Deep Learning for Social Sciences - Group Project: Predicting GINI in Country Networks

Jonas Schrade	Jonah Hartmann	Leander Dietrich	Samuel Rauh
1080887	1477524	952492	1534885

August 31, 2025

Introduction

Income inequality remains one of the most pressing and persistent challenges in global social science, shaping debates on development, justice, and policy. The GINI index measures income inequality, ranging from 0 (perfect equality) to 1 (perfect inequality). Despite its global relevance, GINI data is sparse and inconsistently reported. Focusing on UN member states from 2000 to 2022, we address missingness in the dataset by leveraging multi-modal graph learning. Using global networks across five dimensions—geographic, political, economic, cultural, and linguistic—we apply Graph Attention Networks (GATs) fused by a multi-layer perceptron (MLP) head to estimate missing GINI values and assess inequality patterns. Compared to a flat-featured XGBoost baseline, our approach integrates structured relationships to capture complex drivers of inequality. This study highlights the potential and limitations of graph-based models for inequality prediction and underscores the value of multi-dimensional structural information in development analysis.

Data

Target: GINI

The GINI index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. Conclusively, an index score of 0 represents perfect equality, whereas a score of 100 implies perfect inequality. Data is provided by the World Bank (2024c), covering a time range between 1960 and 2024 with inconsistent annual reporting on 216 unique countries. For the scope of this analysis, the data retrieved is limited to observations of UN member countries between 2000 and 2022 and scaled between 0 and 1. Figure 1 reflects the global distribution of GINI scoring with relative equality across broad parts of Europe, Asia, Northern America and Africa but also presents striking clusters of inequality in the American and African South. Despite our already limited scope, as Figure 2 demonstrates, temporal and spatial missingness is a common pattern in GINI data. Only 166 of the 193 UN countries have at least one GINI index within the time span. This deficiency forms the basis for the objective of this task. Masking missing GINI indices, we utilize global similarity networks in order to propagate information across countries and years, leveraging political, geographic, cultural, and economic ties to generate predictive signals for estimating unavailable values and forecasting inequality dynamics.

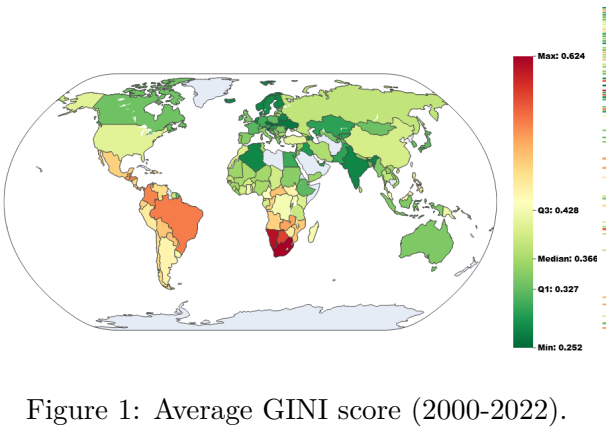


Figure 1: Average GINI score (2000-2022).

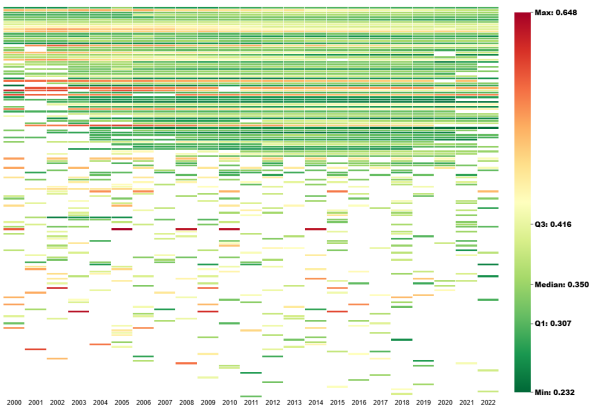


Figure 2: GINI availability per country-year.

Node and Edge Features: Dimensions of Similarity

To capture the diverse channels through which cross-country similarities may shape inequality patterns—and ultimately GINI outcomes—we turn to multiple network representations. Different network types are defined both by their domain focus and the availability of their respective data sources. Combining multiple sources without accounting for inconsistencies in country coverage would introduce substantial missingness, as countries present in one dataset may be absent in another.

Geography:

The nodes and node features for the geographical data is collected via the World Bank API and an open-source dataset. The API provides the country code and the squared kilometers of each country. The second data source furnishes more information on the location of the country, e.g. on which continent the country is and to which subregion it belongs. The categorical data is one-hot encoded. The dataset is derived from the GeoDist project by CEPII Mayer and Zignago, 2011. It provides dyadic variables for pairs of countries, where each row represents a directed country-pair (origin - destination) in a given year. The dataset captures geographic, cultural, and historical linkages between states. The edges are coded using ISO3 representation of countries. The features included are especially distance between capitals or largest city of the country. Beyond geographic distance, the dataset also records structural linkages between states. Contiguity is a binary variable that indicates whether two countries share a land border, while *comlang* identifies whether they share an official or widely spoken language. Historical ties are reflected in the variable colony, which denotes whether a colonial relationship existed between the two states. An advantage is that there are no missing values in both data sets.

