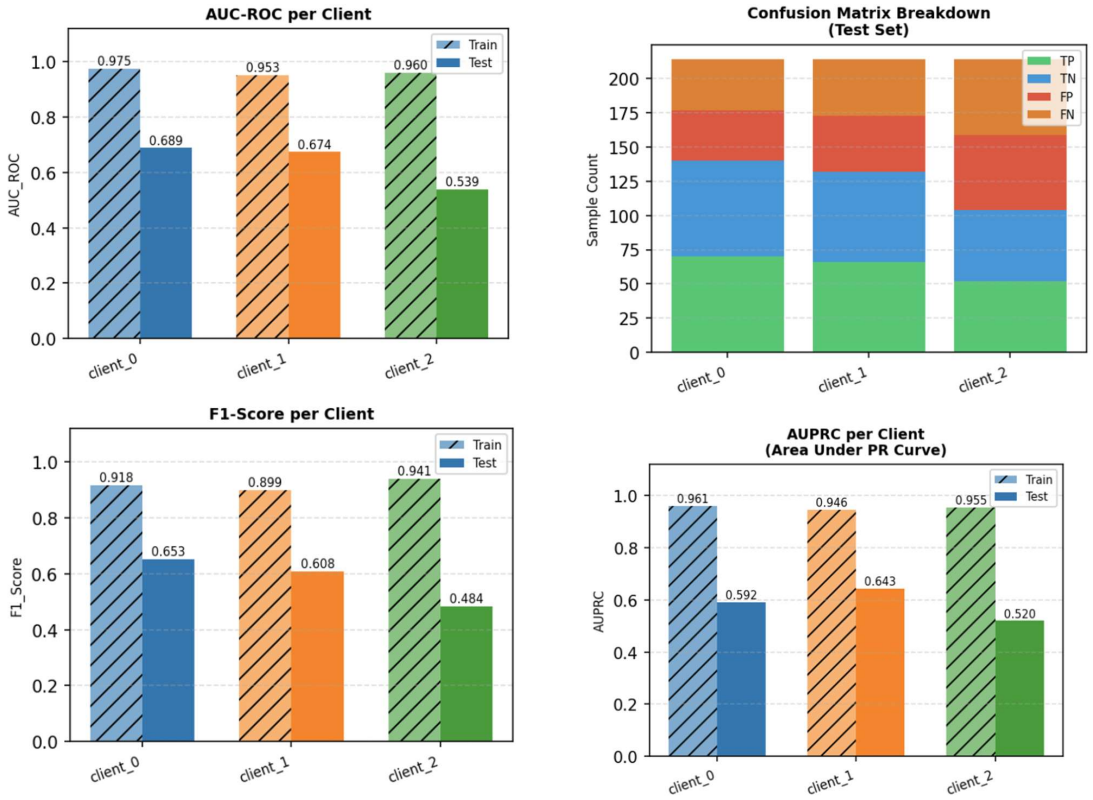


Figure 4: Comprehensive performance evaluation of a federated learning setup using an XGBoost model across three clients, illustrating training vs. testing metrics (AUC-ROC, F1-score, Balanced Accuracy, and AUPRC), overfitting gaps, error rates, and confusion matrix distributions, highlighting variability in generalisation and model performance across decentralised data sources.

B. Confusion Matrices and SHAP Analysis

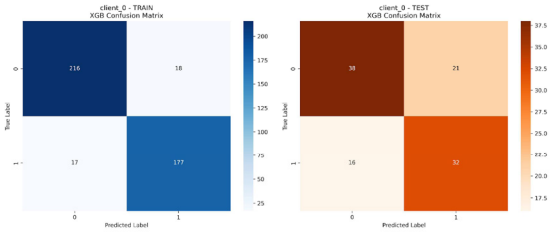


Fig. 5a: Client 0 XGB CM results

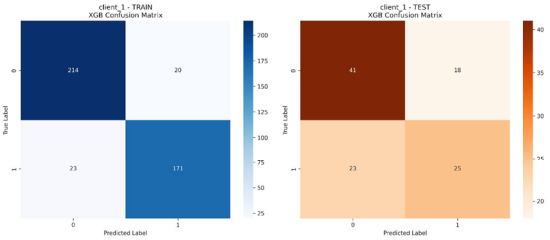


Fig. 5b: Client 1 XGB CM results

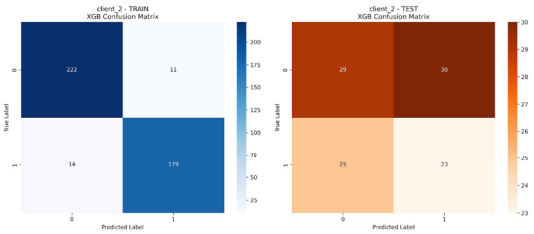


Fig. 5c: Client 2 XGB CM results.

