Deep Learning Insights into Social Determinants of Chronic Disease and Longevity

Ishrat Jaben Bushra • Kin Kwan Liu • Muhammad Zaka Shaheryar • Sazia Afreen

STEM FELLOWSHIP

Toronto Metropolitan University
Data Science & Analytics Graduate Studies

Research Question

1. Which SDoH most significantly influence chronic disease mortality, specifically diabetes, cardiovascular diseases, alzheimer and other dementia, asthma, and kidney diseases, in aging populations?
2. How do these influences differ across income groups and geographic regions?
3. Can deep learning models accurately predict chronic disease outcomes fromm macroscale indicators, and how do interpretable AI techniques (e.g., SHAP) enhance understanding of these relationships?

Background

Social determinants of health (SDoH), for example, access to healthcare, education, and income, have a significant impact on the prevalence of chronic diseases in older populations. Reducing avoidable death and promoting health equity require understanding and addressing these inequities.

Hypothesis

- The mortality of chronic disease vary substantially depending on a country's income level, for example, low-income countries face challenges from healthcare systems and undernourishment; while high-income countries experience challenges from aging and metabolic factors
- Explainable AI techniques (like SHAP) can identify important elements to direct better aging practices, whereas machine learning can model intricate correlations between SDoH and disease

Methods

- **Data Sources:**
  - World Bank data for SDoH indicators [1]
  - WHO Global Health Estimates for the causes of death [2]
  - World Income Inequality Database for income data [3]
- **Data preprocessing:**
  - Merge by country and year
  - Impute missing values using forward/backward fill and inequality group mean based on Palma Ratio [4]
- **Exploratory Analysis:**
  - Trends (2000-2021), correlations, and mortality distribution by disease and region
- **Modeling:**
  - Hyperparameter tuning
  - Linear Regression (baseline)
  - XGBoost
  - LSTM & GRU (for sequential patterns)
- **Interpretability:**
  - SHAP values identified key SDoH driving mortality outcomes

Flowchart for Methods

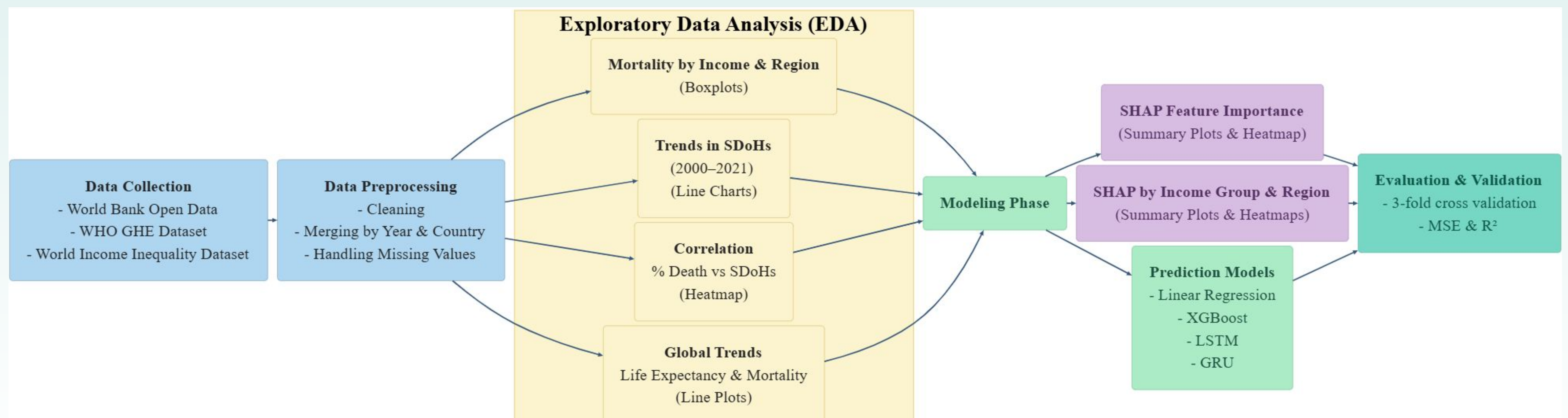

Figure 1: Flowchart of the analysis

Key results

Disease	Test Set MSE				Test Set R^2			
	Regression	XGBoost	LSTM	GRU	Regression	XGBoost	LSTM	GRU
Alzheimer	2.6036	0.3542	0.8937	0.9182	0.5943	0.9447	0.8605	0.8566
Asthma	0.1843	0.0201	0.0400	0.0426	0.3237	0.9264	0.8534	0.8436
Cardiovascular	40.3658	12.6651	25.4781	23.0733	0.7735	0.9290	0.8571	0.8706
Diabetes	4.3467	0.3428	3.9258	1.4542	0.4083	0.9533	0.5327	0.8020
Kidney	1.6912	0.2689	0.7169	0.6528	0.4639	0.9148	0.7727	0.7931