Politics:

Node data for the political graph are drawn from three sources: First, the Annual Democracy Index (2000–2018) of Polity5 Project (2018) measures regime authority for 167 countries across executive recruitment, constraints on executive authority, and political competition, yielding a country–year score from –10 (hereditary monarchy) to +10 (consolidated democracy). Second, the Worldwide Governance Indicators (WGI) of World Bank (2024e) provide governance scores for 200 countries and territories (1996–2023) along six dimensions: Voice and Accountability, Regulatory Quality, Political Stability and Absence of Violence/Terrorism, Rule of Law, Government Effectiveness, and Control of Corruption, ranging from –2.5 to 2.5. Third, five high-level indices from Coppedge et al. (2025) summarize polyarchy (electoral democracy), liberal, participatory, deliberative, and egalitarian democracy for 177 countries annually (1789–2024) on a 0–1 scale. To reduce missingness, values within each country–year series are linearly interpolated across

time, and remaining cases with missing values are excluded (see Figure 3), resulting in a node dataset of 3,785 observations for 165 countries between 2000 and 2022. Edges and edge features are based on the Formal Alliances dataset (v4.1) of the Correlates of War project (Gibler, 2013), which records undirected dyads of countries in formal alliances from 1816 to 2012. For this study, only alliances beginning in or after 1945 are considered, and those without an end date are treated as ongoing. Between 2000 and 2022 this results in 1,614 active alliances, spanning non-exclusive categories including defense pacts, neutrality agreements, nonaggression treaties, ententes, and asymmetric arrangements.

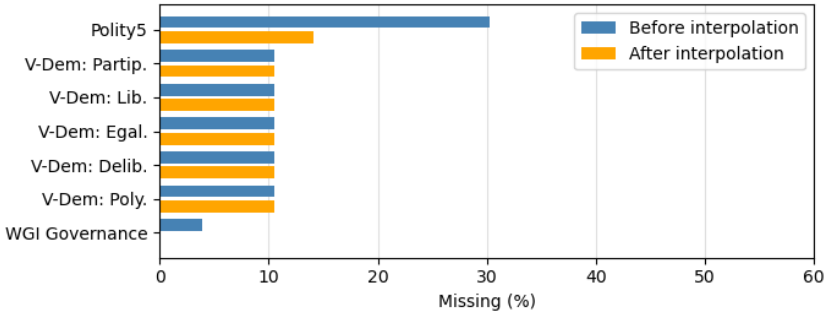


Figure 3: Political node features: Missing data (UN members, 2000-2022).

Economy:

Node data for the economic graph are sourced from the World Bank Open Data platform, focusing on three indicators: GDP (US$) of World Bank (2024a), GDP per capita (US$) of World Bank (2024b), and government expenditure as a percentage of GDP of World Bank (2024d). Only data from the years 2000 to 2022 are used to align with the other graph layers. Aggregate regions and non-country entities (e.g., "World," "Europe & Central Asia") are excluded based on valid ISO3 codes, resulting in 208 country nodes, including microstates and territories with distinct statistical entries (e.g., Monaco, Kosovo, Aruba). Edge data is based on the BACI international trade dataset maintained by CEPII (Gaulier & Zignago, 2010; CEPII, 2025). BACI reconciles UN Comtrade reports to provide consistent directed bilateral flows, where each edge denotes exports from source i to target j. The harmonization process includes estimating Free on Board (FOB)–equivalent values from Cost, Insurance, and Freight (CIF)–reported imports and weighting flows by the reporting reliability of each country. For this study, annual exporter-to-importer trade values from the BACI 202501 release (covering HS92 1995–2023) between 2000 and 2022 (in US$ FOB) are used as edge weights. To reduce skewness and improve model stability, trade values are log-transformed using $\log(1 + x)$. To ensure comparability across years, all three node-level indicators are standardized per year using z-score normalization, centering each feature on its yearly mean and scaling by its standard deviation. Missing values were addressed using linear interpolation within each country–year series. After interpolation, countries with substantial remaining gaps were excluded. A detailed visualization of missingness before and after imputation is provided in Figure 4.

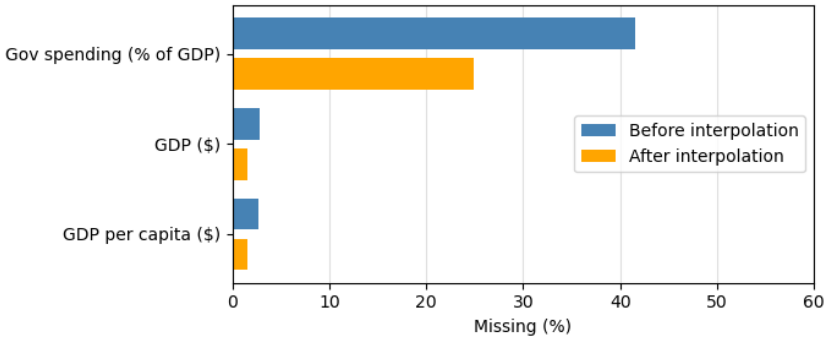


Figure 4: Economics node features: Missing Data (UN Members, 2000-2022).

