Evaluating the Generalisability of Deep Learning Models for Global Fertility Rate Forecasting: A Region-Aware Comparative Analysis

MSc Research Project
MSc Data Analytics: MSCDA_JANOL24

Jair Rodrigues da Silva Junior
Student ID: x23272112

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jair Rodrigues da Silva Junior
Student ID:	x23272112
Programme:	MSc Data Analytics **Year:** 2
Module:	Research Project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	28/01/2026
Project Title:	Evaluating the Generalisability of Deep Learning Models for Global Fertility Rate Forecasting: A Region-Aware Comparative Analysis

Word Count: **6578** **Page Count: 20**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jair Rodrigues da Silva Junior

Date: 28/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluating the Generalisability of Deep Learning Models for Global Fertility Rate Forecasting: A Region-Aware Comparative Analysis

Jair Rodrigues da Silva Junior

Student ID x23272112

Abstract

Global fertility rates have declined considerably in recent decades, leading to ageing populations and shrinking workforces that challenge economic growth and strain healthcare and pension systems. Accurate forecasting of fertility rates is therefore essential for demographic planning and social policy design. This study evaluates the generalisability of seven deep learning architectures (Dense, LSTM, Auto-LSTM, Attention-based LSTM, GRU, BiLSTM, and CNN) using harmonised socioeconomic indicators from 2004 to 2023 across multiple world regions. Spain was used as a reference country for model evaluation and hyperparameter selection. In the Spain pilot, the non-temporal Dense network achieved the lowest prediction error, although several recurrent architectures performed competitively and offered temporal modelling capabilities. A recurrent architecture was therefore selected for cross-regional testing to assess the transferability of learned temporal fertility patterns. When applied to other countries, performance improved substantially under per-country feature rescaling, with weighted MAE reductions ranging from 53% to 97% across regions. The findings show that deep learning models can capture generalisable temporal fertility structures when supported by region-aware preprocessing, and illustrate the policy relevance of improved predictive accuracy: even a small percentage-point reduction in error can translate into thousands of correctly anticipated births in large populations, aiding planning in healthcare, education, and social protection systems.

1 Introduction

Fertility rates have significantly declined worldwide over the past few decades, particularly in Europe and East Asia, where many countries remain well below the replacement level of 2.1 children per woman [1]. Combined with increasing longevity, this decline has accelerated population ageing and placed growing pressure on labour markets, pension systems, and healthcare services [2], [3], [4]. Anticipating these demographic shifts requires reliable fertility forecasts that can support governments and statistical offices in long-term planning.

Recent studies link fertility decline to a combination of socioeconomic and behavioural factors, including rising education levels, urbanisation, labour market participation, housing costs, and access to contraception [2], [3], [4]. These determinants differ markedly across regions, suggesting that fertility dynamics are shaped by nonlinear and region-specific interactions.

Traditional statistical methods, while interpretable, often assume linearity or stationarity and struggle to model complex multivariate fertility relationships. Advances in deep learning offer new opportunities to improve demographic forecasting. Recurrent architectures such as LSTMs, GRUs, and attention-based models have demonstrated strong performance in modelling temporal dependencies in economic and social time series [5], [6], [7], [8], [9], [10], [11], [12]. However, most existing fertility studies evaluate a single model type within a single country, leaving open a key question about how well different architectures generalise across regions, and whether temporal patterns learned in one national context transfer to others.

This study addresses these gaps by conducting a comparative, region-aware evaluation of seven deep learning models (Dense, LSTM, Auto-LSTM, A-LSTM, GRU, BiLSTM, and CNN) using harmonised World Bank indicators from 2004 to 2023. Spain is used as the reference country for model selection, after which the selected model is tested across global regions under two transfer scenarios: strict (using Spain's scaler) and per-country (local rescaling). Performance is assessed using MAE, RMSE, and MAPE.

The central research question guiding this work is: Which deep learning architecture provides the most accurate and generalisable fertility rate forecast across different global regions, and what factors influence model performance?

Through its comparative design, this study provides a benchmark of multiple deep learning architectures for fertility forecasting, a reproducible region-aware transfer framework, and new evidence on how temporal fertility patterns generalise across socioeconomic contexts. It also offers empirical support for integrating deep learning into demographic planning.

The remainder of this report is structured as follows: Section 2 provides the literature review; Section 3 presents the methodology; Sections 4-5 detail design and implementation; Section 6 outlines results and discussion; and Section 7 concludes with implications, limitations, and future research directions.

2 Related Work

2.1 Global Fertility Trends

Global fertility levels have declined significantly across most regions over the past five decades, particularly in Europe and East Asia, where many countries remain significantly under the benchmark of 2.1 children per woman [1], [2]. These combined trends have accelerated population ageing and generated concerns about labour supply, dependency ratios, and long-term sustainability of pension and healthcare systems [3], [4], [13].

The demographic transition has been widely studied. Bongaarts [14] noted that the explosive population growth of the 20th century, fuelled by high fertility and medical advances, was nearing its end, giving way to ageing societies with significant social and economic implications. Fuchs et al. [15] highlighted the pressures this shift places on labour markets, particularly in Europe, where low fertility threatens long-term economic sustainability.

Several studies have explored the socioeconomic drivers of fertility decline. Gotmark and Anderson [3] identified education, economic development, religiosity, and family planning as key factors across six global regions. In China, fertility intentions have been suppressed by rising housing prices, high education costs, and GDP growth [11], [16]. In the United States, Bloome and Ang [17] found that declining marriage and cohabitation rates have contributed to reduced fertility, while Cherlin [18] observed a growing trend of first births occurring outside of marriage among college-educated women, a shift linked to financial constraints and changes in social expectations around family formation.

