



Regular Article

Factors influencing global life expectancy across different income levels using Shapley value decomposition

Levente Sándor Nádas^a, Gábor Petneházi^b, Sándor Kovács^{c,*}, Bernadett Aradi^b^a University of Debrecen, Faculty of Economics and Business, Institute of Economics, Hungary^b University of Debrecen, Faculty of Informatics, Department of Applied Mathematics and Probability Theory, Hungary^c University of Debrecen, Faculty of Economics and Business, Coordination and Research Centre for Social Sciences, Hungary

ARTICLE INFO

JEL classification:

C45
I15
J11
O15

Keywords:

Life expectancy
Machine learning
Light gradient boosting machine (LGBM)
SHAP values
Health convergence
Socioeconomic determinants

ABSTRACT

The convergence of life expectancy across nations is a pivotal subject of inquiry in global health economics. Traditional econometric approaches have difficulty precisely allocating the relative contributions of multiple interacting variables, especially in nonlinear or complex models, which the authors identify as the primary limitation in existing literature. This paper addresses this limitation by using the Shapley value decomposition (SHAP) alongside a Light Gradient Boosting Machine (LGBM) to examine the key factors influencing life expectancy at birth across countries with different income levels. The LGBM model significantly outperformed a traditional linear regression model (R^2 of 0.901 vs. 0.716; MAPE of 2.08% vs. 3.38%). The SHAP analysis revealed that importance of predictive factor varies considerably by income level. For instance, the fertility rate is the most dominant factor in low-income countries, while health expenditure per capita is the most influential variable in high-income regions. The study concludes that model-informed policy considerations must be tailored to the specific economic context of a country, prioritizing basic infrastructure improvements in lower-income nations while focusing on advanced healthcare investments in wealthier ones.

1. Introduction

Global health trajectories have long been characterized by a dual narrative. While aggregate life expectancy has increased worldwide, the disparity between the countries with the longest and shortest life expectancies persists. This has sustained an extensive debate about convergence across nations, which is often linked to the diffusion of medical technology and public health knowledge (Skinner & Staiger, 2015). However, the macroeconomic consequences of this increase are complex. For example, some studies suggest that rapid mortality shocks in the mid-20th century actually led to lower GDP per capita growth rates in certain contexts (Hansen & Lonstrup, 2015). The determinants of life expectancy at birth differ between richer and poorer countries. Preston (1975) demonstrated a non-linear relationship between life expectancy and income, which more recent studies confirm is nuanced and highly inelastic at higher income brackets (Okunade & Osmani, 2020). At lower income levels, the curve is steeper, primarily because reducing infant and child mortality yields substantial gains to average life expectancy, whereas marginal gains from reducing adult mortality

in wealthier nations are smaller (Samarawickrama et al., 2025).

Life expectancy at birth is one of the most widely used indicators of a population's overall health and well-being. It reflects the quality of healthcare systems, as well as broader social, economic, and environmental conditions (Irandoost et al., 2025). However, the factors that shape life expectancy vary significantly between countries of different income levels (Fan & Bo, 2025). High-income countries benefit from advanced medical technology and comprehensive social protection systems, while low- and middle-income countries often struggle with limited healthcare access, poverty, and inadequate infrastructure (Tu et al., 2025). Therefore, identifying the main predictive determinants of life expectancy across income groups is essential for developing effective policies that promote global health equity (Dutta & Dhar, 2024). These results could help design evidence-based, context-specific policy interventions that account for the unique constraints of developing versus developed nations.

* Corresponding author. University of Debrecen, Faculty of Economics and Business, Coordination and Research Centre for Social Sciences, Hungary

E-mail addresses: nadas.levente@econ.unideb.hu (L.S. Nádas), gabor.petnehazi@science.unideb.hu (G. Petneházi), kovacs.sandor@econ.unideb.hu (S. Kovács), aradi.bernadett@inf.unideb.hu (B. Aradi).

<https://doi.org/10.1016/j.ssaho.2026.103704>

Received 30 November 2025; Received in revised form 18 September 2026; Accepted 18 September 2026

Available online 25 September 2026

2590-2911/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1.1. Research gap and motivation

Traditional econometric approaches, such as panel regression, have been widely used to quantify the impact of various factors like GDP, healthcare spending, and education. However, these methods often struggle to accurately allocate the relative contributions of multiple interacting variables, particularly in nonlinear or complex models. This paper addresses this limitation by using the Shapley value decomposition method to examine the factors that influence life expectancy at birth in countries with different income levels. The study highlights the interaction between economic development, healthcare investment, education, and lifestyle factors in predicting population health outcomes (Kousar et al., 2023). Unlike standard feature importance metrics, SHAP enables a game-theoretic decomposition of life expectancy, revealing how the marginal contribution of specific predictors, such as health expenditure, shifts dynamically based on a country's specific development context. This approach's versatility is increasingly recognized in medical and health literature, where SHAP has been successfully used to explain survival predictions in specific clinical scenarios. For example, it has been used to predict life expectancy after thoracic surgery for lung cancer patients (Ali et al., 2024).

Existing literature often struggles to allocate relative contributions in complex, nonlinear models. This study addresses this gap by pursuing four primary objectives. First, it aims to develop a high-precision predictive model for life expectancy using a Light Gradient Boosting Machine (LGBM), which will form the basis for evaluating theoretical expectations H2 and H3. Second, to use Shapley value decomposition (SHAP) to interpret the model and precisely allocate the contribution of interacting variables (addressing theoretical expectation H1). Third, to examine how the importance of predictors (e.g., health expenditure, fertility) varies across different income levels and regions (examining the alignment with H4). Fourth, to derive model-informed context-specific policy recommendations based on the heterogeneous predictive contributions of these factors. The theoretical expectations (H1-H4) of the current study are defined in the literature review. Thus, this study bridges predictive modeling, explanatory analysis, and policy formulation.

This study contributes to the existing literature by integrating the high-dimensional predictive capacity of gradient boosting with the game-theoretic interpretability of SHAP values. Unlike traditional linear models, which often require manual specification of interaction terms and struggle to represent nonlinear patterns, our framework automatically captures these complex dependencies. This allows for a granular decomposition of life expectancy predictors, distinguishing between global drivers and context-specific predictors across different income levels. Unlike, traditional econometric approaches, which prioritize hypothesis testing via p-values and confidence intervals, this study employs a machine learning framework that focuses on minimizing predictive error and extracting non-linear feature importance.

1.2. Research objectives and contributions

The primary motivation of this study is to overcome the limitations of traditional econometric models in predicting life expectancy, specifically their inability to accurately allocate the relative contributions of interacting, non-linear socioeconomic and health determinants across different income levels.

This study makes four key contributions to the existing literature.

- **Predictive modeling:** We integrate the robust predictive capacity of LightGBM with comprehensive global dataset coverage and explicit temporal lag modeling to automatically capture complex, nonlinear dependencies, providing the high-precision empirical foundation necessary to evaluate the alignment of our results with theoretical expectations (H1–H4).

- **Game-theoretic interpretability:** We utilize SHAP values to unlock the “black box” of traditional ML, allowing us to identify the primary predictive features of longevity (H1), the predictive significance of historical socioeconomic conditions (H2), and the negative predictive association of specific health risk factors (H3).
- **Context-specific insights:** We provide a granular breakdown of life expectancy predictors, specifically examining how the importance of these factors is moderated by a country's specific income level (H4).
- **Policy relevance:** We bridge the gap between prediction and policy by translating SHAP importances into tailored, model-informed policy recommendations.

2. Literature review

The existing literature extensively uses econometric and statistical methods to identify the drivers of life expectancy, placing a strong emphasis on economic and health expenditures. Both early and contemporary studies often use panel data methods to model these relationships. For example, Jaba et al. (2014) examined the relationship between per capita health expenditures and life expectancy in 175 countries from 1995 to 2010. These models are often based on the health production function, which treats variables such as income per capita, education, and urbanization as critical inputs for health status (Bayati et al., 2013). Building on this, Ray and Linden (2020) examined the impact of public and private health expenditures on life expectancy at birth and infant mortality in 195 countries from 1995 to 2014. Similarly, longitudinal analyses of specific regions, such as EU accession candidate countries, have reaffirmed that higher GDP per capita and lower infant mortality rates are fundamental prerequisites for increased longevity (Miladinov, 2020).

Beyond health spending, researchers consider a broader range of socioeconomic and environmental factors. For example, Nica et al. (2023) examined the effects of financial development, health expenditure, CO₂ emissions, institutional quality, and energy consumption on life expectancy in Eastern European countries from 1990 to 2021. Using advanced methods (CS-ARDL and quantile regression), the researchers found that increased health spending, renewable energy consumption, and institutional quality improved life expectancy. Meanwhile, CO₂ emissions and fossil fuel consumption had negative effects. These findings align with the prevailing literature (Țarcă et al., 2024; Neofytidou & Fountas, 2020; Ray & Linden, 2020). Furthermore, empirical studies of 138 countries have demonstrated that access to electricity and physician density are statistically significant positive predictors of life expectancy (Mourad, 2021). Interestingly, some evidence suggests that social and behavioral factors, particularly educational attainment, may actually explain a much larger portion of health outcome variance than medical care expenditures alone (Kaplan et al., 2014). Gazilas (2024) conducted a regional analysis of low-income countries from 1995 to 2020. Using fixed-effects panel regression, this study corroborated previous findings, showing that higher GDP per capita, better education, and increased access to healthcare significantly improve life expectancy, while conflict and weak institutions reduce it. Recent studies on OECD countries further highlight that the interplay between physician density, hospital bed availability, and social spending remains a primary driver of longevity in developed health systems (Roffia et al., 2023). Although panel regression is the dominant method in the field, more complex machine learning (ML) methods are increasingly being used for prediction and determining importance of variables. For example, Lipesa et al. (2023) applied supervised ML to predict life expectancy using global health and socioeconomic data for 193 countries. They found that Random Forest and Gradient Boosting models performed best, with key predictors being GDP per capita, infant mortality, and access to healthcare. Similarly, Alvi et al. (2025) demonstrated the effectiveness of recursive partitioning and bagging algorithms on World Health Organization (WHO) data, achieving high predictive accuracy and identifying adult mortality and HIV/AIDS prevalence as critical predictive

characteristics. In regional context, [Țarcă et al. \(2024\)](#) used panel econometric methods to analyze determinants in Eastern Europe. Their finding reinforce the idea that income level, healthcare spending, and education positively influence life expectancy, while unemployment and alcohol consumption negatively affect it.

A crucial aspect of the research focuses on income inequality and how socioeconomic status shapes longevity gaps. For example, [Schwandt et al. \(2022\)](#) examined the relationship between income and life expectancy in California. They demonstrated that the COVID-19 pandemic significantly widened disparities, disproportionately affecting lower-income and minority populations. Similarly, [Walczak et al. \(2021\)](#) examined this link in OECD countries, emphasizing the policy challenges for pension systems given that wealthier groups tend to live longer. [Wang \(2021\)](#) conducted a comprehensive global analysis using both statistical analysis and Machine Learning (ML), confirming that GDP per capita, healthcare access, and education are the strongest positive predictors across continents. However, these standard ML applications largely act as ‘black boxes,’ lacking granular interpretability regarding how predictors interact. Our study differs by utilizing SHAP to dynamically decompose these non-linear interactions across different income levels.

Based on previous literature, the following theoretical expectations (H1–H4) can be formed.

- **H1:** Health expenditure per capita is the most significant predictive feature of life expectancy in various socioeconomic contexts.
- **H2:** Lagged socioeconomic variables, such as access to electricity and GDP per capita, exhibit significant long-term predictive impacts on life expectancy predictions.
- **H3:** Fertility rates and tuberculosis incidence are negatively correlated with life expectancy, with higher values leading to lower predicted life expectancy.
- **H4:** The predictive importance of socioeconomic and health-related variables on life expectancy is significantly moderated by the country's income level.