Culture:

For the culture network, edge features are pairwise cultural distances from the Culturalytics portal (Culturalytics, 2014), computed across multiple dimensions (e.g., Altruism, Norms/Tradition, Politics) and reported for three survey windows (1999–2004, 2005–2009, 2010–2014). For each country pair and target year (2000–2022), we use the value from a matching window; if no exact match exists, the distance from the temporally closest available window is assigned. On the node level, we include Hofstede's cultural dimensions (Hofstede Insights, 2015), expanded to 2000–2022 (e.g., Power Distance (PDI), Individualism vs. Collectivism (IDV), Uncertainty Avoidance (UAI)). Non-numeric entries are treated as missing and imputed using k-nearest neighbors (KNN) with k=5, with all features harmonized to ISO3 codes. The extent of missingness in the Hofstede dimensions before imputation is illustrated in Figure 5.

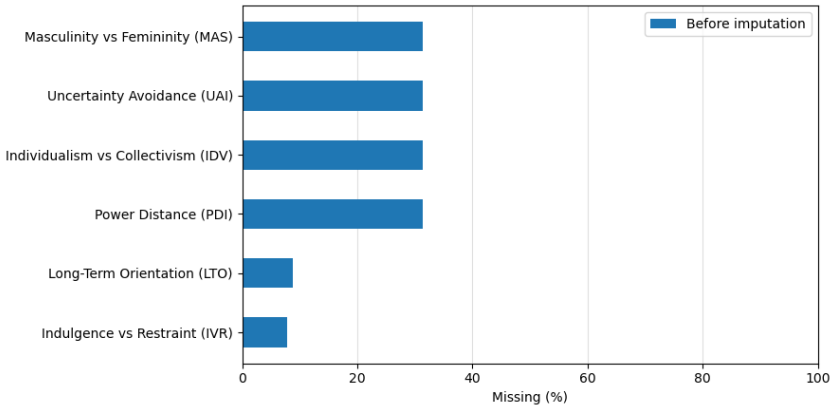


Figure 5: Culture node features (Hofstede): Missing Data (UN Members, 2000-2022).

Language & Religion:

For the linguistic network, edge features are taken from the Domestic and International Common Language (DICL) database (U.S. International Trade Commission, 2024), which provides index measures of linguistic similarity within and between countries. The indices capture common official languages, common native or acquired spoken languages, and linguistic proximity across different languages. Missing values in the language similarity indices are imputed via k-nearest neighbors (k = 5), fitting on all numeric language features and updating only the common spoken language (csl) and linguistic proximity across different spoken languages (lps) indices used downstream. On the node level, we include religious composition data from the Pew Research Center (Pew Research Center, 2020), reported as percentages of major world religions (e.g., Christians, Muslims, religiously unaffiliated). Values from 2010 are assigned to the years

2000–2015, and values from 2020 to 2016–2022. All values are mapped consistently to ISO3 codes.

Similarity Analysis

To visualize similarity relationships between countries across multiple relational layers, we construct smoothed similarity graphs that integrate structural, node-based, and edge-based information. For each of the five layers—geographic, political, cultural, language, and economic—adjacency matrices are derived. Structural similarity is obtained by combining an unweighted adjacency, which records whether a connection exists, with weighted adjacencies that incorporate the strength of numeric edge features; cosine similarity is then applied to these representations. Node attributes and edge feature summaries are likewise standardized and transformed into cosine similarity matrices. Within each layer, these components are merged with equal weight, and the resulting layer-specific similarities are subsequently fused across layers, ensuring consistent alignment of node sets. Finally, temporal smoothing with exponential decay is applied to capture persistent patterns while reducing short-term fluctuations. Figures 6 and 7 display the top-2 similarity graphs for 2018, in which each country has outgoing edges to its two most similar counterparts, with edge thickness indicating the strength of similarity. Figure 6 shows the overall fused graph after temporal smoothing, while Figure 7 depicts the corresponding graphs for each layer individually.

