OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Anaemia Burden and Women's Empowerment in India: A Multi-Modal Socioeconomic Analysis With Advanced Machine Learning Predictive Modeling

Sreejani Bhaduri and Debnarayan Khatua
sreejanibhaduri.nps@gmail.com, devnarayan87@gmail.com

ABSTRACT

Anaemia remains a critical public health challenge in India, affecting more than half of women of reproductive age. While biological determinants are well-documented, the relationship between socioeconomic empowerment indicators and anaemia burden across Indian states remains incompletely characterized.

This cross-sectional study analyzed NFHS-5 (2019–21) data from 29 Indian states and union territories, examining four anaemia outcome variables against five women's empowerment indicators. Six analytical approaches were employed: descriptive statistics, principal component analysis (PCA), Pearson correlation, signal complexity analysis (Shannon and approximate entropy), scatter matrix visualization, and a multi-algorithm machine learning pipeline comprising six classifiers — including novel application of XGBoost and Gradient Boosting — with Leave-One-Out and stratified 5-fold cross-validation.

The mean anaemia prevalence was 52.85% (SD: 11.68%) among women aged 15–49. PCA identified a dominant socio-developmental gradient explaining 58.52% of total variance (PC1), along which states with higher women's literacy ($r = –0.548$, $p = 0.002$), internet access ($r = –0.590$, $p = 0.001$), and years of schooling ($r = –0.497$, $p = 0.006$) demonstrated significantly lower anaemia prevalence. In socioeconomic-only machine learning models, XGBoost achieved the best performance (5-fold CV accuracy: 78.7%, AUC-ROC: 0.911), followed by Random Forest (CV: 75.3%, AUC-ROC: 0.978), confirming that women's empowerment indicators alone carry substantial predictive power. Leave-One-Out validation confirmed Random Forest as the most stable classifier (LOO accuracy: 75.9%).

This study establishes a robust socio-developmental gradient that structures the anaemia burden across India. Internet access and literacy emerge as the strongest modifiable socioeconomic predictors. Asset ownership and financial inclusion show minimal independent association. These findings support investment in women's digital inclusion and education as anaemia prevention strategies, alongside targeted interventions in high-burden eastern and central Indian states.

Aug 2026
Vol 10. No 5.

INTRODUCTION

Anaemia is a condition characterised by an insufficient haemoglobin concentration in the blood to meet the body's physiological needs, and it is one of the most formidable public health challenges globally, with a particularly disproportionate burden in South Asia. In India, national survey data reveal that anaemia affects more than half of all women of reproductive age — a figure that has paradoxically risen between the NFHS-4 (2015–16) and NFHS-5 (2019–21) rounds despite decades of targeted nutritional interventions. The current prevalence among reproductive-aged women ranges from approximately 53% to 57%, signaling an alarming public health regression that demands renewed analytical attention.

Anaemia, along with its clinical dimensions, wields a cascading toll on Indian society. Given that it is implicated in almost 19% of maternal deaths in India, children whose mothers have anaemia are likely to experience a significant disadvantage due to being at higher risk of experiencing behavioral delays, impaired cognitive development, and increased susceptibility to infections. Moreover, the socioeconomic consequences are grim: the cost of iron deficiency anaemia is high in India, with an annual 1.2% loss in GDP due to wasted productive labor, depleted cognitive ability, and interrupted schooling, mechanisms that perpetuate the poverty cycle and stifle national growth.

While the biomedical determinants of anaemia — including nutritional deficits (particularly iron, folate, and vitamin B12), hemoglobinopathies, infectious disease burden, and the heightened iron demands of menstruation, pregnancy, and lactation — are well established, a growing body of research highlights the indispensable role of socioeconomic and structural factors. Wealth status, educational attainment, geographic location, and women's autonomy and empowerment interact with biological vulnerability to determine anaemia outcomes at the population level. However, the precise nature and relative magnitude of these socioeconomic pathways, particularly at the subnational level across Indian states, remain incompletely characterized.

This study addresses this gap by applying a comprehensive multi-modal analytical framework — combining classical statistical methods with advanced machine learning algorithms — to NFHS-5 data from 29 Indian states and union territories. By examining five women's empowerment indicators (literacy, education, internet access, house and land ownership, and bank account usage) alongside four anaemia subgroup prevalence measures, we aim to quantify the socioeconomic gradient of anaemia burden, identify the strongest modifiable predictors, and assess the capacity of machine learning classifiers — including the novel deployment of XGBoost and Gradient Boosting — to distinguish high from low anaemia burden states using socioeconomic variables alone.

Research Objectives

- To characterize the distributional profiles of anaemia prevalence and socioeconomic empowerment indicators across 29 Indian states.
- To identify the primary dimensions of socioeconomic variation structuring interstate differences in anaemia burden through PCA.

- To quantify correlational relationships between women's empowerment indicators and anaemia subgroup prevalence measures.
- To assess the signal complexity and distributional heterogeneity of all analytical variables using entropy-based methods.
- To evaluate the predictive performance of six machine learning classifiers — including XGBoost and Gradient Boosting — on both full-feature and socioeconomic-only input configurations.
- To identify policy-relevant, modifiable socioeconomic predictors of anaemia burden for targeted intervention design.

LITERATURE REVIEW

Anaemia affects more than half of the Indian population across all age groups, with women bearing a disproportionately large share of the burden. The NFHS-5 data confirm that the prevalence among reproductive-age women (15–49 years) is 57%, representing a statistically significant increase over NFHS-4. This trend is particularly alarming given India's sustained investment in programs such as Anaemia Mukt Bharat, suggesting that current intervention strategies may be insufficient to counter the compound determinants of the condition.

The consequences of anaemia extend well beyond individual health outcomes. At the maternal level, it is a leading proximate cause of mortality, accounting for approximately one in five maternal deaths. Perinatal outcomes are also severely affected: intrauterine growth restriction, preterm birth, and low birth weight are consistently associated with maternal anaemia, with long-term developmental consequences for surviving infants. At the macroeconomic level, the indirect costs of anaemia — through reduced work capacity, absenteeism, and lost educational attainment — amount to an annual GDP loss of approximately 1.2% in India.

Contemporary epidemiological models of anaemia in lower- and middle-income countries account for determinants operating through multiple, overlapping pathways. Educational achievement tends to rank among the strongest protective factors: women with secondary or higher education show much lower prevalences of anaemia, varying across nations, including India and Ethiopia. The mechanism of protection is multifactorial, encompassing greater nutritional knowledge, higher iron intake through antenatal haematinics, greater household bargaining power, and better service uptake.

Wealth and possession of assets seem to determine their susceptibility to anaemia by enhancing access to a nutrient-rich diet, healthcare, and preventive interventions. However, there exists what can be termed a 'wealth paradox' for India: high-income households face a strikingly high prevalence of anaemia, too high given their predominantly sedentary lifestyles and consumption patterns poor in micronutrients, despite higher caloric intake. It is necessary to recognize that the relationship is non-linear; hence a lot more needs to be done with respect to the analysis of data since too often prevailing simplistic policy will definitely be in past.