The primary limitation of most statistical models is the difficulty of interpreting the relative contributions of numerous interdependent variables. Recent methodological advances address this gap by utilizing concepts from cooperative game theory. [Bimonte et al. \(2024\)](#) developed an ensemble approach to mortality modeling using Shapley value methods. The present study therefore assigns contribution weights based on Shapley values, evaluating the relative importance of each model in improving prediction accuracy.

3. Methodology

3.1. Theoretical framework

This research is grounded in the socioeconomic determinants of health, which posit that life expectancy is driven by the multifaceted interplay of economic growth, healthcare investment, and infrastructural development. Although traditional econometric models, such as the Preston curve, have established the fundamental link between income and health, they often fail to capture the complex, nonlinear interactions and heterogeneous effects inherent in global demographic data. Consequently, this study uses a machine learning approach with the Light Gradient Boosting Machine (LGBM) to effectively model high-dimensional dependencies and the lagged impacts of historical conditions on current health outcomes. Inclusion of SHAP values resolves the “black box” limitations of such advanced algorithms by providing a game-theoretic basis for interpretability and allowing for the precise quantification of how variable importance shifts across different contexts. Additionally, the analysis incorporates structural development theories, which suggest that the efficacy of determinants such as health expenditure or fertility rates is significantly moderated by a nation's

specific income level. This theoretical synthesis provides a more nuanced understanding of global health disparities by moving beyond static, linear assumptions to dynamic, context-specific predictive modeling.

3.2. Data description and preparation

Based on an extensive literature review, we identified a broad set of factors that could influence life expectancy at birth. Our goal was to include a wide range of variables to allow the machine learning model to select the most relevant ones for longevity. All data were sourced from the World Development Indicators (WDI) database ([World Bank, 2025a](#)). The dependent (target) variable is ‘life expectancy at birth, total (years)’ (SP.DYN.LE00. IN). The major inference behind the selection of the independent variables is grounded in the health production function literature, specifically the human capital framework established by [Grossman \(1972\)](#), which treats macroeconomic indicators, education, and health system capacity as fundamental inputs for population health. We intentionally expanded this theoretical framework to include broader socio-environmental risk factors (e.g., PM2.5 exposure, basic sanitation), drawing on the relationship between income and health outcomes first described by [Preston \(1975\)](#). By mapping each predictor to these established theoretical domains—ranging from direct medical inputs to distal environmental determinants—we ensure that the machine learning model detects diverse, multi-dimensional inferences within a robust economic context. This allows the machine learning model to detect diverse, multi-dimensional inferences without strict a priori constraints. The potential predictors span demographic, economic, social, environmental, and health-related domains and are detailed in [Table 1](#).

We deliberately excluded direct mortality indicators (e.g., crude death rates) as explanatory variables. Since life expectancy is derived from mortality rates, and including mortality as a predictor introduces the risk of tautological reasoning. Instead, we focused on the underlying

Table 1
Potential influencing factors of life expectancy.

Theoretical Domain	Variables & Descriptions	Theoretical Inference/Source
Economic & Infrastructure	GDP_PC_PPP (GDP per capita, PPP); AccessElectricity (% of population); InternetUsersPct (% of population); EmploymentAgriculturePct (% of total); UnemploymentRate (% of total labor force)	Income-Health Link (Preston, 1975). Macroeconomic stability and infrastructure availability.
Health System & Spending	HealthExp_PC (Current health expenditure per capita, US\$); HealthExp_GDP_Pct (% of GDP); HospitalBeds_per_1000; MeaslesImmunization (% of children)	Health Capital Theory (Grossman, 1972). Direct investment in medical services and preventive care.
Demographics	FertilityRate (births per woman); PopulationAge65PlusPct (% of total); PopulationTotal; UrbanPopulationPct (% of total); PopulationDensity (people per sq. km)	Demographic Transition Theory. Reflects dependency ratios and shifts in population structure.
Health Status & Risk Factors	TuberculosisIncidence (per 100 k); AlcoholConsumption (liters per capita); OverweightPrevalence (% of adults); PrevalenceUndernourishment (% of population); HomicideRate (per 100 k)	Epidemiological Transition. Measures biological, behavioral, and external threats to survival.
Social Development & Environment	WaterBasicAccess (% of population); SanitationBasicAccess (% of population); AccessCleanCooking (% of population); PM25_Exposure (mean annual exposure); FemalePrimaryCompletionRate; SchoolEnrollment (Primary/Secondary % gross)	Social Determinants of Health. Distal factors like education and environment impacting health behavior.

socioeconomic, health care, environmental, institutional, and behavioral variables that drive these mortality patterns (Samarawickrama et al., 2025). Data were collected for all available countries and years starting from 1990 to 2023 (a 34-year raw period). However, because we generate up to 8-year lag variables, the earliest years for each country inherently lack historical data and produce missing values (NaNs). After dropping rows with these resulting NaNs to ensure a complete feature matrix, the effective analytical dataset after pre-processing covers approximately 15 years of continuous data for 125 countries. Due to varying initial data availability across nations, the resulting dataset is unbalanced. All the data were then combined into a single panel dataset.

A correlation heatmap of our dataset is available in Fig. A1 in Appendix A. Descriptive statistics are available in Table A2 in Appendix A. Income classifications were applied statically based on the currently available World Bank data to ensure consistent group comparisons throughout the analysis period. While static classification simplifies group comparisons, it represents a minor limitation as it may misclassify transitioning economies that shifted between income groups during our raw 34-year study period (1990–2023). Due to the longitudinal nature of the data, missing values were imputed using forward-filling (Cicek & Aydin, 2025). Although forward-filling is a standard panel imputation technique, we acknowledge that it can partially smooth out or obscure short-term volatility caused by sudden economic or health crises. This specific imputation strategy was chosen to preserve the longitudinal integrity of the panel data and to prevent future data leakage, which would occur if interpolation methods using future data points were applied. Rows with missing data points remaining after these steps were dropped to maintain dataset integrity. Since tree-based models like LGBM are invariant to monotonic transformations, no feature scaling or normalization was applied to the predictors.

To model various temporal dynamics, lagged values (at two, four, and eight years) were created for all predictors to capture delayed effects on life expectancy (De Arcos-Jimenez et al., 2025). These lags were generated by shifting temporal data within each country group, ensuring no data leaked across different nations. Crucially, to prevent data leakage and ensure rigorous validation, all data preprocessing operations—including missing value forward-filling, temporal lag feature creation, and hyperparameter optimization via Optuna—were fitted and executed strictly within the pre-2020 training split prior to evaluating unseen out-of-sample test data (2020–2023).

This approach is critical for time-series panel data as it strictly prevents information leakage from the future into the past, ensuring a realistic assessment of the model's forecasting ability. To quantify the uncertainty of our out-of-sample model performance, we applied a non-parametric bootstrapping approach with 1000 resampling iterations to calculate 95% confidence intervals for our evaluation metrics. Such robust validation techniques are necessary, as research comparing advanced modeling approaches (e.g. Bayesian dynamic linear models versus mixed effects models) has shown that the choice of methodology significantly alters the accuracy of mortality-related analyses (George et al., 2025). The final dataset comprises 125 countries, yielding a total of 1528 observations (1044 in the training set and 484 in the test set). We also incorporated World Bank income classifications (low, lower-middle, upper-middle, and high income) into the analysis to evaluate model performance and feature importance in various economic contexts (World Bank, 2025b).

We applied a pooled regression approach, combining data from different years and countries into a single dataset. We only used the predictors in Table 1 and their lagged values as input variables. This approach enabled us to work with a fairly large dataset and identify fundamental, universal drivers of life expectancy. Using gradient boosting (as detailed in the next section) we can capture complex, nonlinear patterns without further feature engineering. Model training and hyperparameter tuning were executed on a machine equipped with 48,0 GiB RAM and an 11th Gen Intel® Core™ i5-11400 F @ 2.60 GHz ×

12 CPU. The runtime for optimization and SHAP value computation was no more than a few minutes. The Python code necessary to reproduce this analysis is publicly available at [https://github.com/\[anonymized\]/life-expectancy-shap](https://github.com/[anonymized]/life-expectancy-shap).

3.3. Modeling approach

We used the Light Gradient Boosting Machine (LightGBM) framework (Ke et al., 2017) to predict life expectancy. This ensemble method efficiently handles nonlinearities and complex interactions without the need for manual specification of interaction terms, which makes it suitable for our high-dimensional dataset. The model learns generalized patterns by pooling data across countries and years. While pooled modeling maximizes sample size and allows gradient boosting to learn global non-linear dependencies across income-spectrums, it assumes baseline parameter homogeneity, which can partially smooth over unobserved country-specific structural characteristics and localized shocks. To optimize the model's performance, we used the Optuna Python package (Akiba et al., 2019), which employs sequential model-based optimization to efficiently search the hyperparameter space. The key hyperparameters that were tuned included the learning rate, the number of leaves, the feature and bagging fractions (for subsampling to reduce overfitting), and the regularization terms. We used the mean absolute error (MAE) on the training set was used as the objective function and conducted 50 trials to identify near-optimal settings. The final set of optimal hyperparameters used to train the reported model is explicitly listed in Table 2. We ran the experiments using Python 3.10.12, LightGBM 4.6.0, SHAP 0.47.2, and Optuna 4.4.0. Random seeds were fixed to 42 to ensure deterministic reproducibility.

A key advantage of this tree-based ensemble method is its inherent ability to naturally handle multicollinearity and automatically detect complex interactions between variables, which reduces the risk of omitted variable bias that is common in linear models without explicit interaction terms.

3.4. Model validation and interpretation

We assessed model performance on the reserved validation set (2020–2023) to evaluate its predictive accuracy on unseen, future data. To address the research gap identified in the literature review regarding interpretability, we used SHapley Additive Explanations (SHAP) values to understand the contribution of each factor to the model's predictions. The concept of the Shapley value was first introduced in the field of

Table 2
The optimized LGBM parameters as used in the study.

Hyperparameter	Description	Optuna Search Space	Final Value
learning_rate	Shrinkage rate of tree weights	[0.01, 0.20] (Float)	0.0267
num_leaves	Maximum tree leaves for base learners	[20, 200] (Integer)	48
max_depth	Maximum tree depth	[3, 16] (Integer)	8
feature_fraction	Subsampling ratio of columns per tree	[0.6, 1.0] (Float)	0.8308
bagging_fraction	Subsampling ratio of rows per tree	[0.6, 1.0] (Float)	0.6109
bagging_freq	Frequency of bagging	[1, 10] (Integer)	2
min_child_samples	Minimum data points in a leaf	[5, 100] (Integer)	12
lambda_l1	L1 regularization term on weights	[0.0, 5.0] (Float)	0.8143
lambda_l2	L2 regularization term on weights	[0.0, 5.0] (Float)	0.9279

Note. The model was optimized over 50 trials to minimize the Mean Absolute Error (MAE) using 10,000 estimators with an early stopping round of 100.

game theory by Lloyd Shapley in 1953 (Shapley, 1953). For his work in identifying this quantity, Shapley was later awarded a Nobel Prize. It is useful for determining fair credit given to individual players based on their contributions across all possible coalitions. Lundberg and Lee (Lundberg & Lee, 2017) were the first to apply the Shapley value to machine learning; they used it to determine the importance of a feature in a particular prediction. In this case, each feature is treated as a player in a game where the payout is the model's prediction.

Given an input x , the i th feature's Shapley value is calculated by the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

where N is the set of all features, S is a subset of features that does not contain the considered feature (the sum runs over all possible combinations), and $f(S)$ denotes the model's prediction when only the features in S are present (with all other features marginalized). Thus, ϕ_i is the weighted average of the feature's marginal contribution to a prediction. It is critical to emphasize that SHAP values represent marginal contributions to the model's specific predictive output rather than direct real-world causal effect sizes or indicators of policy effectiveness.