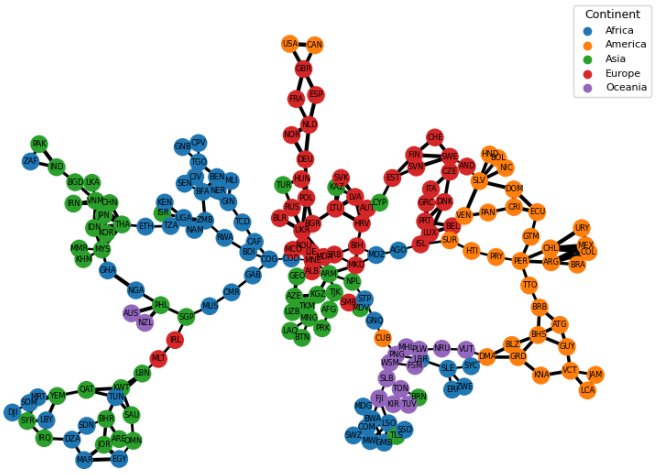


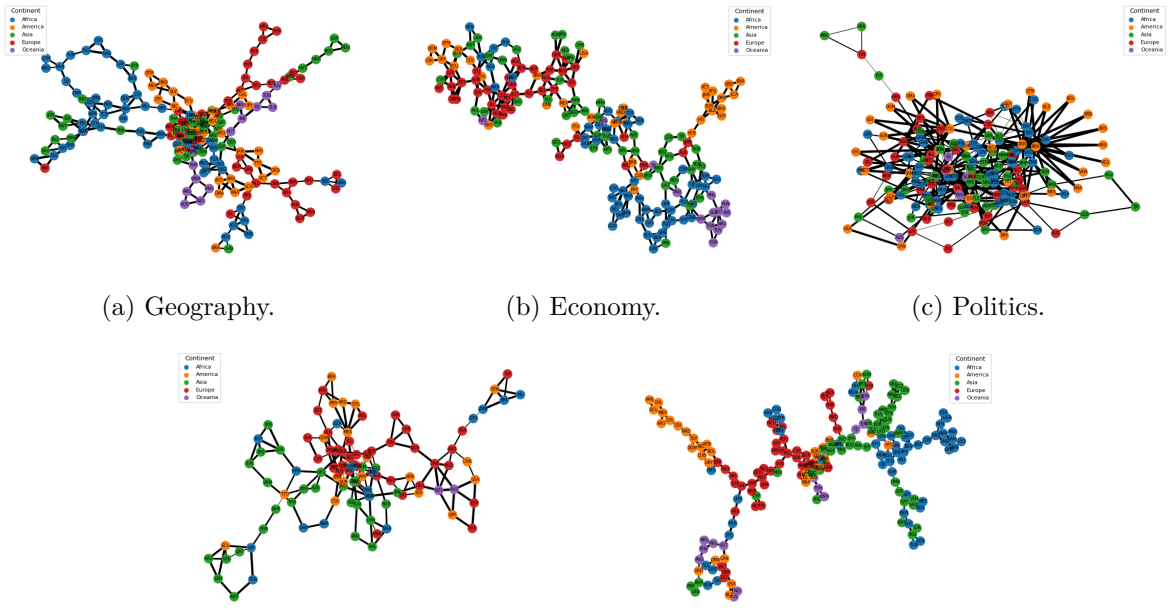
Figure 6: Top-2 Overall similarity graph (2018, smoothed).

Methods

Graph Attention Network in General

Graphs consist of a set of nodes and edges. Nodes represent entities, while edges denote connections between pairs of nodes, indicating that they share some form of relationship. Moreover, both nodes and edges can carry features. These enrich the graph with more specific information about each node and link between the nodes. In our example, countries are represented as nodes, with features such as geographic size, GDP, or spoken language. Edges, in turn, capture relationships between countries, such as diplomatic ties, geographical distance between capitals, or trade volume.

In a Graph Neural Network (GNN), node features are aggregated and combined with the nodes themselves and the overall graph structure to form latent representations, or node embeddings, which are expressed as new multidimensional vectors. The GNN updates these node



(a) Geography. (b) Economy. (c) Politics.

(d) Culture. (e) Language & Religion.

Figure 7: Top-2 similarity graphs (2018, smoothed).

representations only by averaging or summing the feature vectors of the neighboring nodes. By doing so, the model treats every feature of a neighbor as equally important for the specific node. One neighboring node could be mostly irrelevant and would have the same impact on the node embedding as a node which is much more relevant. Generic GNNs do typically not differentiate between neighboring nodes. After updating the node representations through multiple layers, like in other Neural Networks, the resulting embeddings then serve as input for machine learning tasks, such as classification or prediction. A key advantage of GNNs lies in their locality: the graph structure ensures that nodes are more strongly influenced by their immediate neighbors than by distant nodes multiple steps away.

A Graph Attention Network (GAT) addresses some of the limitations of conventional GNNs or Graph Convolutional Networks (GCNs). First, GATs can effectively incorporate edge features, thereby enriching the representation of relationships between nodes. Instead of merely acknowledging the existence of a link between two countries, one can also specify the extent of their relationship—for instance, the degree of cultural similarity, trade intensity, diplomatic cooperation, or geographic proximity. Second, GATs employ an attention mechanism, a concept widely adopted in Transformer architectures and known for its effectiveness in deep learning. Attention enables each node to assign different levels of importance to its neighbors by computing attention scores based on attached edge features. These weights are learnable parameters that are optimized during training. Ultimately, this mechanism produces a more expressive model that captures neighborhood information in a nuanced and flexible manner (Karami, 2023, Turati et al., 2022).