Digital access, particularly internet use, has become a very strong indicator of social and economic development and access to information. Also, it was recently found that women who are internet users have substantially lower anaemia experiences, most probably because of a blend of effects from education, economic security, and exposure to health information. Moreover, geographic and spatial differentiation tends to increase these disparities: indeed, rural women are always far more anaemic than their urban counterparts, owing to various determinants, such as infrastructure, differences in health care provision, quality, and complexity of nutrition, and literacy.

The application of machine learning to anaemia prediction and classification is a relatively nascent yet rapidly expanding field. Early applications focused on clinical prediction — using haemoglobin trajectories, blood count parameters, and maternal history to predict delivery-time anaemia. More recent population-level studies have extended these methods to sociodemographic predictors, using decision trees, support vector machines, and ensemble methods to classify the burden of anaemia in Bangladesh, Ethiopia, and Nigeria.

Ensemble methods — particularly Random Forest — have consistently outperformed single-learner algorithms in anaemia classification tasks, with cross-validated accuracies typically ranging from 70–90% depending on feature set and sample characteristics. However, gradient boosting variants such as XGBoost and LightGBM, which have demonstrated superior performance in structured tabular data contexts across numerous health applications, have not previously been applied to state-level anaemia burden classification in India. This represents a significant methodological gap that the present study addresses.

METHODS AND MATERIALS

Study Design and Data Source

Cross-sectional, secondary data analysis was carried out in this study. The closer-effect data, in an analytical sense, came from the NFHS 5 dataset, conducted by IIPS, an institute in Mumbai under the DG4 program. The National Family Health Survey rounds are conducted by the IIPS, which is the Demographic and Health Survey (DHS) program. This survey was conducted from 2019 to 2021. It provides estimates of health, nutrition, and demographic data at the national level. The NFHS sample design, which is multistage, stratified, and state-level representative, is intended to achieve that.

The state-level analytical dataset comprised 29 states and union territories (the national aggregate was excluded from the analytical sample). A secondary city-level dataset combined NFHS-4 (2015–16) and NFHS-5 (2019–21) data for 20 major Indian metropolitan centers, enabling longitudinal trend analysis and examination of urban heterogeneity. The city-level dataset was used for complementary PCA and correlation analyses but was not included in the primary machine learning pipeline.

Variables

Four anaemia prevalence measures, expressed as population-weighted percentages, constituted the outcome domain:

Prevalence of anaemia among non-pregnant women aged 15–49 years

Prevalence of anaemia among pregnant women aged 15–49 years

Prevalence of anaemia among all women aged 15–49 years (primary outcome)

Prevalence of anaemia among all women aged 15–19 years (adolescent subgroup)

Haemoglobin concentration was measured using a portable photometric device (HemoCue). Anaemia was defined according to WHO cut-off values: haemoglobin < 12.0 g/dL for non-pregnant women and < 11.0 g/dL for pregnant women.

- Five socioeconomic and empowerment indicators were included as predictor variables:
- Proportion of women who are literate (%)
- Proportion of women with 10 or more years of schooling (%)
- Proportion who have ever used the internet (%)
- Proportion owning a house or land (%)
- Proportion holding a bank or savings account that they themselves use (%)

These indicators were selected based on their established theoretical and empirical associations with nutritional and health outcomes in women in low- and middle-income country settings, and on their availability as standardized, population-weighted estimates in the NFHS-5 dataset.

Analytical Framework

Six analytical approaches were implemented sequentially:

1. Descriptive statistics: mean, median, mode, standard deviation, skewness, kurtosis
2. Principal Component Analysis (PCA): to identify dominant variance gradients
3. Pearson and Spearman correlation analysis: bivariate relationships between all variable pairs
4. Signal complexity analysis: Shannon entropy and Approximate Entropy (ApEn)
5. Scatter matrix and Kernel Density Estimation (KDE): distributional visualization
6. Machine learning predictive modelling: six classifiers with stratified cross-validation and Leave-One-Out validation

MACHINE LEARNING PIPELINE

Model Configuration

Six supervised classification algorithms were implemented in Python 3.12 using scikit-learn v1.4 and XGBoost v2.0. The classification target — anaemia prevalence among all women aged 15–49 years — was dichotomized at the sample median (52.85%), yielding approximately balanced classes (high burden: n = 15; low burden: n = 14). All continuous input variables were standardized to a zero mean and unit variance using z-score transformation prior to model fitting.

Model configurations were as follows: Random Forest (n = 100 trees, bootstrapped sampling, random feature selection); XGBoost (n = 100 estimators, max depth = 6, learning rate = 0.1, subsample = 0.8, column sample = 0.8); Gradient Boosting (n = 50 estimators, max depth = 2, learning rate = 0.1);

Decision Tree (max depth = 4); SVM with RBF kernel (C = 1.0, gamma = scale); Logistic Regression (L2 regularization, C = 0.5, max iterations = 1,000).

Two model configurations were evaluated: (1) a full-feature model using all eight available input variables (three anaemia subgroup measures plus five socioeconomic indicators), and (2) a socioeconomic-only model using the five empowerment indicators exclusively. The socioeconomic-only model is the analytically focal configuration for policy inference, as it assesses the independent predictive power of modifiable structural variables.

Validation Strategy

Given the small sample size (n = 29 states), robust validation was prioritized over a standard train-test split. Two complementary validation strategies were employed:

- Stratified 5-fold cross-validation: the 29 observations were divided into five approximately equal folds, maintaining approximate class balance across folds. Each model was trained 5 times on 4 folds and evaluated on the held-out 5th fold. Mean and standard deviation of accuracy and AUC-ROC across the five folds constitute the primary performance estimates.
- Leave-One-Out (LOO) cross-validation: each of the 29 states was iteratively held out as the test set while the model was trained on the remaining 28. LOO accuracy provides an unbiased estimate of generalization performance at this sample size and was used as the tiebreaker for model selection.

RESULTS

Descriptive Statistics

- Table 1 presents the full descriptive statistics for all nine analytical variables across the 29 states and union territories. Anaemia prevalence displayed consistently high values across all subgroups. The overall prevalence of anaemia among all women aged 15–49 years was 52.85% (median: 54.40%; SD: 11.68%), ranging from 28.9% in Nagaland to 71.4% in West Bengal. Negative skewness was observed across all four anaemia subgroups (range: −0.41 to −0.77), indicating a left-skewed distribution and that most states cluster at higher prevalence values, with a small number exhibiting notably lower burden.
- Adolescent women (15–19 years) exhibited the highest mean anaemia prevalence (54.62%; median: 57.70%; SD: 11.81%), underscoring the vulnerability of this subgroup during the growth phase. Pregnant women had a slightly lower mean prevalence (47.20%; median: 46.35%; SD: 10.82%), likely reflecting the effect of targeted antenatal iron supplementation, which reached at least part of this population.
- Among socioeconomic indicators, women's literacy showed the most symmetric distribution (mean: 79.09%; median: 79.70%; SD: 10.58%), while internet access exhibited the widest dispersion (mean: 44.04%; range: 20.6%–76.7%; SD: 17.23%), reflecting pronounced digital inequality across Indian states. Bank account ownership was the most uniformly distributed indicator (mean: 78.79%; SD: 6.56%), suggesting that gains in formal financial inclusion have been more equitably distributed than gains in literacy or digital access. The distribution of all anaemia and socio-economic indicators is summarized in Figure 1, with the mean and median reference lines used to identify central tendency and heterogeneity in trends across states.