The test AUCs for clients 0 and 1 are 0.689 and 0.674, which remain above the practical discrimination floor but show larger gaps between their training and testing results than their RF counterparts at the same partitions. Client 2 reaches a test AUC of 0.539, with a train-test gap of Δ0.421—the worst result across all experiments in this study. The test AUC of 0.539 is statistically similar to random classification, which has an AUC of 0.500. The F1 score of 0.484, together with the Balanced Accuracy score of 0.485 and an FPR that results in FNR equal to 0.515, confirms that complete generalisation failure has occurred. Client 2 maintains a training AUC of 0.960, indicating that the local XGBoost model correctly learned the training partition, as the system failure begins with its generalisation capabilities.

The confusion matrices demonstrate how different performance levels lead to specific clinical results. Client 0 achieves 38 TN, 32 TP with 21 FP and 16 FN: FP (21) outnumbers FN (16), which shows how XGBoost assigns positive labels when it cannot determine outcomes, but the model correctly identifies at-risk children because it detects 32 cases while missing 16 cases, which results in clinically valid positive-class recall. Client 1 shows a broadly similar pattern with slightly elevated errors, which matches its lower performance metrics.

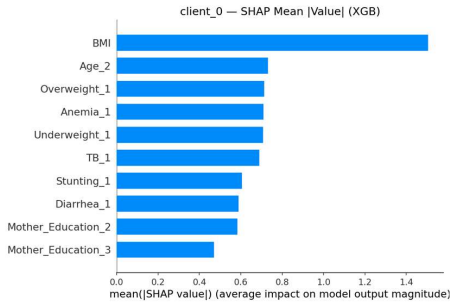


Fig. 6a: SHAP mean absolute value plot for Client 0 in the federated XGBoost model (3-client setup), highlighting the relative importance of features based on their average impact on model predictions, with higher values indicating stronger influence on the output.

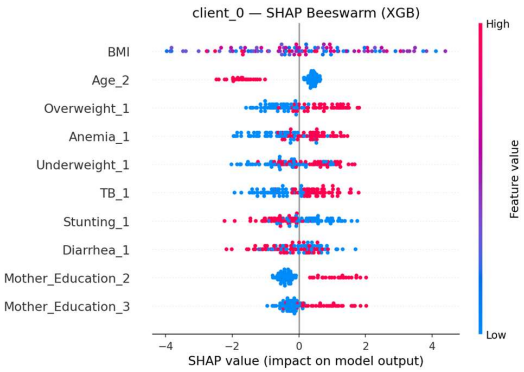


Fig. 6b: SHAP-based feature importance analysis for Client 0 in the federated XGBoost model (3-client setup), illustrating the impact and distribution of key features on model predictions, with colour-coded contributions indicating the direction and magnitude of influence on the output.

The first class boundary fails because FN (25) exceeds TP (23) which shows that more at-risk children are missed than correct identification of at-risk children and FP (30) exceeds TN (29) which shows that more healthy children receive incorrect alerts than proper identification of healthy children. Client 2 displays both FP > TN and FN > TP through its system because the model lacks any recognized trustworthy methods for distinguishing between its data groups. The method, which uses decision thresholds, creates an essential distinction because it produces one type of error improvement while creating another type of error increase. The SHAP mean absolute value plots show two distinct features because the algorithms produce identical results for all clients yet their partitioning methods generate different results. BMI stands as the primary predictor across both clients because it proves BMI functions as a fundamental anthropometric measure that maintains its effectiveness throughout various data grouping methods. The secondary rankings produce different results, which show significant differences between them. Client 2 shows a 0.94 level of anaemia that reaches its highest point because it shows an upward trend which results in the condition becoming the second most common. The Client 2 partition shows an increased incidence of anaemia because scientific evidence shows that iron deficiency anaemia and anthropometric wasting have a reciprocal clinical relationship in the Ethiopian pediatric population. The body condition of individuals with anaemia deteriorates because they lose their appetite which results in them missing vital nutrients needed for growth.