Multi-modal Networks

Our model is based on a multi-modal graph architecture, meaning that we integrate multiple network layers—each representing a distinct dimension of country similarity—to capture their joint influence on inequality outcomes. For each network layer, a separate GAT is trained to generate node embeddings that reflect the structure and relationships within that layer. These layer-specific embeddings are then combined. The resulting joint representation is passed into an MLP that produces the final prediction (see Figure 8). The multi-modal architecture is particularly effective in addressing missing data. Since different network layers may have different

coverage, using multiple sources allows the model to draw on complementary information. For example, when data is missing in one layer (e.g., political or culture networks), the model can still rely on fully observed layers such as geography to build robust country representations.

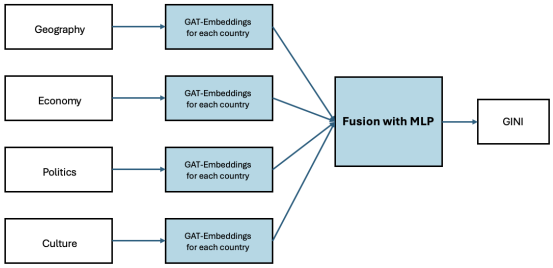


Figure 8: Multi-modal architecture.

To identify the best-performing configuration, we conduct hyperparameter tuning across several components of the pipeline:

- GAT layer parameters
- Embedding fusion method (concatenation vs. attention)
- MLP architecture (shallow, default, deep)
- Learning rate, weight decay, and other optimizer settings

This tuning is performed identically across both approaches described below, resulting parameters are reported in the Training section (Table 1). We approach the above model architecture in two distinct setups, differing in whether the GINI index is used as a node feature during training.

Approach 1: Without GINI as a Node Feature

In this approach, the GINI index is not included in the input features. The dataset is split by year, but not in chronological order. A fixed random seed ensures reproducibility, and the split follows a 60/20/20 ratio for training, validation, and testing. For each network layer and each year, a GAT is trained to generate embeddings. These embeddings are then combined and passed into an MLP to predict GINI values.

Approach 2: With GINI as a Node Feature

Here, the GINI index is included as an input feature for each country in each network to account for real-life applications in which current Gini scores for other (neighboring) countries are available. Because this introduces the risk of information leakage, we apply feature masking instead of temporal data splitting. Specifically, we construct three datasets (training, validation, test) each containing multiple years. In the training set, GINI values are set to zero, and only these zeroed nodes are used in backpropagation. This ensures that the model never learns from actual GINI values during training. As in Approach 1, embeddings from each GAT are combined and passed into an MLP.

Baseline

We implement an Extreme Gradient Boosting (XGBoost) regression model configured with a histogram-based tree-building strategy and the lossguide grow policy, which allows more flexible tree structures by expanding the nodes that contribute most to loss reduction. The model uses a learning rate of 0.03 and up to 2,500 boosting iterations, with early stopping after 50

rounds to prevent overfitting. Each tree is restricted to a maximum of 64 leaves, ensuring controlled complexity, while random subsampling of rows ($subsample = 0.85$) and features ($colsample\_bytree = 0.8$) further regularize the model. Additional shrinkage is applied through L1 ($reg\_alpha = 0.1$) and L2 ($reg\_lambda = 1.0$) penalties. Identical to the main model, the loss function corresponds to the mean absolute error.

To train, validate, and test the model, node-level data from all country–year observations across dimensions are merged into a single dataset, thereby abstracting away the distinctive network structure of individual layers. The final dataset contains 1,733 unique country–year combinations with 51 predictive features (excluding country and year identifiers for generalizability) and the GINI coefficient as the target variable. Data is randomly partitioned into training, validation, and test sets in a 60/20/20 split. Missing feature values are imputed by masking with zero and flagged accordingly, while a standard scaler is fit on the training set and globally applied to scale all numeric features.

Training

During the training process, we employ early stopping and tune a set of hyperparameters for both models. Each training is run for a maximum of 1000 epochs; however, the early stopper consistently terminates training before the maximum is reached. The patience parameter is set to 60, and the minimum delta to 0.0001. The loss is computed using the Mean Absolute Error (MAE). Table 1 presents the hyperparameters applied in both training setups. These values are not the result of manual trial and error but were obtained by using the Optuna framework, which allows to automatically identify the optimal training parameters.

Parameter	**Without GINI**	**With GINI**
Hidden size	148	50
Weight decay	3.41×10^{-7}	1.56×10^{-5}
Learning rate	6.502×10^{-4}	7.833×10^{-4}
Number of GAT layers	1	2
Number of MLP layers	0	1
Combination method	Attention	Concat
Batch size	3	6
MLP variant	Deep	Deep

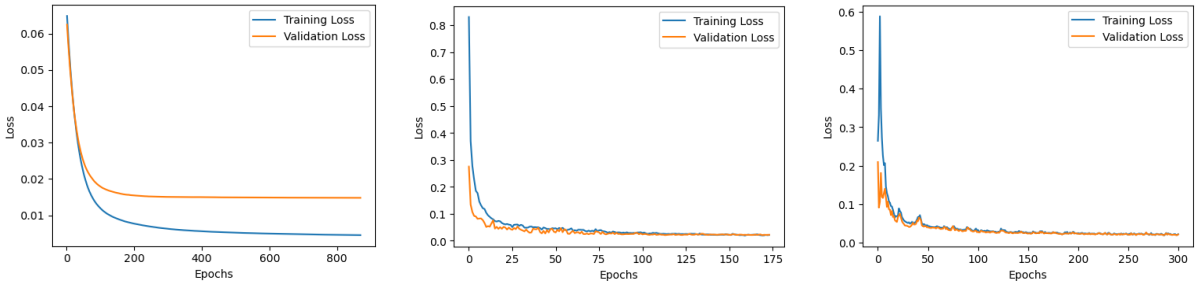
Table 1: Comparison of hyperparameters with and without GINI as predictor.