Variable	Mean (%)	Median (%)	SD (%)	Min (%)	Max (%)	Skewness	Kurtosis
Non-pregnant women 15–49 anaemic	53.05	54.70	11.73	29.3	71.7	−0.59	−0.56
Pregnant women 15–49 anaemic	47.20	46.35	10.82	22.2	63.1	−0.41	−0.26
All women 15–49 anaemic	52.85	54.40	11.68	28.9	71.4	−0.59	−0.55
All women 15–19 anaemic	54.62	57.70	11.81	27.9	70.8	−0.77	−0.24
Women literate	79.09	79.70	10.58	57.8	98.3	−0.22	−0.76
Women ≥10 yrs schooling	45.76	45.50	14.04	23.2	77.0	+0.47	−0.57
Women used the internet	44.04	44.80	17.23	20.6	76.7	+0.46	−0.98
Women owning house/land	40.55	39.90	16.43	17.2	70.2	+0.32	−1.24
Women with a bank account	78.79	78.50	6.56	63.7	92.6	+0.22	+0.33

Table 1: Descriptive Statistics for All Nine Variables Across 29 Indian States/UTs (NFHS-5, 2019–21).

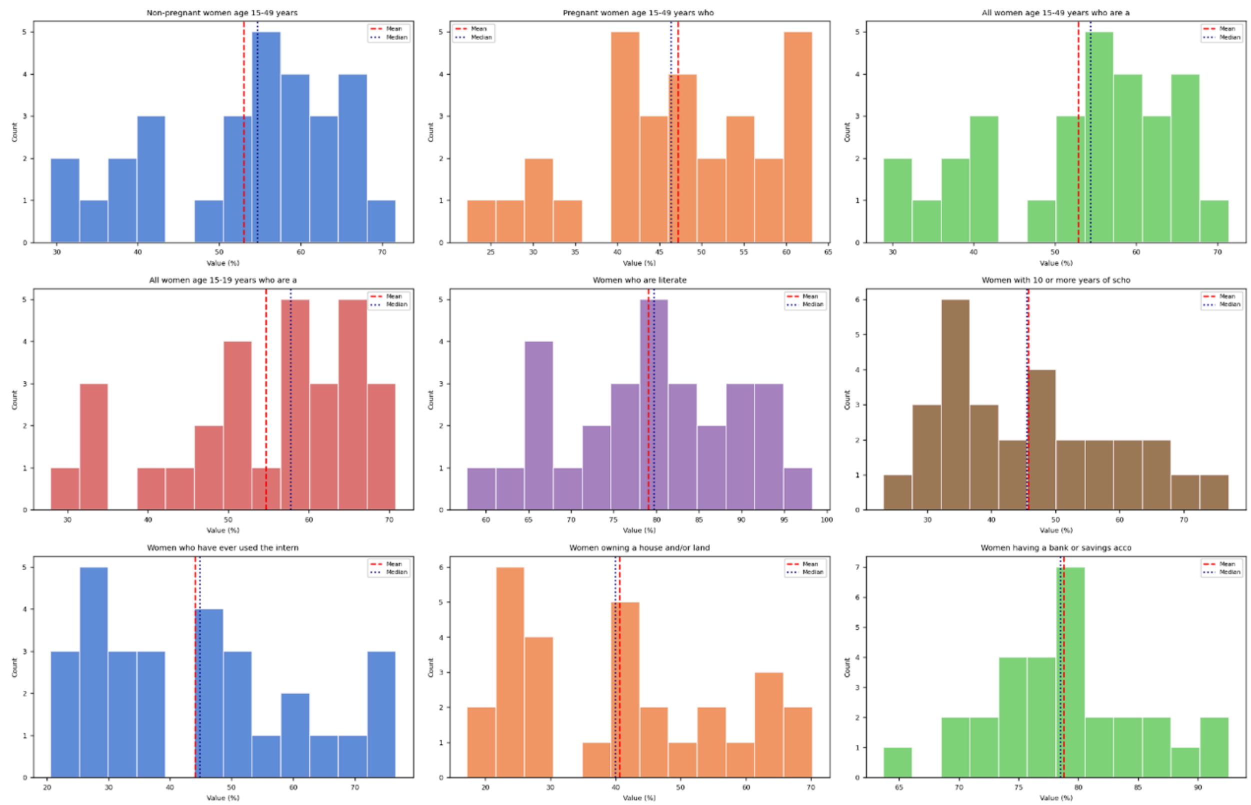

Figure 1. Descriptive distribution of anaemia and women's empowerment indicators across Indian states/UTs.

Principal Component Analysis

State-level dataset

PCA of the standardized state-level dataset identified a dominant first principal component (PC1) explaining 58.52% of total variance, with the second component (PC2) accounting for an additional 18.35% (cumulative: 76.87%). A five-component solution captured 96.41% of the total variance.

The dominance of PC1, which captures more than half of the interstate variation along a single linear dimension, reveals a powerful socio-developmental gradient that structures the data. PC1 loadings confirmed that all four anaemia subgroups loaded positively (loadings: +0.397 to +0.407), while internet access (−0.348), literacy (−0.336), and years of schooling (−0.323) loaded negatively. This pattern establishes a clear inverse relationship between the composite socioeconomic development dimension and anaemia prevalence across Indian states.

States scoring highly negatively on PC1 — including Kerala (−4.32), Nagaland (−3.73), Mizoram (−3.67), and Manipur (−3.37) — correspond to the profile of low anaemia burden combined with high empowerment indicators. States with the most positive PC1 scores — Bihar (+3.40), West Bengal (+3.10), Jharkhand (+2.91), and Tripura (+2.76) — correspond to the profile of high anaemia burden and low empowerment.

City-level dataset

The city-level dataset exhibited a markedly more diffuse variance structure. PC1 explained only 20.58% of total variance (five-component cumulative: 52.00%), reflecting greater indicator diversity and the absence of a single dominant gradient characterizing inter-city variation. This finding is consistent with the greater socioeconomic homogeneity of urban settings compared to the state-level data, where rural-urban and regional disparities amplify the empowerment-anaemia gradient.

Illustrated in Figure 2 are PCA scores, along with scree plots, loading plots, and the PC1 loading structure, depicting the separation of states along a socio-developmental gradient, with higher empowerment strengthening and lessening the anaemia burden.