VI. RF vs. XGBoost: Direct 3-Client Comparison

The section contains a direct algorithm comparison between two algorithms that evaluate three different client partitions through identical data splits used in all test cases. The summary of the experiments is provided below.

A. Quantitative Summary

TABLE II. TABLE II. Full Performance Summary — 3-Client Configuration

Client	Tr. AUC	Te. AUC	ΔAUC	Te. F1	Bal. Acc	FPR	FNR	AUPRC
RF _ C0	0.986	0.721	0.266	0.665	0.665	0.335	0.335	0.657
RF — C1	0.987	0.718	0.269	0.680	0.680	0.320	0.320	0.698
RF — C2	0.989	0.656	0.333	0.553	0.553	0.447	0.447	0.647
XGB — C0	0.975	0.689	0.286	0.653	0.655	0.345	0.345	0.592
XGB — C1	0.953	0.674	0.278	0.608	0.608	0.392	0.392	0.643
XGB — C2	0.960	0.539	0.421	0.484	0.485	0.515	0.515	0.520

b. Tr.=Train, Te.=Test, ΔAUC=Train−Test AUC gap. C0/C1/C2 = Client 0/1/2.

TABLE III. TABLE III. RF vs. XGBoost Head-to-Head per Client

Client	RF Te.AUC	XGB Te.AUC	RF Te.F1	XGB Te.F1	RF ΔAUC	XGB ΔAUC
Client 0	0.721	0.689	0.665	0.653	0.266	0.286
Client 1	0.718	0.674	0.680	0.608	0.269	0.278
Client 2	0.656	0.539	0.553	0.484	0.333	0.421

B. Client 0: Competitive Performance, Divergent Error Profiles

At Client 0, RF and XGBoost produce their closest results of the three partitions. RF achieves a test AUC of 0.721 vs. XGB 0.689, a gap of 0.032. RF F1 (0.665) exceeds XGB F1 (0.653) by 0.012. The train-test gap for RF is Δ0.266 vs. Δ0.286 for XGB—a modest but consistent RF advantage. Despite similar aggregate metrics, the confusion matrices expose a meaningful qualitative difference: RF Client 0 shows FN (20) > FP (15), while XGB Client 0 shows FP (21) > FN (16). RF thus misses slightly more at-risk children but avoids false alarms more effectively, whereas XGB over-refers slightly more healthy children. For a population-level screening tool, the XGB pattern—favouring false positives over false negatives—is the clinically preferable error mode, as over-referral incurs assessment cost while under-detection incurs health cost. Neither model is dominant on clinical grounds alone at this partition.

C. Client 1: F1 Divergence and Ranking Reversal

Client 1 produces the most interesting cross-algorithm asymmetry. The RF system obtained a test AUC score of 0.718, which led to its highest F1 score of 0.680 for this particular client. The XGB system achieved a test AUC of 0.674, yet produced an F1 score of 0.608, which creates an F1 gap of 0.072 that appears much larger than the 0.044 AUC gap. The results show that XGB Client 1 obtains its AUC through precise predictions based on a small number of unambiguous situations, but fails to handle the F1-relevant cases which have uncertain outcomes. The RF Client 1 system achieved a total of 12 false positive errors, which represents the lowest value among all clients in this research study, because its system partitioning contained more definite negative instances, which led to RF system voting processes to reject positive instances. The AUPRC analysis shows that RF Client 1 achieves 0.698 while XGB reaches 0.643, which establishes a 0.055 difference that shows RF maintains better precision throughout all recall levels in this section.

D. Client 2: Catastrophic Divergence

Client 2 is where the algorithms fundamentally diverge. RF achieves test AUC 0.656 and F1 0.553—weak but operationally functional. XGB reaches test AUC 0.539 and F1 0.484—complete generalisation failure. XGB achieves a test AUC of 0.539 and an F1 score of 0.484, which demonstrates an inability to generalise. The AUC gap between the algorithms at Client 2 is 0.117, which is more than three times the gaps at Clients 0 and 1, which are 0.032 and 0.044, respectively. The gap between actual and expected results shows a strong contrast between RF Δ0.333 and XGB Δ0.421, which results in XGBoost developing 26 per cent more overfitting because both models received identical data training. The failure shows an algorithmic failure: the same data at ~910 Client 2 samples,, which creates a functional RF model, does not help XGBoost learn a generalizable decision boundary through its sequential boosting process.