In general, calculating the exact Shapley values (SHAPs) is computationally expensive since it requires considering all possible combinations of features. However, when working with tree-based algorithms, such as LightGBM, we can use the TreeExplainer method from the SHAP Python Package. This is a highly optimized, computationally superior modification of the SHAP method specifically designed to compute exact Shapley values for tree ensembles like LightGBM in polynomial time. We computed SHAP values for the entire dataset to allow for both global and income-group-specific aggregation. We used SHAP to produce both local and global explanations. Local interpretability is guaranteed by the following formula:

$$f(x) = \phi_0 + \sum_{i=1}^{|N|} \phi_i(x) \quad (2)$$

where x is an instance and $\phi_0 = E[f(x)]$ is the expected value of the model output. Thus, the Shapley value can be considered as an additive feature importance measure that divides the model's prediction among its features. In our investigation, we estimated each feature's contribution to the predictions of interest for individual country-years. We also aggregated the SHAP values into global feature importances in order to gain insight into the model's overall behavior.

3.5. Workflow description

To provide a comprehensive overview of the analytical process, Fig. 1 illustrates the end-to-end research workflow employed in this study. This methodological sequence begins with the Data Acquisition and Processing stage, where raw indicators from the World Development Indicators (WDI) database are cleaned, lagged, and split into temporal training and testing sets to strictly prevent data leakage. Following this, the Model Optimization phase utilizes the Optuna framework to fine-tune the LightGBM hyperparameters, ensuring maximum predictive precision. Once the optimal model is established, the interpretability Layer applies TreeExplainer-based SHAP decomposition to the results. This allows us to move from global performance metrics to granular, context-specific insights. Finally, the research concludes with Policy Synthesis stage, where the decomposed contributors are translated into income-specific recommendations.

Fig. 1 provides an end-to-end summary of the research workflow of the current study. The systematic structure of this workflow ensures that the transition from raw socioeconomic data to actionable policy insights is both transparent and reproducible. By separating the predictive modeling from the interpretability analysis, we maintain high accuracy

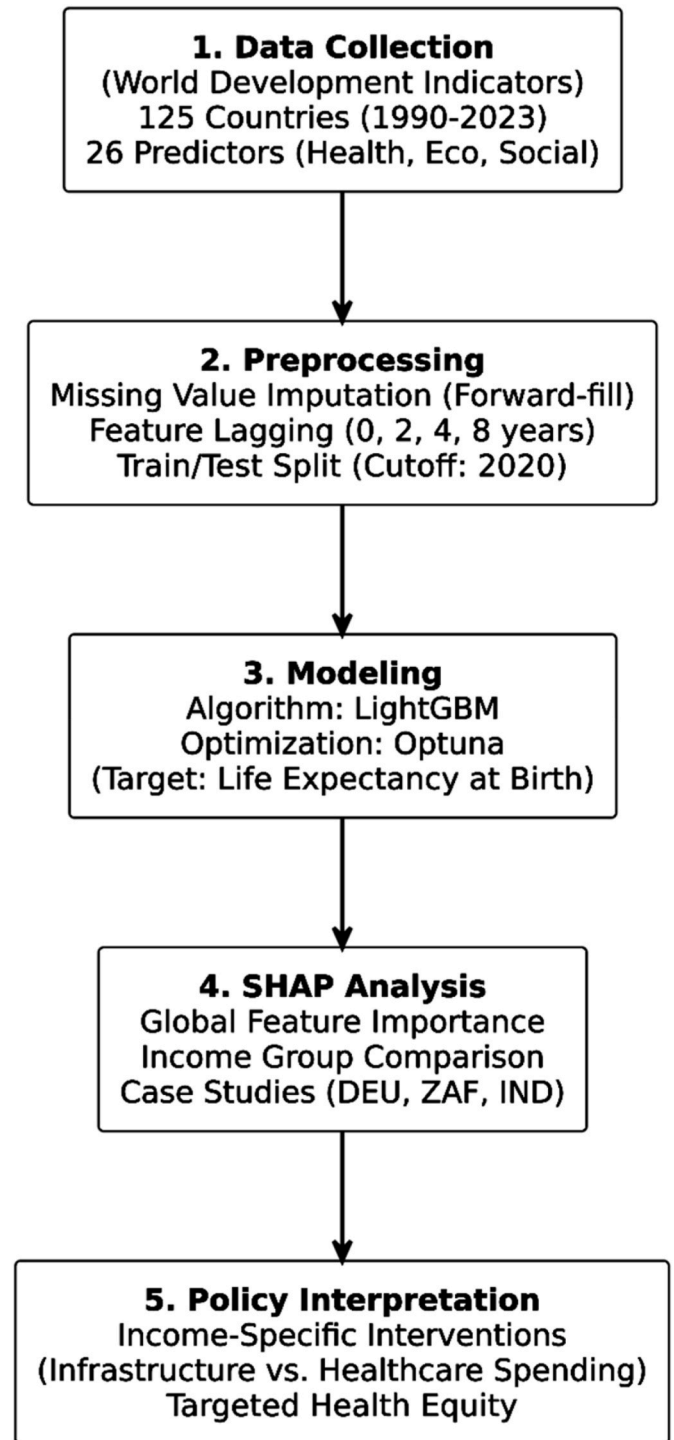


Fig. 1. Research workflow.

while overcoming the typical “black box” limitation of gradient boosting machines. This integrated approach allows for a robust evaluation of the life expectancy determinants, ensuring that the subsequent results (detailed in Section 4) are grounded in a rigorous, multi-stage analytical framework that accounts for both temporal dynamics and income-based heterogeneity.

4. Results

We used a Light Gradient Boosting Machine (LGBM) and, as a benchmark, a linear regression (LR) model to predict life expectancies

using all the input features in Table 1, with lags of 0, 2, 4, and 8 years. To rigorously answer requests for comprehensive evaluation, we expanded our reporting to include additional performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE), and R^2 , along with 95% bootstrapped confidence intervals (Table 3).

The LR model produced a Root Mean Squared Error (RMSE) of 3.316, and an R-squared (R^2) of 0.716. The LGBM model had an RMSE of 1.961, a Mean Absolute Error (MAE) of 1.498, and an R^2 of 0.901, accompanied by tight 95% confidence intervals. Specifically, the LightGBM model's MAE of 1.498 [95% CI: 1.381, 1.616] and R^2 of 0.901 [95% CI: 0.881, 0.917] demonstrate a substantial and statistically robust improvement over the Linear Regression baseline, which exhibited a higher MAE of 2.451 [95% CI: 2.263, 2.675] and a lower R^2 of 0.716 [95% CI: 0.623, 0.786]. Furthermore, evaluating relative and outlier-robust metrics underscores this performance advantage. The LightGBM model achieves a Mean Absolute Percentage Error (MAPE) of 2.1% [95% CI: 2.0%, 2.3%] compared to 3.4% [95% CI: 3.2%, 3.7%] for Linear Regression, indicating that LightGBM's predictions deviate on average by only approximately 2.1% relative to true life expectancy values. Additionally, LightGBM achieves a lower Median Absolute Error (MedAE) of 1.133 years [95% CI: 1.052, 1.264] compared to LR's 1.945 years [95% CI: 1.828, 2.232]. This lower MedAE indicates that typical individual predictions from LightGBM are off by barely over a single year and highlights that LightGBM's performance superiority remains consistent even when insensitive to extreme prediction outliers. The lack of overlap between the 95% bootstrapped confidence intervals across all metrics confirms that the machine learning model's predictive superiority is highly robust and statistically significant. This superior test set performance across all metrics suggests that the more complex machine learning model has a considerable and statistically robust advantage in terms of predictive power compared to the baseline. To rigorously confirm the performance advantage of the LightGBM model over the benchmark linear regression, we conducted statistical comparison tests on the out-of-sample squared errors. A paired *t*-test revealed a highly statistically significant difference in predictive accuracy ($p < 0.001$). Because error distributions can sometimes deviate from strict normality, we also applied a non-parametric Wilcoxon signed-rank test, which further confirmed the significance of the LightGBM model's superior performance ($p < 0.001$). Additionally, a visual comparison of both models' predictive performance across the test set is provided in Fig. A2 in Appendix A. The robustness of the model was further supported by the stability of the SHAP feature rankings observed across exploratory analyses using alternative data preprocessing strategies (e.g., different treatments of missing values). These analyses were conducted during preliminary experimentation but are not reported in this study.

Fig. 2 displays a scatterplot of the actual test set life expectancies against the predicted ones (from LGBM), with the ten largest deviations highlighted.

For our baseline linear regression, the contributions of the features to the predictions are straightforward additive effects, defined by the feature values multiplied by their corresponding coefficients. The raw coefficients of the linear regression are displayed in Table A1 in Appendix A. While the sign of each coefficient correctly indicates whether the impact is positive or negative, the absolute magnitude of the coefficients cannot be used to compare the relative importance of different features because of the wide variation in the scale and units of the input variables.

SHAP allows us to determine feature contributions in the LGBM

model across three distinct analytical levels to avoid visual redundancy: (i) feature-level directionality and distributions across specific lag horizons (Fig. 3), (ii) base-variable importance aggregated across all lag horizons (Fig. 4), and (iii) broad domain/category level contributions (Fig. 5). Fig. 3 specifically details the directional impact (positive/negative) and distribution of SHAP values for the top individual predictor-lag combinations.

The features are listed in order of average importance on the vertical axis (highest importance at the top). The horizontal position corresponds to the SHAP value for a given instance. Color coding reveals the actual feature value: red dots represent high feature values, and blue dots represent low values. According to Fig. 3, current health expenditure per capita (HealthExp_PC) is the most important predictor on average, followed closely by the current access to electricity (AccessElectricity). High feature values (red dots) consistently push the SHAP values into the positive range (to the right of zero). Conversely, low feature values (blue dots) push SHAP values into the negative range (to the left). These results indicate a strong positive correlation: higher health expenditure and greater electricity access are strongly associated with higher predicted life expectancy within the model. The FertilityRate and TuberculosisIncidence features demonstrate a negative correlation. High values (red dots) consistently result in negative SHAP values, which decrease predicted life expectancy. Low values (blue dots) contribute positively. This aligns with expected public health outcomes: higher disease burdens and high fertility rates (often a characteristic of lower-income countries with high infant mortality) are associated with lower overall longevity. The eight-year lag of GDP_PC_PPP is significant, with higher values contributing positively, as expected. Almost all health expenditure per capita variables are among the most important predictors, aligning strongly with economic theory regarding health investment returns. Several lagged variables appear near the top of the list, indicating that past values of the examined factors can also provide valuable assistance in predicting life expectancy by capturing long-term developmental impacts rather than just immediate effects.

Fig. 4 shows the aggregated importance of the base variables, or the importance summed over all lag horizons. A longer bar indicates a higher mean absolute SHAP value, meaning that the given feature group was, on average, more influential in determining the predicted life expectancy. Based on the SHAP values, health expenditure per capita is the most important indicator, followed by access to electricity and fertility rate. Features toward the bottom of the list, such as AccessCleanCooking and PM25_Exposure, have very low mean absolute SHAP values (nearly 0.00). This suggests that, although these variables were included in the dataset, the model identified them as the least influential or useful for making accurate predictions compared to the top variables.

The base variables (i.e., the distinct WDI used as input variables) were grouped into five categories, as indicated in Table 1. The goal was to create categories with roughly the same number of base variables. Fig. 5 displays the aggregated importance of the variable categories. Overall, economic and infrastructure variables are the most important, followed by health system and spending variables. Demographics, health status and risk factors, and social development and environment variables seem to be less important.

4.1. Country groups

To gain a deeper understanding, we grouped countries based on World Bank income classes and regions (World Bank, 2025b). Then, we analyzed the resulting SHAP importances within the groups.

Table 3
Out-of sample test set evaluation results.