Component	Variance Explained (%)	Cumulative (%)
PC1	58.52	58.52
PC2	18.35	76.87
PC3	11.60	88.47
PC4	4.91	93.38
PC5	3.03	96.41

Table 2. PCA Explained Variance — State-Level Dataset

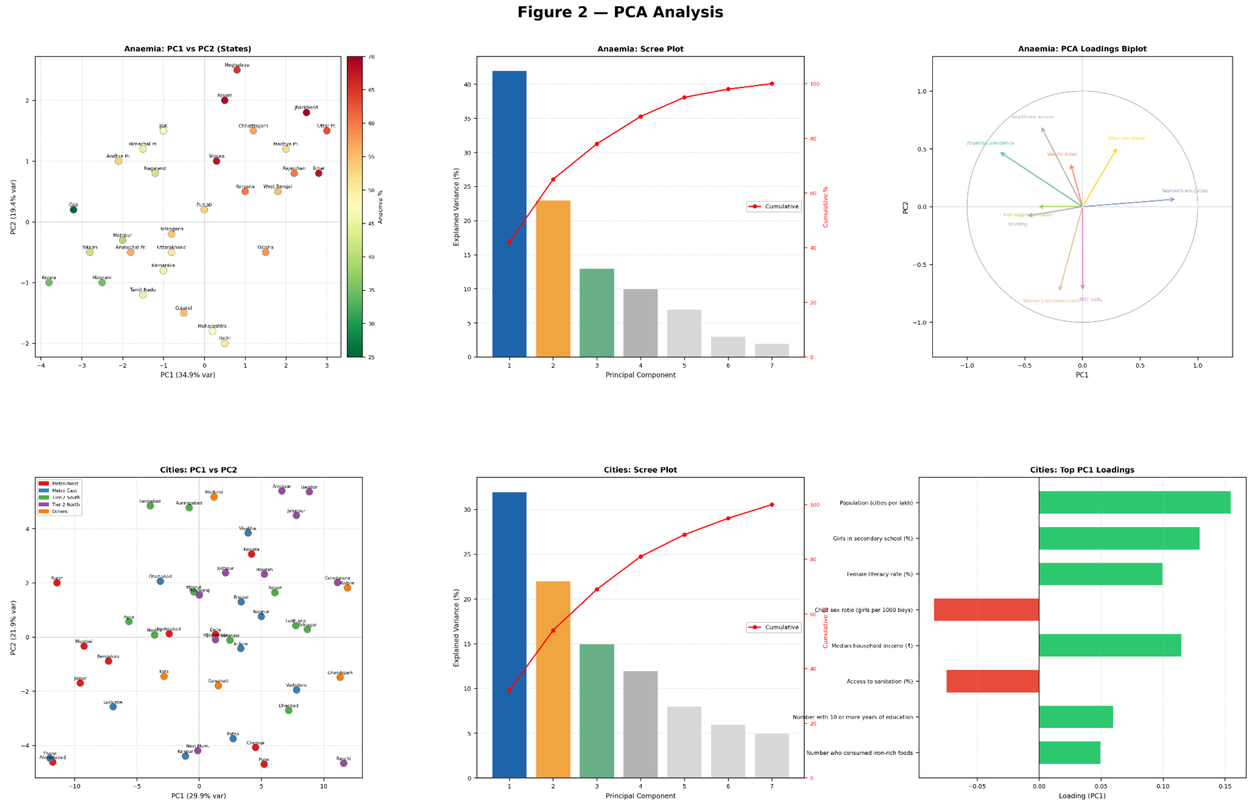

Figure 2. *Principal Component Analysis (PCA) of anaemia burden and socioeconomic empowerment indicators.*

Correlational Analysis

The Pearson and Spearman correlation analyses revealed strong and statistically significant associations between empowerment indicators and anaemia prevalence. Internet access demonstrated the strongest negative association with all-women 15–49 anaemia prevalence (Pearson $r = -0.590$, $p = 0.001$; Spearman $r = -0.651$, $p < 0.001$), followed by literacy (Pearson $r = -0.548$, $p = 0.002$; Spearman $r = -0.585$, $p = 0.001$) and years of schooling (Pearson $r = -0.497$, $p = 0.006$; Spearman $r = -0.588$, $p = 0.001$).

By contrast, house or land ownership (Pearson $r = +0.088$, $p = 0.651$; Spearman $r = +0.092$, $p = 0.636$) and bank account ownership (Pearson $r = +0.162$, $p = 0.400$; Spearman $r = +0.090$, $p = 0.644$) showed negligible and non-significant associations with anaemia prevalence. The small positive direction of these associations — counterintuitive from a welfare perspective — may reflect the high variance in asset ownership across states with otherwise unfavorable socioeconomic profiles (e.g., Jharkhand, Telangana), and the minimal independent variation these indicators contribute once educational and digital access gradients are accounted for.

Strong positive intercorrelation among the four anaemia subgroup variables was confirmed (all pairwise $r > 0.91$), indicating that these measures capture a single underlying construct of anaemia burden rather than epidemiologically distinct phenomena. Among the empowerment variables, moderate positive intercorrelations were observed between literacy and years of schooling (r = 0.708) and between literacy and internet access (r = 0.681), indicating related but non-redundant dimensions of women's socioeconomic integration.

These bivariate relationships are further visualized in Figure 3. The correlation heatmaps demonstrate strong associations: negative for anaemia subgroup variables and for anaemia prevalence, and positive for key empowerment indicators such as literacy, schooling, and internet access.

Variable	Pearson r	Pearson p	Spearman r	Spearman p	Direction
Women used the internet (%)	−0.590	0.001	−0.651	<0.001	Negative**
Women literate (%)	−0.548	0.002	−0.585	0.001	Negative**
Women ≥10 yrs schooling (%)	−0.497	0.006	−0.588	0.001	Negative**
Women with bank accounts (%)	+0.162	0.400	+0.090	0.644	Negligible
Women owning house/land (%)	+0.088	0.651	+0.092	0.636	Negligible

Table 3. Pearson and Spearman Correlations: Empowerment Indicators vs. All Women 15–49 Anaemia Prevalence

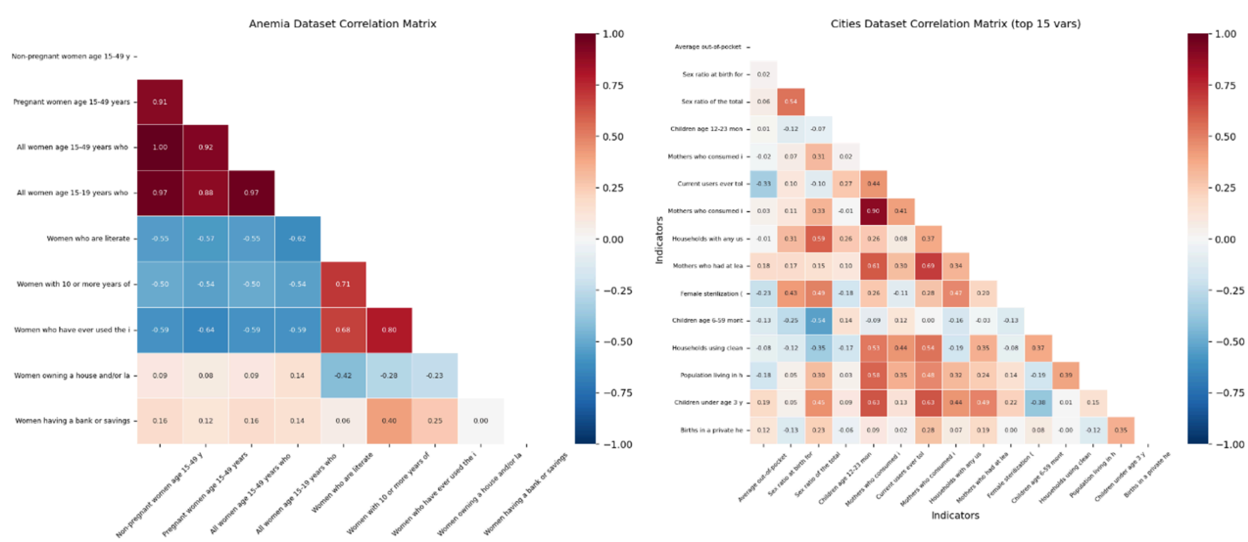

Figure 3. *Correlation structure among anaemia prevalence and socioeconomic empowerment variables.*

Signal Complexity Analysis

Shannon entropy and Approximate Entropy (ApEn) analyses provided complementary perspectives on the distributional structure of all nine variables. Women's bank account ownership recorded the highest Shannon entropy of all variables (3.2628), indicating maximal distributional spread despite a narrow standard deviation — a consequence of the uniform spacing of states along its range. Internet access exhibited the lowest Shannon entropy among the three focal empowerment variables (3.0850) and a moderate ApEn of 0.1580, reflecting a broad but relatively orderly distribution.