E. Inter-Client Variability: Consistency and Predictability

A clinically important property of any federated deployment is the consistency of performance across nodes: a framework whose nodes exhibit unpredictable performance variation is harder to govern than one with uniform degradation. For RF, the inter-client test AUC range is 0.065 (0.721 to 0.656) and the F1 range is 0.127 (0.680 to 0.553). For XGBoost, the test AUC range is 0.150 (0.689 to 0.539) and the F1 range is 0.169 (0.653 to 0.484). XGBoost's inter-client variability is 2.3× larger on AUC and 1.3× larger on F1. This higher variability makes XGBoost's federated performance harder to predict and guarantee across nodes—a significant governance concern for a public health deployment context where minimum service standards must be maintained at every participating facility.

VII. Discussion

A. Data Scarcity as the Binding Constraint

Training AUC exceeds 0.953 for every client in both algorithms, while test performance degrades substantially relative to the 1-client baseline reported in the supplementary material. This consistent pattern—near-perfect local training fit paired with meaningful test-set degradation—unambiguously identifies per-client data scarcity as the binding constraint on 3-client federated performance, not the ProxyFL architecture, the gossip communication protocol, or algorithmic incompatibility with the federated setting. The test performance degradation occurs because increasing client partitions creates an inherent privacy protection and statistical generalisation trade-off, which exists in decentralised healthcare learning systems. The Random Forest performance drop occurs at a slow pace, which

demonstrates that the suggested system maintains its ability to make clinical distinctions between patients through the entire process. The ~910 samples available at each 3-client node represent the lower boundary of what is practically viable for tree-based ensemble classifiers: RF retains marginal but functional performance; Low-data decentralised conditions cause XGBoost to show instability because its sequential boosting systems depend on two specific factors, which include partition-level feature sparsity and local distributional variation, more than they do with Random Forest's bagging-based ensemble method.

B. Algorithm Selection: RF as the Robust Choice

For this federated deployment scenario with 3-client comparison, within the evaluated low-data decentralised setting, RF demonstrate greater robustness and calibration stability than XGBoost across all client partitions. It is unique in three ways. First, RF degrades predictably; its performance loss is consistent across partitions and proportional to the data reduction, allowing facility planners to predict capability for a specific node size. Second, while XGBoost's weakest client achieves nearly random discrimination, RF avoids catastrophic failure: even its weakest client (AUC 0.656) maintains operational utility. Third, RF is calibrated; unlike XGBoost Client 2, its well-separated correct/incorrect confidence distributions allow threshold-based triage that is deployable even at the 3-client scale. Future deployments in which per-node sample sizes exceed roughly 10,000 records achievable through multi-facility data pooling within a single federated node remain a viable option for XGBoost, as its superior asymptotic performance would outweigh the overfitting risk.

C. SHAP Consistency and Clinical Validity

The BMI measurement functions as the top SHAP predictor for both algorithms across all client tests, which proves its status as the main anthropometric risk factor. The secondary SHAP rankings demonstrate clinical coherence through age group indicators, which show developmental stage dependency at Z score thresholds, while XGB Client 2 uses infectious disease features, which include Anemia_1, Diarrhea_1 and TB_1 to demonstrate the partition's high comorbidity burden and established biological pathways that link these conditions with malnutrition. The SHAP analysis provides model explainability through its demonstration that the learned feature weights of the model match the clinical epidemiology for each local data subset, while it shows indirect validation of partition composition.

D. Experimental Results

The experimental setup held up well across all client setups. Performance stayed stable regardless of how the data was split, which is the core thing a decentralised system needs to demonstrate.

Both Random Forest and XGBoost showed a train-test gap, indicating small overfitting, as each client only sees a slice of the full dataset, so the models don't get enough variety to generalise cleanly. There is a direct trade-off between splitting data across nodes (which protects privacy) but hurts generalisation.

The collaborative rounds clearly help metrics, don't collapse the way they would if each client were trained in full isolation. So the observation is that decentralised federated learning works, but it incurs a generalisation cost. In settings where sharing raw data isn't an option, that costs generalisation.

E. Limitations

The scope of these findings is limited by three factors. (1) Dataset scale: The dataset makes the 3-client configuration the simulation's practical limit; the failure threshold for XGBoost and the lower viable boundary for RF are represented by the ~428 samples per node. The 5-client and 10-client regimes could be evaluated in future work using multi-facility datasets larger than 10,000 records. (2) Binary classification target: multi-class prediction across WHO Z-score severity categories would enable more focused intervention protocols and provide more precise clinical routing. The malnutrition risk label combines stunting, wasting, and underweight into a single binary outcome.