Model	RMSE [95% CI]	MAE [95% CI]	MAPE (%) [95% CI]	MedAE (Years) [95% CI]	R^2 [95% CI]
Linear Regression	3.316 [2.877, 3.784]	2.451 [2.263, 2.675]	3.4 [3.2, 3.7]	1.945 [1.828, 2.232]	0.716 [0.623, 0.786]
LightGBM	1.961 [1.825, 2.098]	1.498 [1.381, 1.616]	2.1 [2.0, 2.3]	1.133 [1.052, 1.264]	0.901 [0.881, 0.917]

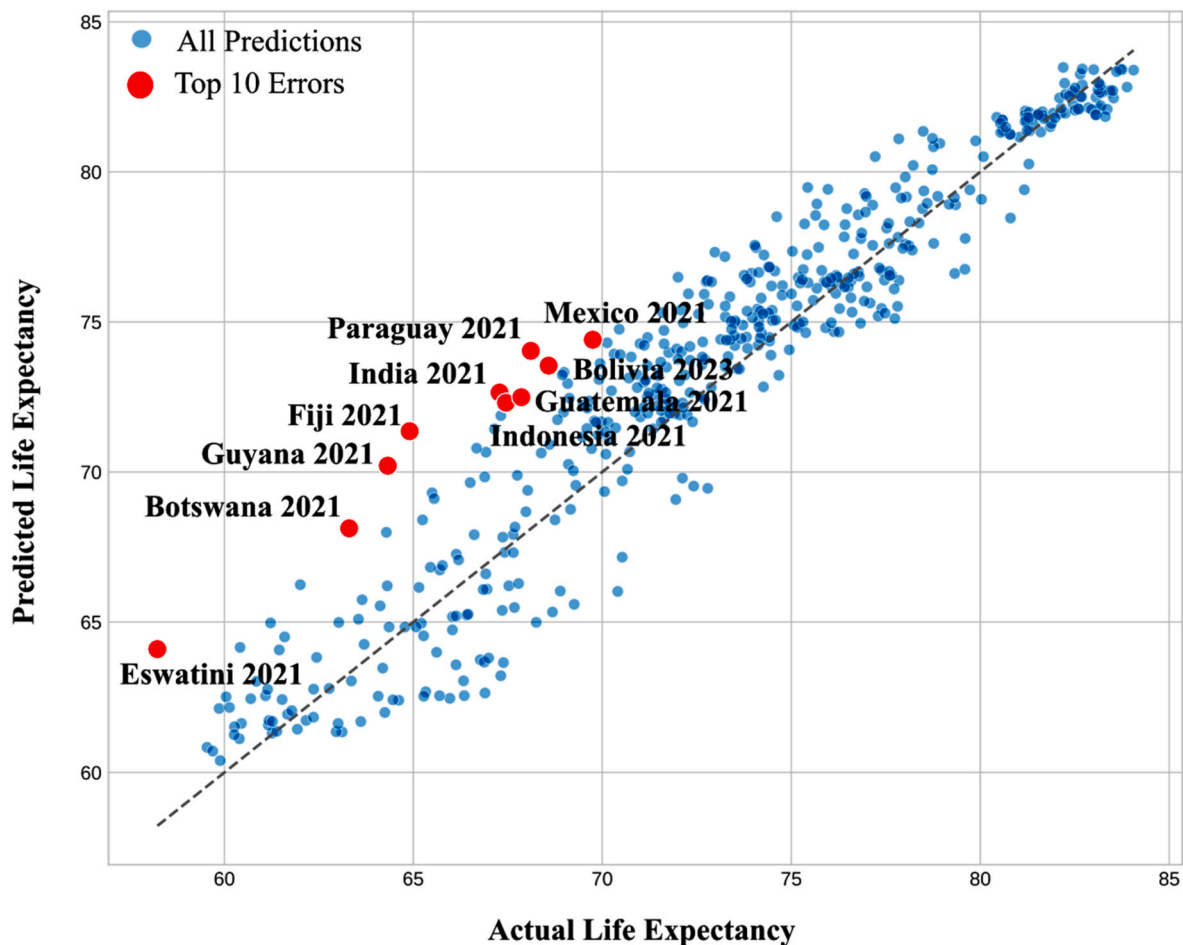


Fig. 2. LGBM test set predictions against true values with 10 largest errors highlighted.

Fig. 6 illustrates the relative importance of various socioeconomic and health-related variables at different income levels. The results reveal significant differences in the factors influencing the model's predictions across income groups, emphasizing the need for income-specific policies.

Among low-income countries (orange bars), access to electricity exhibits the highest mean absolute SHAP value, suggesting that fundamental infrastructure has dominant explanatory power in these contexts. Fertility rate is the second most important feature for low-income countries, emphasizing the importance of demographic factors. Health expenditure per capita is the most influential variable for high-income (blue bars) and upper-middle-income (red bars) countries, reflecting the importance of advanced healthcare systems in achieving marginal longevity gains in wealthier nations. Internet usage contributes moderately across most income categories, underscoring the universal impact of infrastructure and technology on outcomes. Although variables such as tuberculosis incidence and homicide rate have relatively low SHAP values overall, they are more important in certain income groups. The homicide rate is particularly important in high-income regions, while the incidence of tuberculosis is more important in non-high-income areas. These findings suggest that, as primary health concerns are addressed, specific issues, such as safety, become more influential in determining longevity in wealthier societies. Interestingly, the proportion of the population aged 65 and over modestly influences longevity, particularly in high-income regions. This is consistent with global demographic aging trends.

Fig. 7 shows the mean absolute SHAP values of the variable categories by income group. Considerable variation in SHAP importance exists across different groups of countries. Demographic, economic,

social and environmental factors seem to be more important predictors of life expectancy in less developed countries and regions. In more developed areas, factors related to the health system seem to be more important predictors.

Table 4 summarizes the top three predictors for each income level based on mean absolute SHAP values to facilitate a quick comparison of the dominant drivers across different economic contexts.

The model's predictive power varies substantially depending on the economic contexts. It generally performs best in high-income regions, where data quality is often higher and health trajectories are more uniform. Conversely, predictive accuracy is notably lower in upper-middle-income countries. This group comprises highly heterogeneous transition economies, including Latin American nations, Eastern European transition states, and Southern African countries. This structural diversity, coupled with distinct regional health crises, contributes to the reduced predictive power in this bracket. Table 5 shows the performance metrics by income level, including Mean Absolute Error (MAE) alongside R^2 and RMSE for a more comprehensive view of out-of-sample performance.

In order to systematically evaluate the predictive behavior and address the potential impact of omitting the HIV/AIDS prevalence variable, we first acknowledge that directly including an HIV/AIDS prevalence predictor was not feasible due to severe data limitations (substantial missing values in the World Bank WDI database across our global longitudinal panel), which would have greatly reduced the analytical sample size. To evaluate the effect of omitting this variable and other localized factors on the model, we performed a two-part sensitivity analysis on the upper-middle-income group. As Eswatini is classified as a lower-middle-income country by the World Bank, the only Southern African countries in our upper-middle-income dataset are

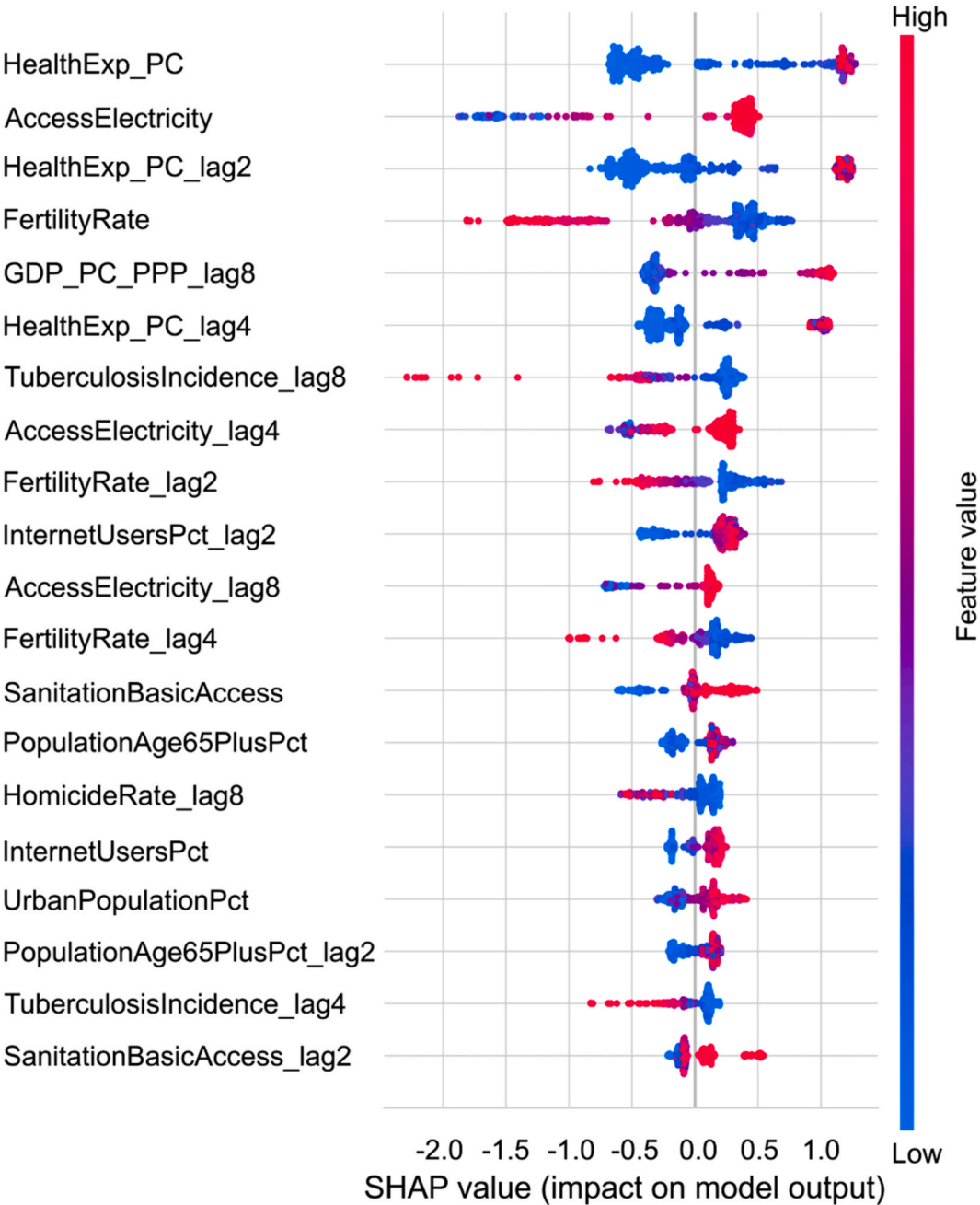


Fig. 3. Global SHAP Summary Plot. The features are ranked by mean absolute SHAP value (importance). **Note.** Each dot represents a country-year observation; red dots indicate high feature values, and blue dots indicate low feature values. The horizontal position shows the impact on predicted life expectancy (positive or negative).

South Africa and Botswana.

In the first part of the sensitivity analysis (Sensitivity Test A), we excluded these two countries. Following their exclusion, the average predictive error metrics remained virtually unchanged (root mean square error (RMSE) decreased marginally from 2.2050 to 2.1949 and mean absolute error (MAE) changed from 1.7138 to 1.7148), while R^2 decreased to 0.4046. This drop in R^2 is a mathematical consequence of restricting the variance of the dependent variable, as these Southern African nations represent outliers in terms of lower-bound longevity in our dataset due to the historical HIV crisis. The stable RMSE and MAE confirm that the baseline explanatory power of the model in this group ($R^2 = 0.5420$) is not solely driven by these Southern African countries. In the second part of the sensitivity analysis (Sensitivity Test B), we excluded the three countries with the largest prediction errors in the test

set: Fiji, Moldova and Indonesia. This resulted in a notable improvement in performance: the R^2 increased to 0.5788, the RMSE decreased to 2.0117 and the MAE decreased to 1.5577 (compared to a baseline MAE of 1.7138).