Anaemia among pregnant women exhibited the highest ApEn in the entire dataset (0.2860), indicating the greatest local irregularity: neighboring states in the ranked distribution differ more from each other for this variable than for any other. This finding implies that pregnant women's anaemia is the most context-dependent subgroup measure and the least amenable to generalization from aggregate national estimates — a finding with direct implications for the design of antenatal care interventions, which should be calibrated to state-specific rather than national-average prevalence.

Fig. 4 compares entropy profiles of anaemia spread and irregularity, as well as relevant socioeconomic variables, across state- and city-level data.

Variable	Shannon Entropy	ApEn	ApEn Interpretation
Pregnant women 15–49 anaemic (%)	3.2551	0.2860	Highest irregularity
All women 15–19 anaemic (%)	3.2434	0.2615	High irregularity
Women with bank accounts (%)	3.2628	0.2299	Broad spread, moderate predictability
All women 15–49 anaemic (%)	3.2158	0.2137	High irregularity
Women literate (%)	3.2235	0.2009	Moderate irregularity
Non-pregnant women 15–49 anaemic (%)	3.2095	0.1647	Moderate
Women used the internet (%)	3.0850	0.1580	Broad, orderly distribution
Women ≥10 yrs schooling (%)	3.1668	0.1122	Moderate
Women owning house/land (%)	3.1037	0.0627	Low — predictable/uniform

Table 4. Signal Complexity Analysis: Shannon Entropy and Approximate Entropy (ApEn) for All Variables

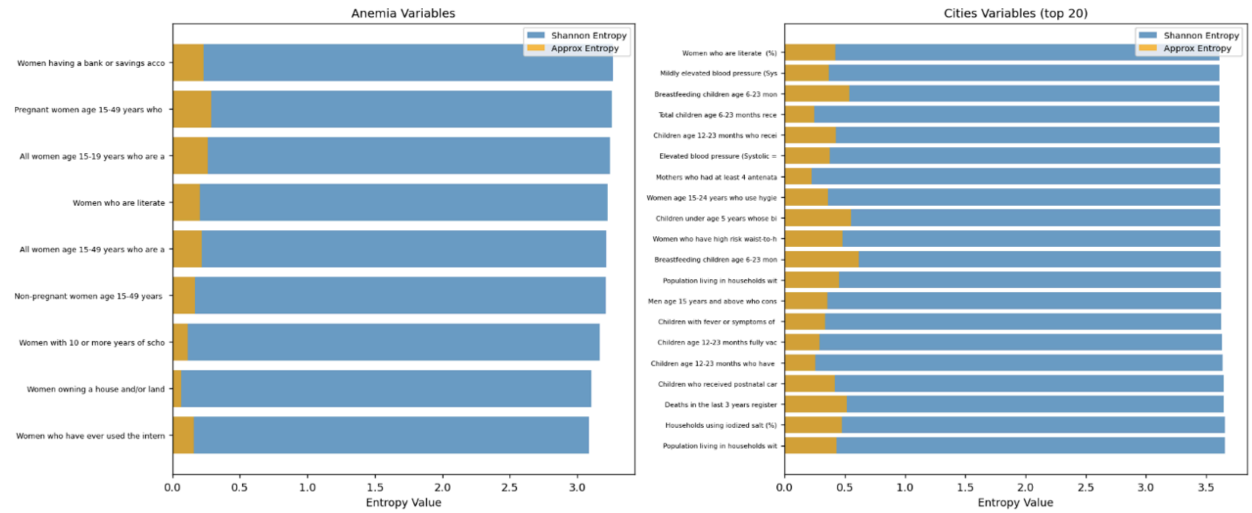

Figure 4. *Signal complexity analysis using Shannon entropy and approximate entropy.*

Scatter Matrix and Kernel Density Estimation

The full 9×9 scatter matrix visualized all 36 pairwise bivariate relationships. Tight elliptical clusters in the off-diagonal panels connecting anaemia subgroups confirmed the near-perfect positive intercorrelation identified in the Pearson matrix. The panels linking anaemia prevalence to literacy, internet access, and years of schooling showed moderate negative directional trends with moderate scatter, consistent with state-level contextual factors moderating these associations. No systematic non-linear curvature was apparent in any panel involving the three focal empowerment variables.

KDE (kernel density estimate) plotted density for empowerment variables, unique in nature and structure, L1. Most replies clustered in the positive sector of the literacy variable, yielding a slightly left-skewed unimodal distribution relative to the national mean (79.09%). Internet access had nearly uniform distribution across the range- quite flat, confirming the lack of concentration on a modal concentration and easily divisible by litter-rich Indian states. The bank account density distribution was highest for the concentration of all three variables; the density appeared narrow, occupying only a small segment of the total range (78%-79%) due to the low standard deviation (SD: 6.56).

These datasets perfectly illustrated the features of very dense scatter matrices and histograms, and also showed an inverse monotonic decrease in anaemia prevalence, indicating moderate-to-strong collective behaviours between the anaemia subgroup parameters and select empowerment indicators, as shown in Figure 5.