Optuna

Optuna is a Python library for automated hyperparameter optimization, providing an efficient framework to identify the most suitable training parameters. It performs a predefined number of trials, where each trial consists of training and evaluating the model with a new set of parameters. The combinations of parameters tested are stored in an SQL database, allowing convenient access to the best trial along with all of its corresponding parameters.

Training Curves

The training curves display the MAE loss on the y-axis and the epochs on the x-axis. Figure 9b and Figure 9c illustrate a steep initial decline in the loss followed by an overall smooth convergence. Early stopping was triggered in the model without the GINI index as a predictor at epoch 174, with a validation loss of 0.021. The training curve shows minor signs of overfitting, which does not appear problematic at this stage but should be monitored in future experiments. For the model including the GINI index as a predictor, early stopping occurred later, at epoch 301,



(a) Baseline. (b) Without GINI as predictor. (c) With GINI as predictor.

Figure 9: Training curves.

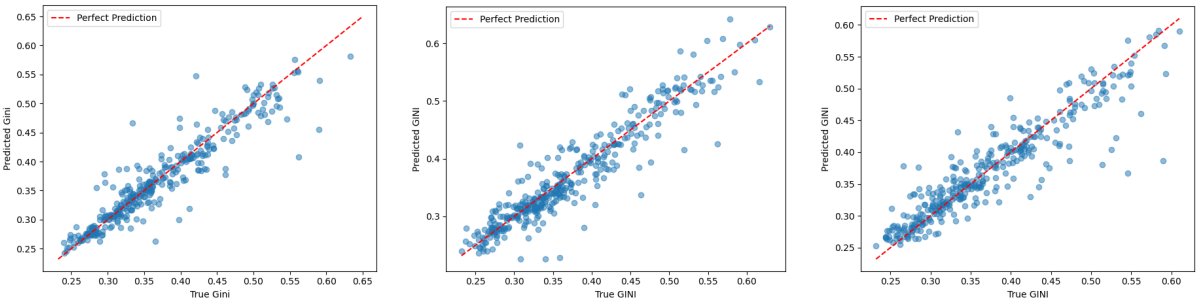
with a validation loss of 0.020. Training for the baseline (Fig. 9a also exhibited a smooth, steep gradient descent, triggering early-stopping after 867 epochs at a validation MAE of 0.015.

Results

Metric	Baseline	Without GINI	With GINI
Mean Absolute Error (MAE)	0.017	0.020	0.021
Root Mean Squared Error (RMSE)	0.028	0.030	0.033
Mean Absolute Percentage Error (MAPE)	4.41%	5.76%	5.91%
R^2 Score	0.883	0.874	0.850

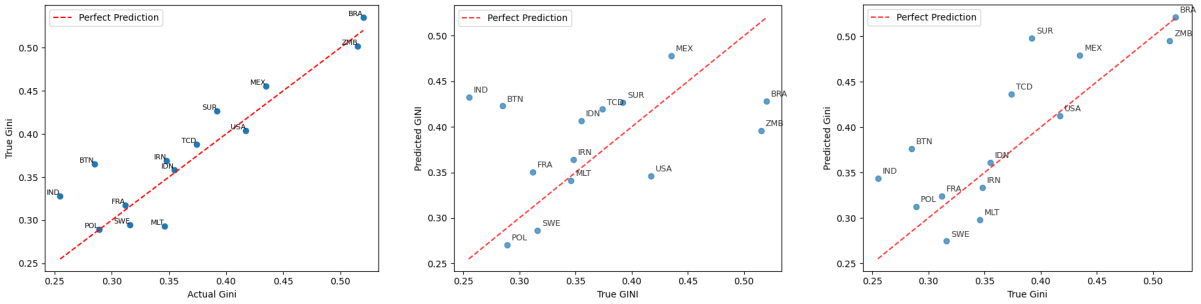
Table 2: Prediction performance on the test set.