To systematically evaluate the stability of our model and SHAP feature importance rankings, model training was repeated across 10 distinct random seed perturbations. The overall predictive performance demonstrated high stability, with tight standard deviations across iterations (e.g., R^2 of 0.8818 ± 0.0016 ; MAE of 1.6390 ± 0.0139). However, the exact global SHAP rankings exhibited moderate variability, yielding a mean Spearman rank correlation of $\rho = 0.69$ across the perturbations. Due to the presence of highly collinear lagged variables (e.g., current versus lagged health expenditure), the gradient boosting algorithm frequently swapped the importance of correlated features across

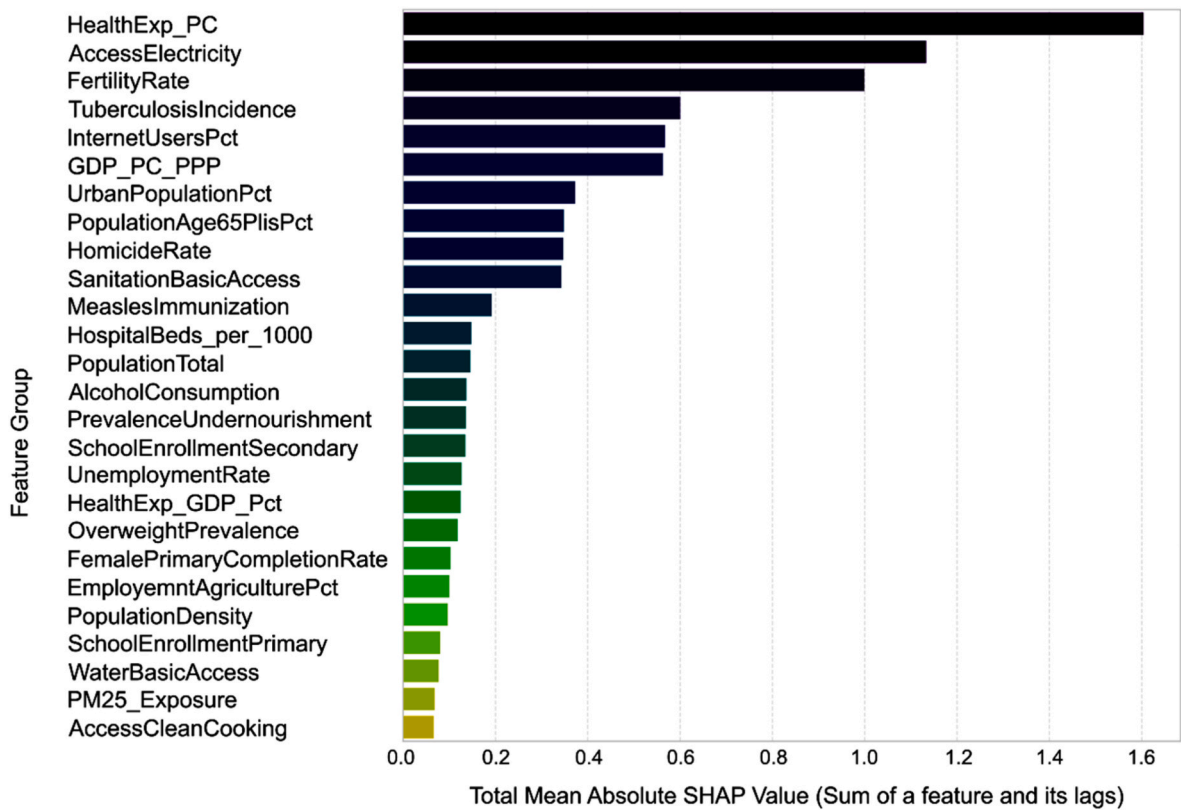


Fig. 4. Ranked feature importance derived from Shapley analysis, showing all 26 base features.

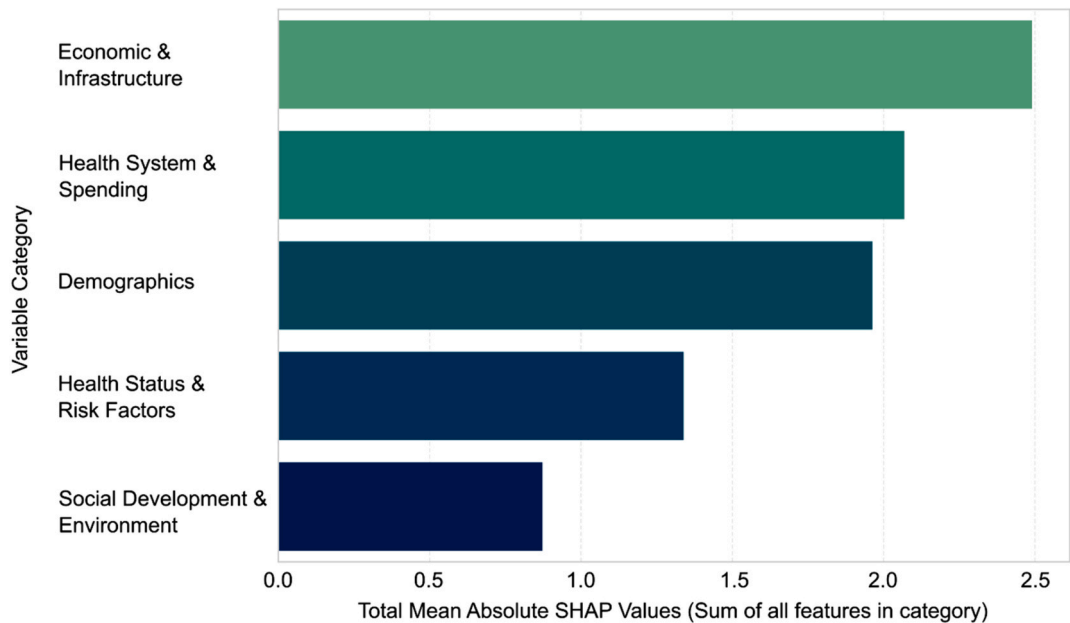


Fig. 5. Aggregated SHAP Feature Importance by variable category.

different random seeds. Table 6 provides a structured summary of the predictive performance across these robustness checks, combining the random seed perturbations with the sensitivity testing of the upper-middle-income cohort.

4.2. Countries case studies

To illustrate how specific factors predict individual outcomes and

move beyond aggregate analysis, we selected three countries as distinct case studies for further granular analysis: Germany, South Africa, and India. We made this deliberate choice to represent diverse points across the World Bank's income classification spectrum and to highlight fundamentally different public health profiles and policy contexts. These criteria ensure that we cover the full range of development statuses. Germany is a representative high-income country with a highly developed public health infrastructure and high life expectancy. South Africa

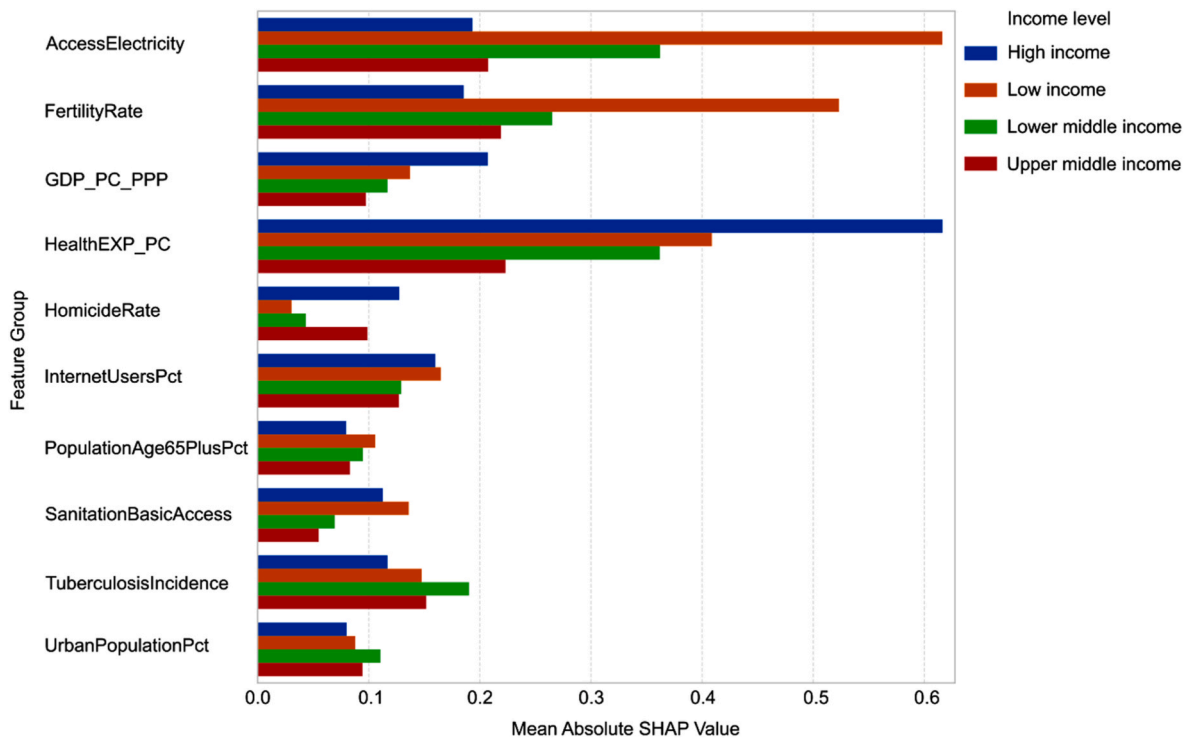


Fig. 6. Comparative SHAP Feature Importance by Income level.

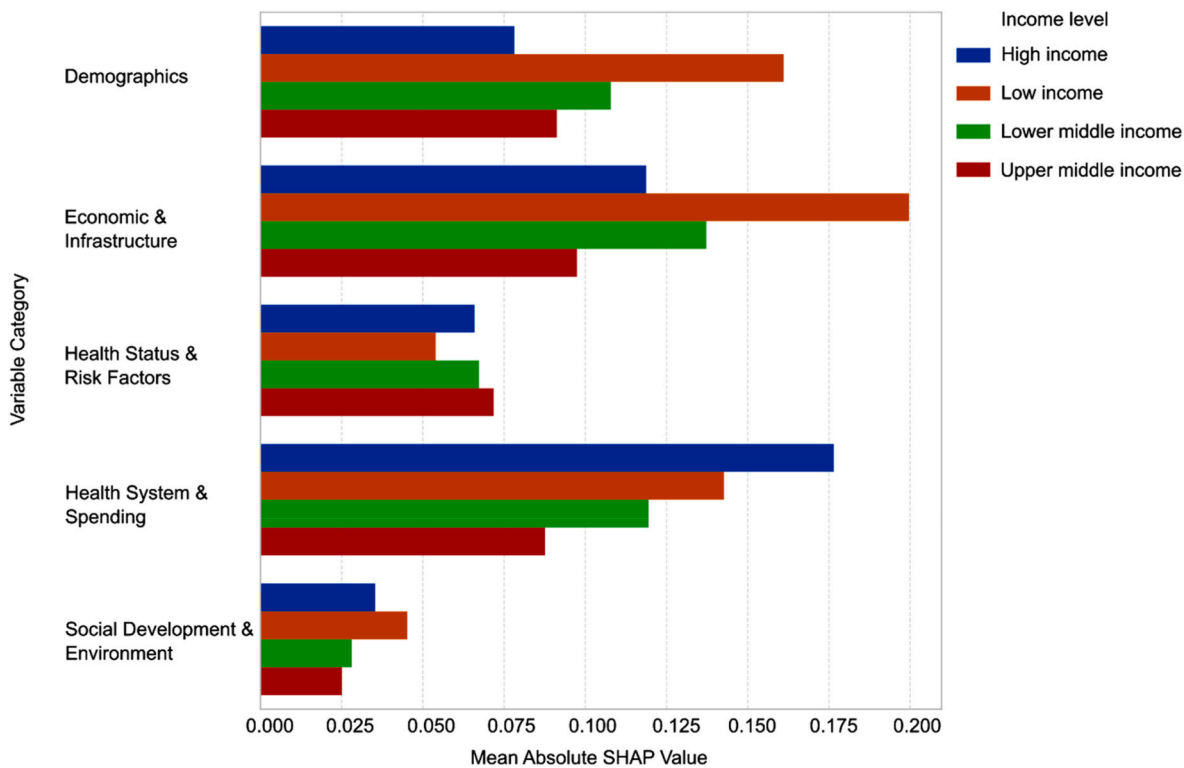


Fig. 7. Comparative (aggregated) SHAP Feature Importance by Income level.

is a crucial and complex case of an upper-middle-income country grappling with a dual health burden. Although it has a significant economic base and infrastructure, its health trajectory has been profoundly shaped by specific public health crises. India is an important example of a large, rapidly developing lower-middle-income country currently undergoing significant demographic and epidemiological transitions. By

analyzing these three diverse contexts, we ensure that our findings cover a broad spectrum of global health determinants. This allows us to validate our model's interpretations across different income levels and policy landscapes.