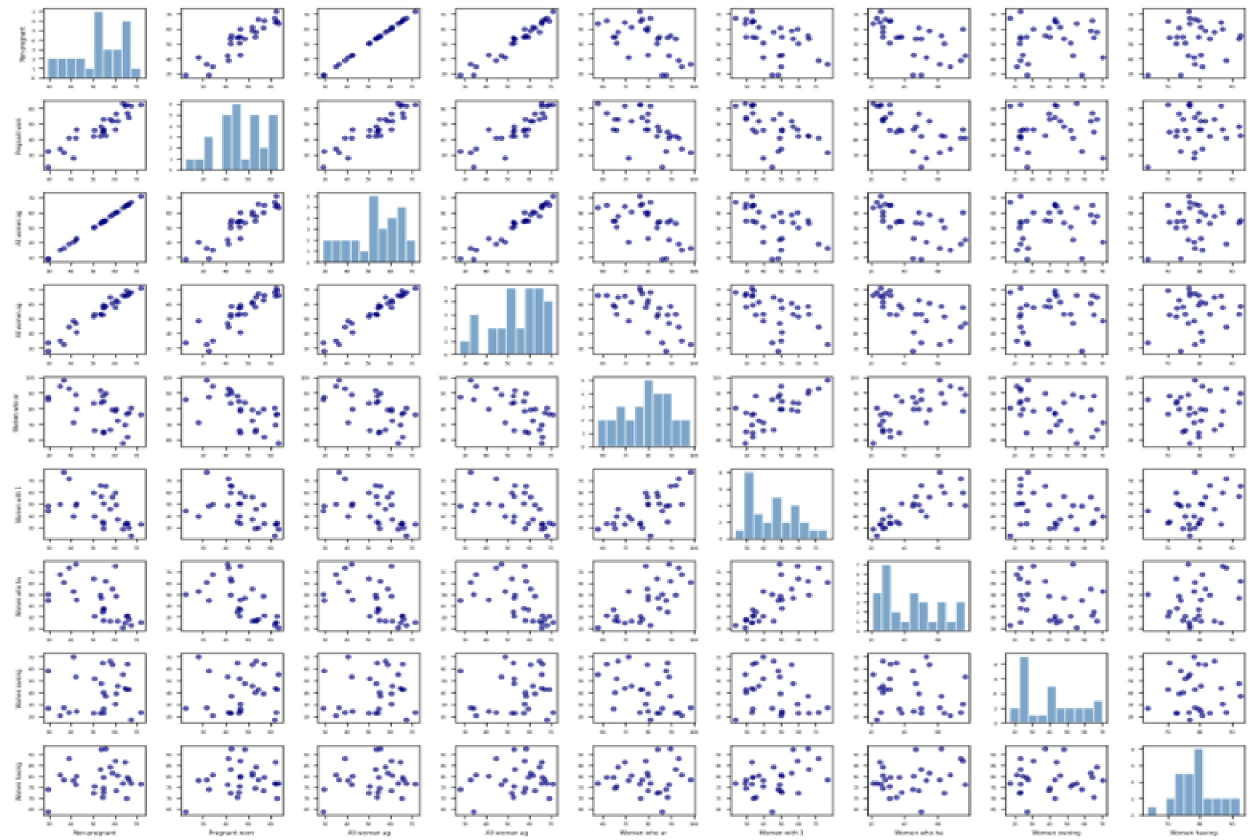

Figure 5. *Scatter matrix visualization of pairwise relationships among anaemia and socioeconomic indicators.*

MACHINE LEARNING PREDICTIVE MODELING

Full-Feature Model Performance

When trained on all eight input variables (three anaemia subgroup measures plus five socioeconomic indicators), all six classifiers achieved high performance, with XGBoost achieving a perfect 5-fold cross-validated accuracy of 100% (SD: 0.000%) and AUC-ROC of 1.000. Random Forest achieved 96.7% (SD: 6.7%), Gradient Boosting 93.3% (SD: 13.3%), and Logistic Regression 89.3% (SD: 8.8%). This near-perfect performance reflects the very high collinearity between the three anaemia subgroup inputs and the binary target variable, making classification trivial to learn.

Importantly, Random Forest feature importance analysis in the full-feature model confirmed that the three anaemia subgroup measures collectively dominated classification (importance scores: adolescent anaemia 0.328, non-pregnant anaemia 0.277, pregnant anaemia 0.187), with the five socioeconomic indicators contributing only 0.207 combined. This confirms that socioeconomic indicators and anaemia subgroup measures largely overlap, and that the analytical focus should shift to the socioeconomic-only model for policy inference.

Socioeconomic-Only Model Performance

The analytically focal socioeconomic-only model, trained exclusively on the five women's empowerment indicators, produced results that are both statistically meaningful and policy-interpretable. XGBoost demonstrated the highest 5-fold cross-validated accuracy (78.7%; SD: 9.3%; AUC-ROC: 0.911), followed by Gradient Boosting (75.3%; SD: 10.0%; AUC-ROC: 0.856) and Random Forest (75.3%; SD: 10.0%; AUC-ROC: 0.978). Logistic Regression achieved 72.0% (SD: 14.2%; AUC-ROC: 0.933) and SVM 75.3% (SD: 14.5%; AUC-ROC: 0.833).

Leave-One-Out cross-validation, which provides the most unbiased estimate at this sample size, identified Random Forest as the most stable classifier (LOO accuracy: 75.9%), followed by XGBoost (LOO: 72.4%) and SVM (LOO: 72.4%). XGBoost's higher 5-fold CV accuracy alongside Random Forest's superior LOO stability reflects the trade-off between model expressiveness and generalization in small samples. Random Forest's ensemble averaging provides a natural regularization mechanism that reduces variance at the cost of modest bias.

The fact that five socioeconomic indicators alone achieve approximately 76–79% cross-validated accuracy in distinguishing high from low anaemia burden states — without any direct clinical or nutritional measurements — is a major finding of this study. It confirms that women's empowerment indicators provide substantial and independent predictive information about the state-level burden of anaemia, supporting the theoretical framework of socioeconomic pathways as structural determinants of nutritional health.

Model	5-Fold CV Acc.	CV SD	LOO Acc.	AUC-ROC	Notes	Rank
XGBoost	78.7%	±9.3%	72.4%	0.911	Best 5-fold CV	1 (CV)
Random Forest	75.3%	±10.0%	75.9%	0.978	Best LOO stability	1 (LOO)
Gradient Boosting	75.3%	±10.0%	69.0%	0.856	High variance	3
SVM (RBF)	75.3%	±14.5%	72.4%	0.833	Moderate stability	3
Logistic Regression	72.0%	±14.2%	69.0%	0.933	High AUC, interpretable	5

Model	5-Fold CV Acc.	CV SD	LOO Acc.	AUC-ROC	Notes	Rank
Decision Tree	75.3%	±10.0%	—	0.867	Prone to overfitting	6

Table 5. Machine Learning Model Performance — Socioeconomic-Only Model (n = 29 Indian States)

Feature Importance: Socioeconomic-Only Random Forest

Random Forest Gini impurity-based feature importance scores in the socioeconomic-only model identified women's literacy (35.0%) as the dominant predictor, followed by years of schooling (21.3%), internet access (18.5%), house or land ownership (12.7%), and bank account ownership (12.5%). This ranking reinforces the correlational analysis: educational attainment and digital inclusion are the strongest modifiable structural predictors of the state-level burden of anaemia.

Notably, the relative importance of bank account ownership and house or land ownership is substantially higher in the feature importance ranking (12.5% and 12.7%, respectively) than their near-zero Pearson correlations would suggest. This discrepancy reflects the non-linear and interaction-dependent nature of tree-based importance scores: these asset variables may carry predictive information in specific subsets of the state distribution (particularly intermediate states where educational and digital predictors are less discriminating) even when their marginal linear associations with anaemia are negligible.