Table 1: Model Performance for Disease Prediction

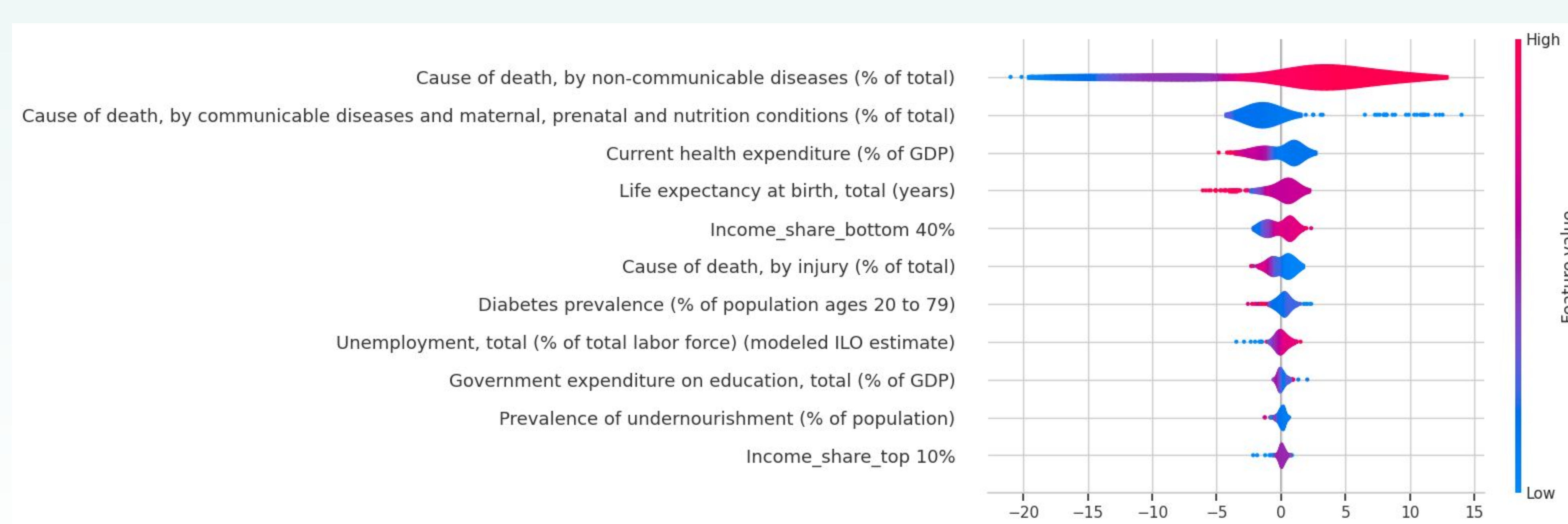

Figure 2: SHAP for SoDH Importance Cardiovascular Diseases

Interpretation

- Diabetes and cardiovascular disease mortality were primarily influenced by income distribution, highlighting the intersection of metabolic risk and socioeconomic inequality
- For kidney disease, diabetes prevalence and injury-related mortality were key contributors, pointing to the role of both metabolic comorbidities and physical health risks
- Alzheimer's mortality was uniquely associated with government education expenditure, possibly reflecting the protective role of lifelong cognitive enrichment
- Asthma mortality was most strongly associated with undernourishment, suggesting vulnerability to nutritional and environmental stressors

Key Findings

- **Top Predictors:** Life expectancy, diabetes prevalence, inequality, and unemployment
- SDoH impacts vary by income level and region, calling for tailored health strategies
  - In **high-income** countries, chronic disease mortality primarily relates to aging and metabolic factors
  - **Low-income** countries experience greater influence from structural vulnerabilities, including communicable diseases, unemployment, and systemic infrastructure gaps.
  - Diabetes prevalence drives mortality in **East Asia & Pacific** and **Latin America**, whereas **South Asia** faces chronic disease burden and undernutrition.
  - Cardiovascular disease mortality reflects aging impacts in Europe and structural factors like unemployment in the **Middle East & North Africa**
- 95% explanation of variance is shown for by Diabetes using **XGBoost** (highest R^2)
- XGBoost achieved the highest accuracy across all diseases with lowest MSE among all models
- XGBoost achieved the highest explanation of variance across all diseases ($R^2 > 0.91$)

Challenges

- Missing data in limited resources countries.
- Absence of genetic and individual-level data.
- Possible unmeasured confounding factors (e.g., healthcare quality)
- Limited temporal resolution data (Only for selected years)

Future Directions

- Incorporate individual-level and geospatial data to improve predictions
- Use attention-based deep learning models (e.g., transformers) for policy impact and interpretability.
- Develop time-aware and longitudinal models for monitoring the disease progression

References

[1] World Bank Group. World bank open data [Internet], 2025. [cited 2025 May 14]
[2] World Health Organization. Global health estimates: Leading causes of death [Internet], 2025. [cited 2025 May 14]
[3] United Nations University World Institute for Development Economics Research. Unuwider: World income inequality database [Internet], 2025. [cited 2025 June 11]
[4] José Gabriel Palma. Has the income share of the middle and upper-middle been stable around the '50/50 rule', or has it converged towards that level? the 'palma ratio' revisited. *Development and Change*, 45(6):1416–1448, 2014.