Figs. 8–10 show the SHAP force plots for Germany, South Africa, and India, respectively. Each prediction starts at a baseline value, which is

Table 4

Top-ranked predictive variables by income level.

Income Level	#1	#2	#3
High income	HealthExp_PC	GDP_PC_PPP	AccessElectricity
Upper-middle income	HealthExp_PC	FertilityRate	AccessElectricity
Lower-middle income	HealthExp_PC	AccessElectricity	FertilityRate
Low income	AccessElectricity	FertilityRate	HealthExp_PC

Table 5

LGBM test set performance across income levels.

Income Level	R-squared (R^2)	RMSE	MAE
High income	0.8185	1.5581	1.1335
Low income	0.7754	1.7837	1.4686
Lower middle income	0.7942	2.0470	1.6063
Upper middle income	0.5420	2.2050	1.7138

Note. RMSE: Root Mean Squared Error; MAE: Mean Absolute Error.

Table 6

Summary of model robustness and sensitivity analyses.

Analysis/Scenario	Dataset Scope	R^2	RMSE	MAE
Baseline	Full Test Set (2020–2023)	0.9007	1.9609	1.4976
LightGBM	Full Test Set (2020–2023)	0.8818	2.1393	1.6390
Seed	Perturbations (10 Runs)	(± 0.0016)	(± 0.0143)	(± 0.0139)
Sensitivity Test A	Excl. South Africa & Botswana	0.4046	2.1949	1.7148
Sensitivity Test B	Excl. Fiji, Moldova, Indonesia	0.5788	2.0117	1.5577

the average predicted life expectancy across the dataset (denoted by $E[f(x)]$). Feature values that increase the individual prediction are shown in

red (positive SHAP values), and feature values that decrease the prediction are shown in blue (negative SHAP values). The magnitude of the impact is represented by the length of the colored segment. Predicted life expectancy is denoted by $f(x)$. In Germany (Fig. 8), where the predicted life expectancy (81.31 years) is well above the global average (72.89 years). The primary drivers are rooted in a high-income context.

The main positive factor is the healthcare system. High HealthExp_PC values are associated with substantial positive SHAP values (a large red segment), pushing the prediction significantly higher than the baseline. For example, high SanitationBasicAccess provides a smaller, albeit positive, contribution. This shows that extensive healthcare investment is the key factor for exceptional longevity in highly developed nations. The pattern in South Africa (Fig. 9) highlights the dominance of specific public health challenges that lower life expectancy.

Tuberculosis Incidence (8 years prior): This is the most dominant negative factor; it alone reduces the prediction by 2.21 years relative to the average. The value from four years prior also has a significant negative impact (-0.67 years), which highlights the disease's persistent and destructive effect on lifespan. The homicide rate (and its eight-year lag) is also a significant negative factor, reducing life expectancy by 0.37 and 0.34 years, respectively. Close-to-average access to electricity shows a varied impact. Internet users (75% of the population) move the prediction in a positive direction. South Africa's predicted life expectancy is 67.26 years, which is approximately 5.6 years less than the average of 72.89 years. It is important to acknowledge that our feature set did not include HIV/AIDS prevalence. For Southern African countries, HIV/AIDS was the primary driver of life expectancy slumps between 1990 and 2023. The model attempts to proxy this massive health crisis through other health metrics (like Tuberculosis incidence), but the omission of a direct HIV/AIDS indicator limits the model's granular accuracy in this specific region. This omission directly accounts for the large residual errors observed in the Southern African cluster in Fig. 2, underscoring a major limitation in interpreting specific country-level case studies in regions where key epidemiological variables are missing.

Fig. 10 shows the SHAP waterfall plot, which serves to explain the individual prediction for India in 2023. It illustrates how the model's average prediction (the default global average) was influenced by the

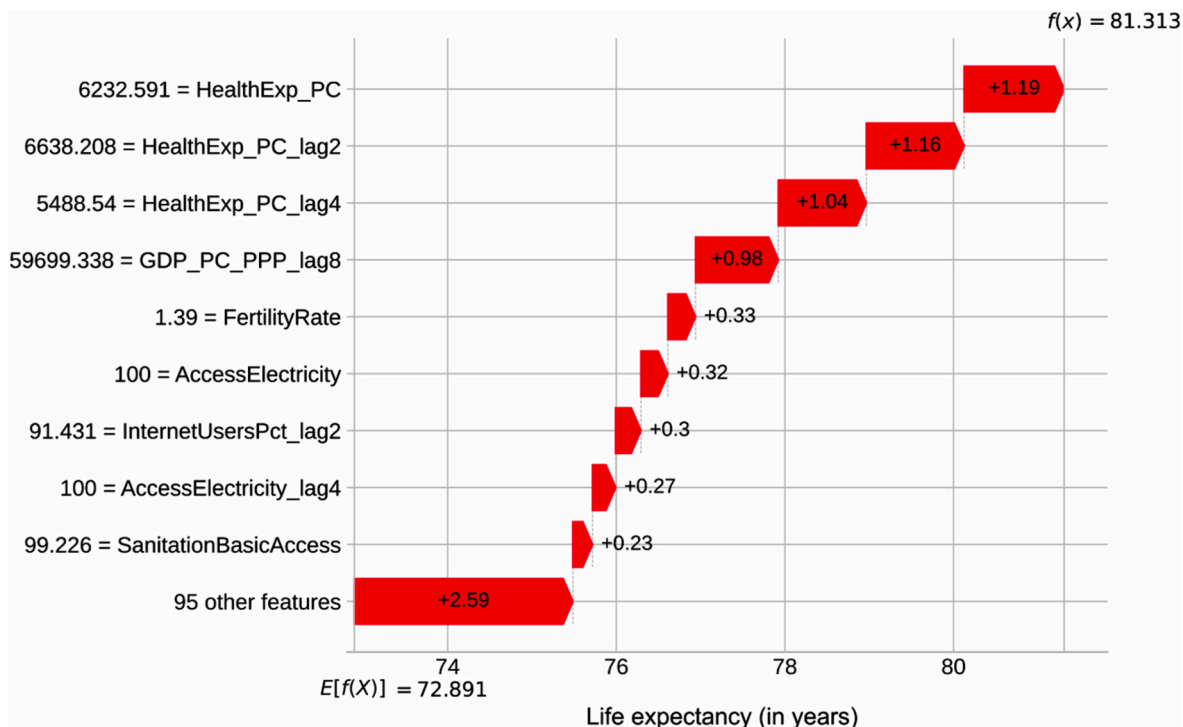


Fig. 8. SHAP waterfall plot for Germany (2023).

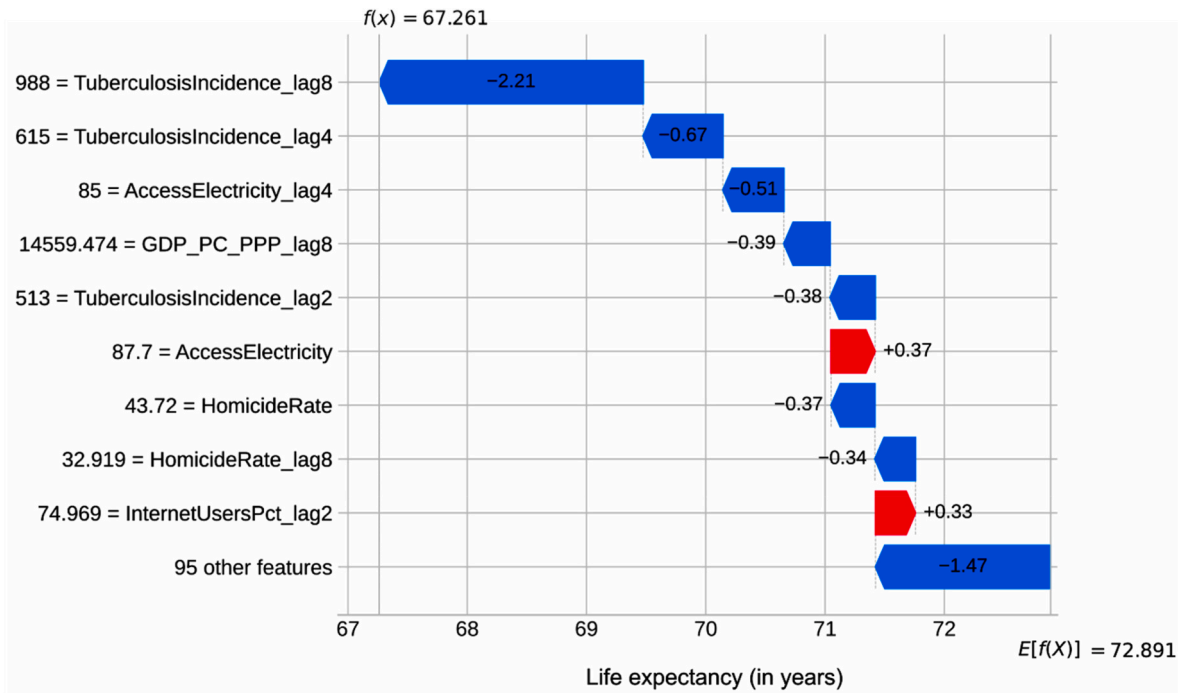


Fig. 9. SHAP waterfall plot for South Africa (2023).

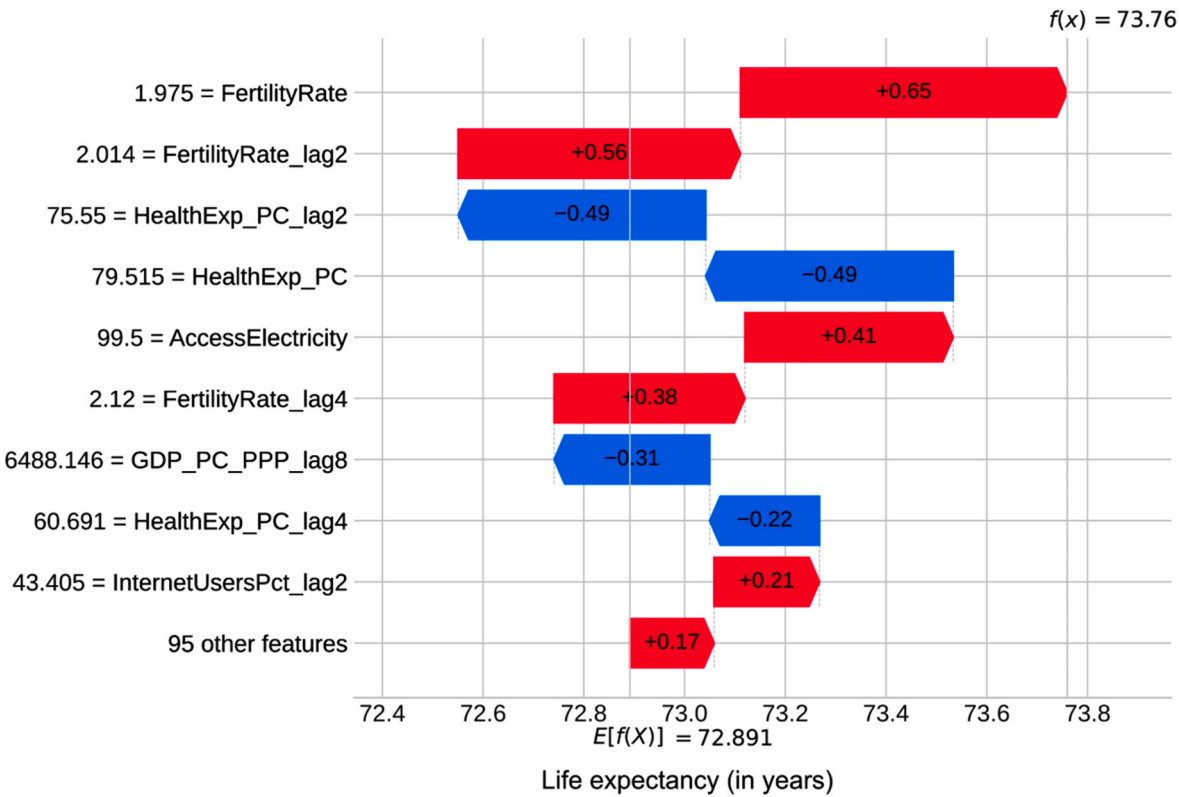


Fig. 10. SHAP waterfall plot for India (2023).

unique characteristics of India, leading to the final prediction for that country. Fig. 10 explains why the model predicts a life expectancy of 73.76 years for India, which is only slightly above the global average (72.89 years). The pattern demonstrates a balance between positive and negative effects.