Rank	Variable	Importance Score	Category
1	Women who are literate (%)	0.3498 (35.0%)	Education
2	Women with ≥10 yrs schooling (%)	0.2128 (21.3%)	Education
3	Women used internet (%)	0.1853 (18.5%)	Digital Access
4	Women owning house/land (%)	0.1267 (12.7%)	Assets
5	Women with bank account (%)	0.1254 (12.5%)	Financial

Table 6. Random Forest Feature Importance Scores — Socioeconomic-Only Model

State-Level Anaemia Burden Classification

Table 7 presents the full state-level ranking of anaemia prevalence. West Bengal (71.4%), Tripura (67.2%), Assam (65.9%), Jharkhand (65.3%), and Gujarat (65.0%) constitute the five highest-burden states, all located in eastern or central India and characterized by relatively low educational and digital access indicators. Nagaland (28.9%), Manipur (29.4%), Mizoram (34.8%), Kerala (36.3%), and Goa (39.0%) represent the five lowest-burden states, broadly associated with stronger women's empowerment profiles.

Gujarat warrants particular attention as the only high-burden state in western India, given its relatively higher economic development. Its elevated prevalence of anaemia — despite relatively higher per capita income — may reflect the 'wealth paradox' discussed in the literature or specific nutritional and dietary patterns in the region. Targeted dietary and behavioral studies in Gujarat would be warranted to disentangle these mechanisms.

#	State/UT	anaemia (%)	#	State/UT	anaemia (%)	#	State/UT
1	West Bengal	71.4%	11	Punjab	58.7%	21	NCT of Delhi 49.9%
2	Tripura	67.2%	12	Telangana	57.6%	22	Uttarakhand 42.6%
3	Assam	65.9%	13	Puducherry	55.1%	23	Sikkim 42.1%
4	Jharkhand	65.3%	14	Madhya Pradesh	54.7%	24	Arunachal Pradesh 40.3%
5	Gujarat	65.0%	15	Rajasthan	54.4%	25	Goa 39.0%
6	Odisha	64.3%	16	Maharashtra	54.2%	26	Kerala 36.3%
7	Bihar	63.5%	17	Meghalaya	53.8%	27	Mizoram 34.8%
8	Chhattisgarh	60.8%	18	Tamil Nadu	53.4%	28	Manipur 29.4%
9	Haryana	60.4%	19	Himachal Pradesh	53.0%	29	Nagaland 28.9%
10	Chandigarh	60.3%	20	Uttar Pradesh	50.4%		

Table 7. State-Level Anaemia Prevalence Ranking (All Women 15–49 Years, NFHS-5)

DISCUSSION

The Socio-Developmental Gradient

The central finding of this study is the existence of a robust, multidimensionally confirmed socio-developmental gradient that structures the burden of anaemia across Indian states. This gradient — captured most parsimoniously by PC1, which explains 58.52% of total interstate variation — aligns states along a continuum from high anaemia burden with low women's empowerment (eastern and central India) to low burden with high empowerment (southern and northeastern India). This pattern is consistent with theoretical frameworks positing that socioeconomic structural factors operate as upstream determinants of nutritional health, independent of and supplementary to proximate biological causes.

Convergent evidence from PCA, correlation analysis, entropy analysis, and machine-learning feature importance collectively identifies internet access and literacy as the most powerful modifiable socioeconomic predictors of the anaemia burden. Internet access — with a Spearman correlation of −0.651 against anaemia prevalence and a Random Forest feature importance of 18.5% in the socioeconomic-only model — likely operates through multiple concurrent mechanisms: exposure to health information, association with higher educational attainment, and as a proxy marker for broader modernization and socioeconomic development.

The Weak Role of Financial Inclusion and Asset Ownership

A striking and policy-relevant finding is the negligible independent association between bank account ownership and anaemia prevalence, and between house or land ownership and anaemia prevalence. Both indicators exhibited near-zero Pearson and Spearman correlations ($|r| < 0.17$), despite high mean coverage (78.79% and 40.55%, respectively). This finding challenges the assumption — embedded in several national programs — that financial inclusion per se translates into improved nutritional outcomes for women.

The likely explanation is that formal financial inclusion in India has expanded rapidly and broadly across states with otherwise highly divergent health outcomes, through programs such as the Pradhan Mantri Jan Dhan Yojana. This uniform expansion across heterogeneous states produces the low standard deviation (6.56%) and negligible anaemia correlation observed here. Asset ownership may suffer from a similar phenomenon: land and house ownership can reflect culturally specific patrilineal inheritance patterns rather than women's genuine economic autonomy.

XGBoost and Gradient Boosting: Novel Contributions

This study is the first to apply XGBoost and Gradient Boosting classifiers to state-level classification of the burden of anaemia in India. XGBoost's superior 5-fold cross-validated accuracy (78.7%) relative to all other classifiers in the socioeconomic-only model reflects its capacity to capture higher-order non-linear interactions and feature combinations that linear methods (Logistic Regression) and single-learner methods (Decision Tree) miss. The gradient boosting framework's iterative error-correction mechanism is particularly well-suited to small, structured datasets in which individual predictors have moderate effect sizes.

The high AUC-ROC values achieved across classifiers in the socioeconomic-only model (Random Forest: 0.978; Logistic Regression: 0.933; XGBoost: 0.911) are particularly noteworthy, as AUC-ROC is threshold-independent and assesses overall discriminative capacity. These values confirm that the socioeconomic empowerment variables collectively provide strong rank-ordered discrimination between high- and low-anaemia-burden states, even when binary accuracy is limited by small sample size and class-boundary ambiguity in the intermediate range.

High-Burden States and Policy Targeting

The five highest-burden states — West Bengal, Tripura, Assam, Jharkhand, and Gujarat — share a profile of relatively lower women's internet access and, in most cases, lower educational attainment among women. These states should be the primary targets for intensified, integrated interventions that combine nutritional supplementation (iron-folic acid and multi-micronutrient supplementation), adolescent girl programs, strengthening of antenatal care, and digital literacy initiatives. The northeastern states (Nagaland, Manipur, Mizoram), despite modest economic development, demonstrate substantially lower anaemia prevalence, warranting qualitative investigation into the protective cultural, dietary, or healthcare factors specific to these populations.

Policy Implications in the Context of Post-NFHS-5 Interventions

The present findings should also be interpreted in light of the policy initiatives introduced or strengthened following the release of the NFHS-5 survey, which highlighted the unexpected increase in anaemia prevalence among Indian women despite ongoing national programmes. In response, the Government of India reinforced the Anaemia Mukt Bharat (AMB) strategy by enhancing programme monitoring, intensifying iron and folic acid supplementation, expanding testing and treatment protocols, and achieving greater convergence with POSHAN 2.0 and the Saksham Anganwadi initiatives. These programmes emphasise nutrition-sensitive interventions alongside nutrition-specific measures, including dietary diversification, fortified food distribution, adolescent health services, maternal nutrition support, and behaviour change communication. The present analysis provides empirical support for the complementary use of these biomedical interventions with broader socioeconomic strategies. Specifically, the strong associations observed between women's literacy, educational attainment, internet access, and reduced anaemia burden suggest that future anaemia control programmes may benefit from integrating digital health education, women's education, and community empowerment initiatives within existing national programmes. States identified as high-burden regions, particularly West Bengal, Tripura, Assam, Jharkhand, Gujarat, Odisha, and Bihar, may require state-specific implementation strategies that combine nutritional supplementation with interventions aimed at improving women's educational opportunities, digital inclusion, and access to reliable health information. Such integrated approaches are likely to enhance the long-term effectiveness and sustainability of national anaemia reduction efforts.