Both the GNN ensembles and the baseline perform strongly on the test set for GINI prediction: MAE is in the low hundredths, mean relative deviation is $\leq 6\%$, and R^2 is high, indicating that all three models explain most of the variance in the GINI index (see Table 2). However, the results suggest that enriching country-year information with layer-specific network structure does not improve performance over the baseline. In predictions for selected 2022 test countries with known GINI values, the baseline aligns more closely with the identity line, while both ensemble models—with and without neighboring GINI inputs—exhibit greater scatter (Figure 11). When predicting missing GINI values (Figure 12), only the baseline accurately identifies the high-inequality cluster in Southern Africa (cf. Figure 1). In contrast, the model without GINI appears to underestimate inequality globally and especially in this region, while the model incorporating neighboring GINIs shifts predicted inequality toward Central Africa (specifically South Sudan) and Southeast Asia. These discrepancies highlight key limitations of the structure-based approach.



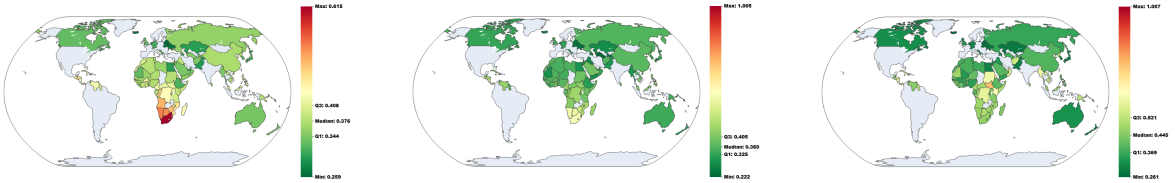
(a) Baseline. (b) Without GINI as predictor. (c) With GINI as predictor.

Figure 10: Predicted vs. true GINI.



(a) Baseline. (b) Without GINI as predictor. (c) With GINI as predictor.

Figure 11: Predicting 2022 test countries.



(a) Baseline. (b) Without GINI as predictor. (c) With GINI as predictor.

Figure 12: Predicting missing GINI values in 2022.

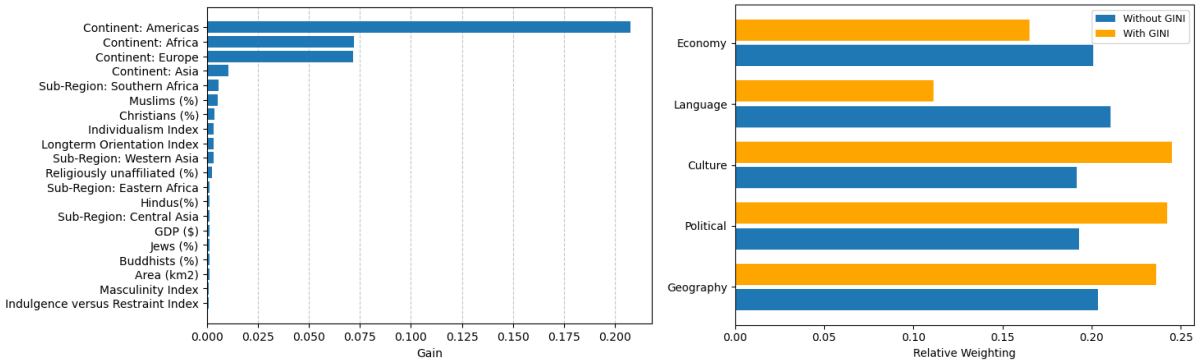
Discussion & Implications for Development Economics Analysis

Because GNNs rely on the structural integrity of country subgraphs constructed from domain-relevant similarities, missingness within clusters or among relatively isolated countries can induce systematic information loss and bias inequality estimates—evident in parts of Africa, where both GINI and predictor data is sparse. Despite leveraging more than two decades of data and multiple similarity dimensions, these structural constraints could not be fully overcome. This highlights a comparative strength of the XGBoost baseline, which—as a tree-based model well-suited for tabular data—can often outperform neural architectures when relational structure is weak or incomplete.

A second limitation concerns evaluation design. Our results pertain to predicting GINI values within the observed historical range: the models were trained on data from both before and after the test years, which favors interpolation. This setup does not mirror genuine forecasting, where only past data is available at inference time. Performance on purely future observations may therefore be lower—especially in regions or periods with sparse or shifting predictors—compounding the structural sensitivity noted above.

Finally, the results are not country-independent. Even without an explicit country identifier, the models can infer country identity from quasi-invariant attributes (e.g., geography, institutions, or other slowly changing features), enabling implicit memorization. This raises the risk of overfitting to country-specific signatures and spurious stability, which can inflate in-sample fit yet weaken generalization—particularly in data-sparse settings, further exacerbating the challenges for true out-of-sample prediction.

Despite its limitations, our approach offers valuable insights that motivate further research. Figure 13a illustrates the baseline model's reliance on geographic features, reflecting the greediness of XGBoost, which prioritizes information gain predominantly from spatial data. Its focus on the Americas and Africa implies that colonial history remains a central driver of global inequality patterns. In contrast, Figure 13b shows the relative modality weights assigned by the MLP prediction head, approximating the importance of each GNN layer. The uninformed variant (without neighboring GINI values) assigns similar importance across all five modalities, while



(a) Features by baseline gain.

(b) Relative MLP weighting of graph layers.

the informed variant emphasizes cultural, political, and geographic dimensions, showing reduced sensitivity to language and economic information.