In the case of India, the observed power decline in fertility rates

functions as a positive predictor; the model associates this downward demographic trend with broader improvements in public health outcomes and the progression of the demographic transition. This factor is the most dominant positive factor, increasing the prediction by 0.65 years. Lagged fertility rate variables are also important, confirming that sustained shifts provide cumulative gains to longevity. Additionally,

access to electricity (99.5% of the population) provides a positive contribution, reflecting the successful expansion of basic infrastructure. However, these positive effects are largely counteracted by health expenditure variables, which have a large negative impact on the prediction because they remain lower than the global average. This suggests that while India has successfully leveraged demographic and infrastructural gains, its path toward higher life expectancy is currently constrained by relatively low healthcare investment. In summary, the positive demographic predictors are mostly offset by these expenditure gaps, keeping the overall prediction above, yet near, the baseline average.

5. Discussion

Our study used a robust gradient boosting framework, the Light Gradient Boosting Machine (LGBM), to predict life expectancy at birth using a comprehensive, pooled dataset. This dataset included 125 countries, spanned an effective 15 years after preprocessing (derived from the 34-year raw data period after lag truncation and the removal of missing data), and utilized 26 predictor variables, each with four different lags (0, 2, 4, and 8 years). Validation results indicate that this model produces more accurate predictions than linear regression. LGBM demonstrated clear statistical superiority over linear regression ($R^2 = 0.901$ vs. 0.716 ; $MAPE = 2.08\%$ vs. 3.38%). These results are consistent with the finding that machine learning models, such as gradient boosting, can effectively handle complex data structures in life expectancy predictions (Ahmad et al., 2025). SHAP values reveal the model's "black box," by measuring the precise contribution of each variable to the predictions, thereby enhancing transparency (Khan & Rafat, 2025; Zhao & Song, 2025). This methodology mirrors recent advancements in geriatric health modeling, where SHAP values have been used to identify depressive symptoms and physical frailty as dominant predictors of dysfunction in older adults (Lin et al., 2025). Our results show that economic and infrastructure variables are the most predictive categories, which aligns with theoretical expectation H1. Focusing on the interpretation of these heterogeneous patterns, the observed shift in predictive importance across income groups (Fig. 7) empirically validates the nonlinear relationship described by the Preston curve (Preston, 1975). Our findings suggest that, in a low-income context, life expectancy is primarily predicted by fundamental infrastructure and demographic factors, such as access to electricity and fertility rates. However, as nations transition to higher income levels, the predictive power of these basic determinants diminishes. In high-income regions, the model's predictive focus decisively shifts toward health system efficiency and specialized expenditures. This transition suggests that once a society surpasses a certain developmental threshold, the longevity of the healthcare system becomes crucial for managing non-communicable diseases and geriatric care. This aligns with the flattening of the Preston curve, where income alone becomes a less potent predictor than the specific allocation of resources within the healthcare sector.

Further analysis of lagged socioeconomic variables revealed their significant long-term impact on life expectancy predictions. This finding aligns with the work of Ramos-Herrera and Sosvilla-Rivero (2025), who examined the enduring effects of socioeconomic policies on health outcomes. Therefore, these results align with expectation H2.

The study shows that a higher fertility rate and a higher incidence of tuberculosis are negative correlated with life expectancy. Our SHAP analysis corroborates this finding, suggesting that substantial disease burdens and demographic challenges can negatively impact longevity. This aligns with Woolf (2023), who documented socioeconomic disparities affecting life expectancy across different regions. Thus, our findings are consistent with theoretical expectation H3.

We found considerable variation in SHAP contributions among different groups of countries, which is consistent with expectation H4; our case studies of Germany, South Africa, and India demonstrate this. In

high-income regions such as Germany, where life expectancy gains are often driven by continuous improvements in geriatric care and the management of non-communicable diseases, healthcare expenditure is paramount. However, its importance is reduced in other countries. Selecting South Africa allowed us to explore a scenario in which average life expectancy is low despite middle-income status due to the high burden of specific infectious diseases. This demonstrates how dominant context-specific factors can offset general economic gains. India served as a case study for transitional determinants, illustrating the balance of positive and negative forces, showing that basic infrastructure improvements can have significant impact. These observations align with those of Rapi et al. (2025), who explored the nuanced relationship between income levels and insurance demand. Our findings suggest that health expenditure is unlikely to lead to meaningful improvements in life expectancy without meeting basic needs, reflecting a hierarchy of developmental priorities.

Model performances varies substantially by income group (Table 5). The relatively lower R^2 observed in upper-middle-income countries (0.5420) compared to the overall model (0.901) suggests that life expectancy in transitioning economies is driven by highly heterogeneous, context-specific factors. This group comprises a structurally diverse mix of nations, including Latin American countries, Eastern European transition states, and Southern African nations. Although Eswatini is classified in the lower-middle-income bracket, other Southern African countries like South Africa and Botswana fall into the upper-middle-group. In these countries, historical public health crises, particularly the HIV/AIDS epidemic, have dramatically impacted life expectancy. The direct inclusion of HIV/AIDS prevalence variable was prevented by data limitations (missing values). Our sensitivity analysis revealed that while removing these Southern African nations left the average predictive error (RMSE and MAE) unchanged, excluding the three countries (Fiji, Moldova, and Indonesia) with the largest prediction errors improved the R^2 and reduced the RMSE. This confirms that the lower baseline performance of the model in the upper-middle-income category is not caused by a single regional outlier or disease omission. This sensitivity analysis shows that lower baseline performance in the upper-middle-income group is not concentrated in a single geographic region or caused solely by the omission of a single disease. Rather, it reflects the systemic, multi-regional structural heterogeneity of transition economies, ranging from Eastern European post-Soviet states such as Moldova to small island developing states such as Fiji and rapidly industrializing Asian nations such as Indonesia. These economies experience highly diverse regional, demographic and epidemiological trajectories that are difficult to fully capture with a single, static model.

Overall, our study confirms the vital role of socioeconomic investments and historical factors in predicting life expectancy, highlighting the necessity of targeted policy interventions. SHAP values offer actionable insights into complex predictions, providing a robust foundation for informed decision-making and resource allocation aimed at improving health outcomes in diverse socioeconomic contexts. These results suggest that policy planning may benefit from targeting country-specific predictive drivers, highlighting a strong predictive association between basic infrastructure improvements and predicted life expectancy in lower-income countries, and advanced healthcare investments in wealthier ones.

6. Conclusion

In terms of methodological contributions, the application of advanced machine learning techniques, particularly Light Gradient Boosting Machine (LGBM), has proven to be a pivotal in accurately predicting life expectancy. The model's capacity to process intricate datasets and offer insightful insights via SHAP values represents a substantial improvement over conventional methods such as linear regression. This study has extensively explored the interplay of socioeconomic and health-related variables, revealing their substantial impact on life

expectancy across different income groups and geographic contexts.

Regarding empirical findings, the results provide strong empirical alignment with our formulated theoretical expectations, underscoring the methodological strength of combining nonlinear machine learning with interpretable SHAP analysis on a global longitudinal dataset. H1, which posits that health expenditure per capita is the most influential predictor, is reflected in the dominant SHAP values observed globally, highlighting a strong predictive association between healthcare investment and predicted longevity. Similarly, the findings are consistent with H2, confirming the significant long-term predictive value of lagged socioeconomic variables, such as access to electricity and GDP. This supports the need for policies that consider past conditions and their enduring associations with population health. Our results also align with H3, indicating that negative correlations between fertility rates, disease burdens, and life expectancy must be addressed through targeted interventions in regions with high disease prevalence. Lastly, the evidence supports H4, confirming that moderating the predictive associations of variables by income levels requires tailoring socioeconomic strategies to a country's specific economic context. High-income regions benefit more from healthcare investments, while lower-income countries may need to focus on demographic and infrastructure improvements.

Based on these predictive insights, several model-informed policy considerations emerge. However, these must be interpreted within strict predictive boundaries rather than as causal guarantees. First, Empirical Findings: Higher healthcare expenditure exhibits a strong predictive association with life expectancy in high-income regions, while basic infrastructure access (e.g., electricity) is the leading predictive factor in lower-income settings. Methodological Contributions: By combining tree-based gradient boosting with multi-tier SHAP decomposition, this study overcomes the black-box constraint of ML models while managing complex non-linear lag dynamics. Model-Informed Policy Implications & Boundaries: It is crucial to emphasize that SHAP values reflect predictive associations within the model architecture rather than direct real-world causal mechanisms. Policy recommendations derived from these models serve as complementary, data-driven signals rather than deterministic evidence that specific expenditures will guarantee exact health gains. Future Directions: Future research should extend this framework by incorporating time-varying coefficient models and dynamic panel structures to better isolate localized causal mechanisms.

Variables associated with reduced fertility rates and controlled disease burdens also align with higher predicted life expectancy, especially in areas where these factors represent major constraints. For future research directions, our framework highlights the importance of analyzing predictive patterns sensitive to a country's historical and current economic conditions to better understand its unique structural challenges.

Despite the model's high predictive accuracy, several limitations warrant consideration. First, our reliance on secondary data (WDI) means findings are associative and predictive rather than causal. Inherent data quality issues may exist, and the use of static income classifications may misclassify countries whose income levels changed over the study period. Next, while forward-filling is a standard approach for maintaining the integrity of longitudinal panel data, it may partially obscure short-term volatility during sudden economic shocks or rapid health crises. Our selection of lag structures (0, 2, 4, and 8 years) was designed to capture broad developmental trends, though we acknowledge that certain intermediate temporal dynamics might not be fully represented. Furthermore, it is critical to emphasize that purely predictive modeling lacks causal inference. SHAP values represent marginal contributions to the model's specific predictive output and should not be interpreted as indicators of direct real-world causality or guaranteed policy effectiveness. The omission of important epidemiological variables (such as HIV/AIDS prevalence, which could not be directly included due to panel data gaps) and cultural/institutional factors limits the model's granular accuracy in regions where these are primary mortality factors, as demonstrated by our sensitivity analysis on the

upper-middle-income cohort. Additionally, the unbalanced nature of our panel dataset implies that countries with more comprehensive historical records may exert a disproportionate influence on the global model's learned patterns. Finally, our pooled panel data approach assumes a degree of general parameter homogeneity and may partially obscure unobserved country-specific heterogeneity; therefore, future research could further refine these insights by utilizing time-varying coefficients or regime-switching models. Finally, while we formally verified the statistical significance of the overall performance gap between our models, our evaluation framework primarily focuses on point predictions and feature interpretability. As demonstrated by our perturbation analysis, the presence of highly correlated temporal lags introduces moderate variability in exact SHAP feature rankings across random seeds. While we formally estimated 95% confidence intervals for overall model performance via bootstrapping, we did not generate uncertainty estimations for individual point predictions and their corresponding local SHAP values; therefore, measuring the local confidence intervals of specific feature importance remains a valuable avenue for future methodological research.