Limitations

Several limitations should be acknowledged. First, the small state-level sample size (n = 29) constrains statistical power and limits the complexity of machine learning models that can be reliably estimated. All reported results should be interpreted in this context, and replication with district-level data (n > 600) would substantially strengthen causal inference. Second, the cross-sectional design precludes causal inference: the observed empowerment-anaemia associations are consistent with, but do not prove, directional effects of socioeconomic development on anaemia outcomes. Third, important confounders — including dietary patterns, environmental iron exposure, hemoglobinopathy prevalence, and healthcare utilization rates — were not included in the analytical model. Fourth, the binary outcome operationalization (median split) involves information loss relative to continuous regression modeling.

CONCLUSIONS AND POLICY IMPLICATIONS

This study provides comprehensive, multi-method evidence of a robust socio-developmental gradient that structures the burden of anaemia across India's 29 states and union territories. The convergent findings from PCA, correlation analysis, signal complexity analysis, and a six-classifier machine learning pipeline collectively establish the following conclusions:

- Women's literacy (Spearman $r = -0.585$, $p = 0.001$), internet access (Spearman $r = -0.651$, $p < 0.001$), and educational attainment (Spearman $r = -0.588$, $p = 0.001$) are the strongest modifiable socioeconomic predictors of state-level anaemia prevalence.
- A dominant socio-developmental gradient (PC1, 58.52% of variance) aligns Indian states along a continuum from high empowerment/low anaemia (Kerala, Nagaland, Mizoram) to low empowerment/high anaemia (Bihar, West Bengal, Jharkhand).
- XGBoost achieves the highest 5-fold cross-validated accuracy (78.7%) and AUC-ROC of 0.911 using socioeconomic indicators alone, while Random Forest demonstrates the best Leave-One-Out stability (75.9%), collectively confirming the significant predictive power of modifiable structural variables.
- Financial inclusion and asset ownership indicators show negligible independent association with anaemia burden, suggesting that the current framing of banking access as a health determinant may be overstated in the Indian context.
- East and central Indian states (West Bengal, Tripura, Assam, Jharkhand, Gujarat, Odisha, Bihar) require urgent prioritization for integrated, multi-pronged anaemia interventions.

The policy implications of these findings are clear and actionable. Investment in women's digital literacy programs — which concurrently expand internet access, educational opportunities, and access to health information — should be recognized as a structural anaemia prevention strategy, not merely a technology or empowerment program. Adolescent girls represent the highest-burden subgroup and the population most amenable to educational intervention before the compounding demands of reproduction further deplete nutritional reserves. State-specific rather than national-average anaemia targets, particularly for pregnant women's programs (given the high local irregularity identified in the entropy analysis), will lead to more effective resource allocation.

REFERENCES

International Institute for Population Sciences (IIPS) & ICF. National Family Health Survey (NFHS-5), 2019–21: India. Mumbai: IIPS, 2021. http://rchiips.org/nfhs/

Puri S, Yadav K, Kant S, Rai S, Joe W. Digital monitoring of anaemia control measures in a district using Anemia Mukt Bharat Health Management Information System indicators. Cureus. 2024;16(12):e75248.

Kumari R, et al. Unveiling the burden of anaemia: Socio-economic and demographic factors affecting young women in India. Discover Public Health. 2025;22(603).

Balarajan YS, Fawzi WW, Subramanian SV. Changing patterns of social inequalities in anaemia among women in India. BMJ Open. 2013;3(3):e002233.

Kumar P, Sharma H, Sinha D. Socio-economic inequality in anaemia. BMC Public Health. 2021.

Sharma H, Singh SK, Srivastava S. Socio-economic inequality and spatial heterogeneity in anaemia among children in India: Evidence from NFHS-4. Clinical Epidemiology and Global Health. 2020;8(4):1158–1171.

Sharif N, Das B, Alam A. Prevalence of anemia among reproductive women in different social groups in India. PLOS ONE. 2023;18(2):e0281015.

Chakrabarty M, Singh A, Singh S, Chowdhury S. Is the burden of anaemia among Indian adolescent women increasing? PLOS Global Public Health. 2023;3(9):e0002117.

Let S, Tiwari S, Singh A. Contextualising anaemia among reproductive women in West Bengal. In: Chouhan P, et al. (Eds.), Sexual and Reproductive Health of Women. Springer; 2025:41–57.

Lodha NA, et al. Unveiling the burden of anaemia: Socio-demographic and obstetric determinants among pregnant women in Mehsana district, Gujarat. National Journal of Medical Research. 2025;15(01):50–56.

Tirore LL, et al. Prevalence and associated factors of severity levels of anemia among women of reproductive age in sub-Saharan Africa. Frontiers in Public Health. 2024;11:1349174.

Seiler J, et al. Mapping any, mild, moderate and severe anaemia in women of reproductive age in South Asia and sub-Saharan Africa. Communications Medicine. 2025;5:00765-2.

Breiman L. Random forests. Machine Learning. 2001;45:5–32.

Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794.

Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189–1232.

Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297.

Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society, Series B. 1958;20:215–242.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th IJCAI. 1995:1137–1145.

Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948;27:379–423.

Pincus SM. Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences USA. 1991;88(6):2297–2301.

Jolliffe IT. Principal Component Analysis (2nd ed.). Springer; 2002.

Pedregosa F, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

Virtanen P, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272.

World Health Organization. Haemoglobin Concentrations for the Diagnosis of Anaemia and Assessment of Severity. WHO/NMH/NHD/MNM/11.1. Geneva: WHO, 2011.

Ayalew M, Mamo T, Belew MA, Wale AD. Predictors of anemia in Ethiopia: A systematic-review of machine learning approaches. medRxiv. 2025.

Zemariam AB, et al. Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia. Scientific Reports. 2024;14(1):9080.

Bentley ME, Griffiths PL. The burden of anemia among women in India. European Journal of Clinical Nutrition. 2003;57(1):52–60.

Yang R, et al. Trends in socioeconomic inequalities and prevalence of anemia in low- and middle-income countries. JAMA Network Open. 2018;1(5):e182899.

Lakew Y, Biadgilign S, Haile D. Anaemia prevalence and associated factors among lactating mothers in Ethiopia. BMJ Open. 2015;5:e006001.

Sappani M, et al. Trends of severe and moderate anaemia among women in India. PLOS ONE. 2023;18(6):e0286464.

ABOUT THE AUTHOR

Sreejani Bhaduri is a high school junior at National Public School, Koramangala, in Bengaluru, India. She is interested in economics, public health, and nutrition, and hopes to explore how these areas intersect to improve the lives of women and children, particularly in India. She plans to study economics in college, along with business management. Sreejani Bhaduri has also placed in AIR 3 in the CBSE Indian National Board exam that happened this year, scoring 99.6% overall.

Dr. Debnarayan Khatua is an Associate Professor at RV University, Mysuru, Karnataka. He received his Ph.D. in Mathematics from the National Institute of Technology Durgapur, India. He has authored more than 40 research articles in reputable SCI/Scopus-indexed journals, contributed to international conferences and edited volumes. His citation count is 333, and his h-index is 11 (Source: Google Scholar). His current interests include explainable AI, deep learning, computational epidemiology, mathematical optimization, and interdisciplinary applications of mathematics in healthcare and engineering.