By integrating multiple similarity dimensions, the model provides a more balanced perspective on the predictors of inequality. While geographic factors may continue to explain persistent clustering, evolving inequality trajectories in developing regions—alongside political and economic shifts—suggest the growing relevance of a multi-dimensional explanatory framework. Hence, a multi-modal approach to structural information not only enriches future research but also informs more nuanced policy strategies that move beyond purely geographic or post-colonial narratives toward a broader understanding of inequality's diverse roots.

Appendix

Code, pre-processed data and visualizations can be retrieved at our puplic GitHub repository.

References

CEPII. (2025). *Baci: International trade database at the product-level (version 202501)* [Last update: 2025-01-30. HS92 (1995–2023), HS96 (1996–2023), HS02 (2002–2023), HS07 (2007–2023), HS12 (2012–2023), HS17 (2017–2023), HS22 (2022–2023).]. Retrieved August 26, 2025, from https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A., Fish, M. S., Fox, L., Gastaldi, L., Gjerløw, H., Glynn, A., Good God, A., Grahn, S., Hicken, A., Kinzelbach, K., Krusell, J., ... Ziblatt, D. (2025). *V-dem [country-year/country-date] dataset v15*. https://doi.org/10.23696/vdemds25

Culturalytics. (2014). *Cultural distance data (1999–2014)* [Survey-based pairwise cultural distances across dimensions such as altruism, norms, and politics. Waves: 1999–2004, 2005–2009, 2010–2014.]. Retrieved August 29, 2025, from https://world.culturalytics.com

Gaulier, G., & Zignago, S. (2010). *Baci: International trade database at the product-level. the 1994–2007 version* (Working Paper No. 2010-23) (CEPII Working Paper No. 2010-23). CEPII. https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=3

Gibler, D. M. (2013). Formal alliances (v4.1) [Dataset recording all formal alliances among states, 1816–2012. Version 4.1 updates Gibler (2009), *International Military Alliances, 1648–2008*.]. https://correlatesofwar.org/data-sets/formal-alliances

Hofstede Insights. (2015). *Hofstede's 6d model of national culture* [Six cultural dimensions: Power Distance (PDI), Individualism (IDV), Masculinity (MAS), Uncertainty Avoidance (UAI), Long-Term Orientation (LTO), Indulgence (IVR).]. Retrieved August 29, 2025, from https://www.kaggle.com/datasets/tarukofusuki/hofstedes-6d-model-of-national-culture

Karami, F. (2023). Understanding graph attention networks: A practical exploration [Accessed: 2025-08-27]. https://medium.com/@farzad.karami/understanding-graph-attention-networks-a-practical-exploration-cf033a8f3d9d

Mayer, T., & Zignago, S. (2011). *Notes on cepii's distances measures: The geodist database* (Working Papers No. 2011-25). CEPII. https://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877

Pew Research Center. (2020). *Religious composition by country, 2010–2020* [Shares of major world religions by country: Christians, Muslims, religiously unaffiliated, Buddhists, Hindus, Jews, other. Data for 2010 and 2020.]. Retrieved August 29, 2025, from https://www.pewresearch.org/religion/feature/religious-composition-by-country-2010-2020/

Polity5 Project. (2018). *Political regime characteristics and transitions, 1800–2018* [Data and codebook available from the Center for Systemic Peace. Polity5 refined data mainly cover 1946–2018; data for 1800–1945 are Polity IV values.]. http://www.systemicpeace.org/inscr/p5v2018.xls

Turati, A., Boennighausen, P., & Shiv, R. (2022). Incorporating edge features into graph neural networks for country gdp predictions [Accessed: 2025-08-27]. https://medium.com/stanford-cs224w/incorporating-edge-features-into-graph-neural-networks-for-country-gdp-predictions-1d4dea68337d

U.S. International Trade Commission. (2024). *Domestic and international common language (dicl) database* [Bilateral indicators of shared official, common, and spoken languages between countries.]. Retrieved August 29, 2025, from https://www.usitc.gov/data/gravity/dicl.htm

World Bank. (2024a). *Gdp (us$)* [Indicator ID: NY.GDP.MKTP.CD. Gross Domestic Product in US dollars.]. Retrieved August 24, 2025, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

World Bank. (2024b). *Gdp per capita (us$)* [Indicator ID: NY.GDP.PCAP.CD. GDP divided by midyear population, in US dollars.]. Retrieved August 24, 2025, from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Bank. (2024c). *Gini index (world bank, poverty and inequality platform)* [Indicator ID: SI.POV.GINI. Based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are mostly from the Luxembourg Income Study database. License: CC BY-4.0.].

World Bank. (2024d). *Government expenditure (% of gdp)* [Indicator ID: GC.XPN.TOTL.GD.ZS. General government total expenditure as a share of GDP.]. Retrieved August 24, 2025, from https://data.worldbank.org/indicator/GC.XPN.TOTL.GD.ZS

World Bank. (2024e). *Worldwide governance indicators, 2024 update* [Accessed on 2024-10-30]. http://www.govindicators.org