In conclusion, this study reinforces the importance of leveraging advanced machine learning techniques to understand the multifaceted determinants of life expectancy. Aligning these insights with targeted policy interventions can drive significant improvements in global health outcomes and reduce disparities between different socioeconomic groups.

Generative AI usage

Generative artificial intelligence (AI) tools were used to support parts of this research. Google's Gemini Pro 3.0 large language model was used to aid in code generation and refinement of manuscript text. WOS Core Collection's AI supported research assistant was used for literature search. The authors used DeepL Write ([deepl.com](https://www.deepl.com)) in order to improve readability and language. All outputs produced by these tools were critically reviewed, verified, and edited by the authors to ensure accuracy, originality, and adherence to academic standards. The final responsibility for the content of this article rests entirely with the authors.

CRediT authorship contribution statement

Levente Sándor Nádasí: Writing – review & editing, Writing – original draft, Conceptualization. **Gábor Petneházi:** Formal analysis, Data curation, Conceptualization. **Sándor Kovács:** Writing – review & editing, Software, Methodology, Funding acquisition, Formal analysis. **Bernadett Aradi:** Supervision, Data curation.

Data availability

The underlying dataset sourced from the World Bank World Development Indicators (WDI) and the Python reproduction scripts are available from the corresponding author upon reasonable request and via the GitHub repository.

Ethics statement

Ethical approval is not applicable to this manuscript.

Research funding

This research did not receive funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors of the study would like to thank to the University of Debrecen Program for Scientific Publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ssaho.2026.103704>.

References

- Ahamad, S. S., Kumar, K. P., & Koneru, S. V. (2025). Machine learning-based life expectancy prediction in developed and developing regions. *IEEE Access*, 13, 69520–69531. <https://doi.org/10.1109/access.2025.3560890>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631). <https://doi.org/10.1145/3292500.3330701>
- Ali, M. N. Y., Faruq, O., Ahammed, F., Mahzabin, K., Dihan, S. S., & Jahan, I. (2024). Predicting post-thoracic surgery life expectancy for lung cancer using machine learning with SHapley additive exPlanations. In *2024 27th international conference on computer and information technology (ICCIT)* (pp. 334–339). IEEE. <https://doi.org/10.1109/iccit64611.2024.11022499>
- Alvi, M. B., Alvi, M., & Fatima, N. (2025). Life expectancy prediction using recursive partitioning and bagging algorithms. *EAI Endorsed Transactions on Scalable Information Systems*, 12(5), 1–18. <https://doi.org/10.4108/eetis.8959>
- Bayati, M., Akbarian, R., & Kavosi, Z. (2013). Determinants of life expectancy in eastern mediterranean region: A health production function. *International Journal of Health Policy and Management*, 1(1), 57. <https://doi.org/10.15171/ijhpm.2013.09>
- Bimonte, G., Russolillo, M., Shang, H. L., & Yang, Y. (2024). Mortality models ensemble via shapley value. *Decisions in Economics and Finance*, 1–29. <https://doi.org/10.1007/s10203-024-00455-z>
- Cicek, Z. I. E., & Aydin, Z. E. (2025). Air quality forecasting in presence of missing data. In *Intelligent systems and pattern recognition, ISPR 2024, part II* (pp. 250–261). https://doi.org/10.1007/978-3-031-82153-0_19
- De Arcos-Jimenez, J. C., Quintero-Salgado, E., & Briseno-Ramirez, J. (2025). Population-level SARS-CoV-2 RT-PCR cycle threshold values and their relationships with COVID-19 transmission and outcome metrics: A time series analysis across pandemic years. *Viruses*, 17(1). <https://doi.org/10.3390/v17010103>
- Dutta, N., & Dhar, D. (2024). Understanding medical technology innovation in low- and middle-income countries: Factors, impact, and a model proposal. *She Ji: The Journal of Design, Economics, and Innovation*, 10(2), 192–222. <https://doi.org/10.1016/j.sheji.2024.07.002>
- Fan, C. X., & Bo, L. (2025). Determinants of life expectancy in China: A comparative analysis of economic and non-economic factors. *African Journal of Reproductive Health*, 29(3), 160–167. <https://doi.org/10.29063/ajrh2025/v29i3.18>
- Gazilas, E. T. (2024). Factors influencing life expectancy in low-income countries: A panel data analysis. *Journal of Applied Economic Research*, 23(3), 580–601. <https://doi.org/10.15826/vestnik.2024.23.3.023>
- George, B., Waititu, G. A., & Awichi, R. O. (2025). Comparison of bayesian dynamic linear and mixed effects interrupted time series models in assessing Uganda's neonatal mortality reduction since the SDGs onset. *Communications in Mathematical Biology and Neuroscience*, 2025, 1–19. <https://doi.org/10.28919/cmbn/9440>, 127.
- Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy*, 80(2), 223–255. <https://doi.org/10.1086/259880>
- Hansen, C. W., & Lonstrup, L. (2015). The rise in life expectancy and economic growth in the 20th century. *Economic Journal*, Article 2340926. <https://doi.org/10.1111/econj.12261>
- Irandoost, K., Daroudi, R., & Yaseri, M. (2025). Global and regional impact of health determinants on life expectancy and health-adjusted life expectancy, 2000–2018: An econometric analysis based on the global burden of disease study 2019. *Frontiers in Public Health*, 13. <https://doi.org/10.3389/fpubh.2025.1566469>
- Jaba, E., Balan, C. B., & Robu, I. B. (2014). The relationship between life expectancy at birth and health expenditures estimated by a cross-country and time-series analysis. *Procedia Economics and Finance*, 15, 108–114. [https://doi.org/10.1016/s2212-5671\(14\)00454-7](https://doi.org/10.1016/s2212-5671(14)00454-7)
- Kaplan, R. M., Spittel, M. L., & Zeno, T. L. (2014). Educational attainment and life expectancy. *Policy Insights from the Behavioral and Brain Sciences*, 1(1), 189–194. <https://doi.org/10.1177/2372732214549754>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Khan, N. A., & Rafat, A. M. S. (2025). Enhancing fetal health assessment using machine learning and XAI techniques. *Baghdad Science Journal*, 22(7), 2469–2478. <https://doi.org/10.21123/2411-7986.5010>
- Kousar, S., Ahmed, F., & Segovia, J. E. T. (2023). Is government spending in the education and health sector necessary for human capital development? *Humanities and Social Sciences Communications*, 10(1). <https://doi.org/10.1057/s41599-023-01514-3>
- Lin, F. B., Liu, C., & Liu, H. (2025). Development and validation of a predictive model for activities of daily living dysfunction in older adults. *JMIR Medical Informatics*, 13, Article e73030. <https://doi.org/10.2196/73030>
- Lipesa, B. A., Okongo, E., Omolo, B. O., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. *SN Applied Sciences*, 5(7), 189. <https://doi.org/10.1007/s42452-023-05404-w>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU accession candidate countries. *Genus*, 76(1), 2. <https://doi.org/10.1186/s41118-019-0071-0>
- Mourad, M. (2021). Impact of socio-economic variables on life expectancy: An empirical study for 138 countries. *Journal of Public Administration and Governance*, 11(1). <https://doi.org/10.5296/jpag.v11i1.18437>
- Neofytidou, A., & Fountas, S. (2020). The impact of health on GDP: A panel data investigation. *The Journal of Economic Asymmetries*, 21, Article e00139. <https://doi.org/10.1016/j.jeca.2019.e00139>
- Nica, E., Poliakova, A., Popescu, G. H., Valaskova, K., Burcea, S. G., & Constantin, A. L. D. (2023). The impact of financial development, health expenditure, CO2 emissions, institutional quality, and energy mix on life expectancy in eastern Europe: CS-ARDL and quantile regression approaches. *Heliyon*, 9(11). <https://doi.org/10.1016/j.heliyon.2023.e21084>
- Okunade, A. A., & Osmani, A. R. (2020). Effects of life expectancy on economic growth: New results using the flexible box-cox power transformation model. *Applied Economics Letters*, 27(20), 1681–1684. <https://doi.org/10.1080/13504851.2020.1713976>
- Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies*, 29(2), 231–248. <https://doi.org/10.1080/00324728.1975.10410201>
- Ramos-Herrera, M. D. C., & Sosvilla-Rivero, S. (2025). Pension expenditures and old-age poverty in OECD countries. *National Accounting Review*, 7(1), 28–58. <https://doi.org/10.3934/nar.2025002>
- Rapi, K., Priyansono, D. S., & Bakhtiar, T. (2025). The non-monotonic relationship between income and life insurance demand. *Economics*, 13(1), 4. <https://doi.org/10.3390/economics13010004>
- Ray, D., & Linden, M. (2020). Health expenditure, longevity, and child mortality: Dynamic panel data approach with global data. *International Journal of Health Economics and Management*, 20, 99–119. <https://doi.org/10.1007/s10754-019-09272-z>
- Roffia, P., Bucciol, A., & Hashlamoun, S. (2023). Determinants of life expectancy at birth: A longitudinal study on OECD countries. *International Journal of Health Economics and Management*, 23(2), 189–212. <https://doi.org/10.1007/s10754-022-09338-5>
- Samarawickrama, P. A., Fernando, G., & Pathirana, U. (2025). Macroeconomic determinants of child mortality in low and lower-middle-income nations. *International Journal of Sociology & Social Policy*, 45(5/6), 586–605. <https://doi.org/10.1108/ijsspp-02-2025-0128>
- Schwandt, H., Currie, J., Von Wachter, T., Kowarski, J., Chapman, D., & Woolf, S. H. (2022). Changes in the relationship between income and life expectancy before and during the COVID-19 pandemic, California, 2015–2021. *JAMA*, 328(4), 360–366. <https://doi.org/10.1001/jama.2022.10952>
- Shapley, L. (1953). A value for n-person games. In H. Kuhn, & A. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton: Princeton University Press.
- Skinner, J., & Staiger, D. (2015). Technology diffusion and productivity growth in health care. *The Review of Economics and Statistics*, 97(5), 951–964. https://doi.org/10.1162/rest_a.00535
- Țarcă, V., Țarcă, E., & Moscalu, M. (2024). Social and economic determinants of life expectancy at birth in eastern Europe. *Healthcare*, 12(11), 1148. <https://doi.org/10.3390/healthcare12111148>
- Tu, Y., Chen, B., & Xu, B. (2025). Inequality in infrastructure access and its association with health disparities. *Nature Human Behaviour*, 9(8). <https://doi.org/10.1038/s41562-025-02208-3>
- Walczak, D., Wantoch-Rekowski, J., & Marczak, R. (2021). Impact of income on life expectancy: A challenge for the pension policy. *Risks*, 9(4), 65. <https://doi.org/10.3390/risks9040065>
- Wang, Y. (2021). The greatest factors affecting life expectancy: A research based on different continents and countries. In *2021 3rd international conference on machine learning, big data and business intelligence (MLBDBI)* (pp. 531–541). IEEE. <https://doi.org/10.1109/mlbdbl54094.2021.00107>
- Woolf, S. H. (2023). Falling behind: The growing gap in life expectancy between the United States and other countries. *American Journal of Public Health*, 113(9), 970–980. <https://doi.org/10.2105/ajph.2023.307310>
- World Bank. (2025a). World development indicators. Retrieved from <https://databank.worldbank.org/source/world-development-indicators>, 2025 November.
- World Bank. (2025b). Country classifications by income level for 2024–2025. Retrieved from <https://blogs.worldbank.org/en/opendata/world-bank-country-classifications-by-income-level-for-2024-2025>, 2025 November.
- Zhao, H. J., & Song, G. S. (2025). AVP-GPT2: A Transformer-powered platform for de novo generation, screening, and explanation of antiviral peptides. *Viruses*, 17(1), 14. <https://doi.org/10.3390/v17010014>