VIII. CONCLUSION

In the ProxyFL framework, which functions as a decentralised federated learning system to predict the risk of childhood malnutrition, the study directly compared Random Forest and XGBoost. For the experiments, the study used three clients, each with about 428 training samples.

This comparison reveals four main conclusions. First, RF performs better than XGBoost on every test-set metric across all three client partitions. The advantage ranges from a small 0.032 AUC gap at Client 0 to a disastrous 0.117-point gap at Client 2, where RF maintains operational performance (AUC 0.656) while XGBoost collapses to near-random classification (AUC 0.539). Second, on test AUC, RF's inter-client performance variability is 2.3times lower than XGBoost's, indicating that RF generates more consistent, governable federated performance across nodes—a crucial characteristic for minimum-standard public health deployments. Third, a useful confidence-thresholded clinical triage protocol is made possible by RF's well-calibrated confidence distributions, which consistently distinguish between accurate predictions (confidence 0.75–0.95) and errors (0.50–0.65) across all three clients. Fourth, the ProxyFL gossip architecture is verified to be algorithm-agnostic and operationally stable throughout all 20 communication rounds at the 3-client scale for both algorithms. This confirms that the observed performance differences are solely due to data scarcity and algorithmic generalisation properties rather than the federated communication mechanism. These results confirm that the suggested configuration for the short-term implementation of federated malnutrition screening in Ethiopian health settings with limited resources is RF-based ProxyFL. The research results thus show that decentralised proxy-based learning maintains clinically accurate prediction capabilities because it does not need either raw data transmission or gradient sharing, or central coordination, and ProxyFL establishes a promising solution for healthcare AI systems that need to protect user privacy while they operate in environments with restricted infrastructure capacity. The framework introduced a distinct implementation of FL that operates without centralised aggregation or gradient sharing. The traditional approaches, including FedAvg [1] and SCAFFOLD [6], depend on a central server to conduct distributed optimisation through the exchange of parameters and gradients, which results in excessive communication needs, creates a single point of failure and allows potential privacy breaches. The differential privacy-based FL methods use

noise injection to decrease gradient leakage, but this approach usually results in reduced model performance. The decentralised and gossip-based federated learning methods used by Lalitha and his colleagues [7] eliminate the need for a central server, but their system still requires parameter sharing, which creates a risk of exposing confidential data. Thus the work combines decentralisation, communication efficiency, privacy preservation, and explainable healthcare AI within a unified framework suitable for resource-constrained clinical environments.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. AISTATS*, 2017. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

[2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9084352

[3] R. Shokri and V. Shmatikov, "Privacy-Preserving Distributed Stochastic Gradient Descent," *Proc. ACM CCS*, 2015. [Online]. Available: https://www.cs.cornell.edu/~shmat/shmat_ccs15.pdf

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] P. Kairouz *et al.*, "Advances and Open Problems in Federated Learning," 2022. [Online]. Available: https://www.researchgate.net/publication/352779989_Advances_and_Open_Problems_in_Federated_Learning

[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," *Proc. ICML*, 2020. [Online]. Available: https://proceedings.mlr.press/v119/karimireddy20a.html

[7] A. Lalitha, O. S. Shekhar, and T. Javidi, "Fully Decentralized Federated Learning," 2022. [Online]. Available: https://bayesiandeeplearning.org/2018/papers/140.pdf

[8] Z. Zhang, Y. Yang, Z. Yao, Y. Yan, J. E. Gonzalez, K. Ramchandran, and M. W. Mahoney, "Improving Semi-supervised Federated Learning by Reducing the Gradient Diversity of Models," *Proc. IEEE Big Data*, 2021. [Online]. Available: https://www.researchgate.net/publication/373194295_Zhang_et_al_2023

[9] Ethiopian Public Health Institute, "Child Malnutrition Dataset (Ethiopia)," Kaggle Public Repository, 2022.

[10] World Health Organization, "WHO Child Growth Standards: Methods and Development," Geneva, 2006. [Online]. Available: https://www.who.int/tools/child-growth-standards/standards

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD*, 2016. [Online]. Available: https://dl.acm.org/doi/epdf/10.1145/2939672.2939785