Policy responses have varied. Wang et al. [19] examined China's transition from a one-child to a three-child policy, noting limited success in reversing fertility decline. In Europe, Kashnitsky et al. [4] emphasised the economic risks of ageing populations and the need for proactive demographic planning. Xie et al. [6] proposed digital inclusive finance as a tool to support family formation and mitigate fertility-related risks.

While these studies offer valuable insights, most focus on isolated national contexts or specific policy interventions. Few provide scalable data-driven forecasting tools that account for regional variation and nonlinear dynamics, a gap this study aims to address through comparative modelling across global regions.

2.2 Traditional Statistical Approaches to Fertility Forecasting

Statistical demographic forecasting methods remain widely used due to their interpretability and long-standing theoretical foundations. ARIMA and cohort-component methods are commonly applied to fertility and demographic time series [8], [15], offering simplicity and transparent parameterisation. However, they usually struggle to capture the multivariate and nonlinear nature of fertility dynamics. ARIMA assumes linearity, stationarity, and univariate structure, which limits its ability to incorporate multiple covariates or capture nonlinear dependencies [11], [10], [20].

The demographic literature also includes a range of more advanced approaches. The Lee-Carter model [21] decomposes age-specific fertility or mortality schedules into a long-term trend and age-specific sensitivity, providing a powerful tool for forecasting age-structured rates. Functional data approaches extend this idea by smoothing fertility curves and applying time series models to the resulting functional components [22]. These methods can capture rich demographic patterns but still rely on relatively rigid parametric assumptions.

Probabilistic models in the Bayesian hierarchical tradition provide an alternative by pooling information across countries to generate uncertainty distributions for future fertility [23]. These frameworks have been adopted in major global population exercises and offer a principled way to quantify uncertainty. Nevertheless, both functional and Bayesian approaches remain constrained by strong structural assumptions and may struggle with high-dimensional multivariate predictors or sudden structural changes.

In general, traditional statistical models provide interpretability but are limited in modelling nonlinear, multivariate, and region-specific dynamics underlying modern fertility trends, motivating the use of machine and deep learning methods.

2.3 Machine Learning Approaches

Machine learning (ML) approaches have been applied to fertility and population studies due to their flexibility in modelling complex relationships. Classification-based studies have used support vector machines, random forests, and logistic regression to identify factors associated with fertility intentions or parity transitions, revealing the importance of variables such as education, employment, and socioeconomic status [5], [10], [13].

At the macro-level, supervised ML algorithms have been used to predict total fertility rates using economic and social indicators. Although these models often outperform linear regressions, many still treat forecasting as a static task with limited ability to capture temporal dependencies [8]. Moreover, most ML studies focus on a single country or region, limiting their ability to evaluate cross-country generalisation.

While ML can accommodate sets of predictors, it does not inherently model sequential structure, highlighting the need for deep learning architectures designed for temporal data.

2.4 Deep Learning Approaches

Deep learning (DL) offers powerful tools for modelling sequential multivariate data, particularly in contexts where traditional methods fall short. Long Short-Term Memory (LSTM) networks have shown strong performance in fertility forecasting, outperforming generic Recurrent Neural Networks (RNNs) in capturing long-term dependencies [8], [12]. However, LSTM models can struggle with feature complexity, especially when multiple socioeconomic indicators interact in nonlinear ways.

To address this, hybrid architectures have emerged. One example is BRP-Net, which integrates attention mechanism with LSTM layers to enhance predictive accuracy in high-dimension datasets [8]. This model was applied to granular data from 361 Chinese cities, enabling fine-grained fertility predictions. While effective, such models rely on detailed local inputs that are not available in many regions, limiting their generalisability.

Other studies have explored alternative enhancements. Wu [9] combined factor analysis with neural networks to reduce dimensionality and improve forecasting precision, while Tang et al. [10] applied a backpropagation neural network to long-term population data, using a multi-phase training and validation strategy. Farahani [24] incorporated genetic algorithms to select key economic indicators before feeding them into an artificial neural network, demonstrating improved accuracy in population projections.

Riiman et al. [12] found that LSTM models performed best when trained on localised data, such as county-level data in the United States. Their findings suggest that fertility forecasting benefits from region-specific training, rather than broad continental aggregations. Similarly, early work by Wu [11] used a time series neural network to predict fertility trends, but relied on a single-variable input, limiting its ability to capture complex interactions.

Despite these advances, most deep learning studies focus on single architecture, typically LSTM, and are tailored to specific national datasets. The literature lacks comprehensive, multi-architecture evaluations comparing models such as GRU, CNN, BiLSTM, or attention-based networks. Existing model comparison studies in fertility forecasting remain limited to small-scale experiments or single-region datasets, offering little insight into how different architectures generalise across different socioeconomic contexts. This gap reinforces the need for multi-architecture benchmarking and motivates the comparative, region-aware evaluation undertaken in this study.

2.5 Summary and Research Gap

The literature demonstrates the growing use of machine learning and deep learning in fertility forecasting, with clear advantages over traditional statistical methods. These methods are better suited to modelling nonlinear relationships, capturing temporal dependencies, and incorporating multivariate socioeconomic indicators. However, existing studies are often constrained by narrow geographic scope, limited model diversity, or reliance on highly granular datasets that restrict wider applicability.

A key limitation is the lack of comparative research evaluating multiple deep learning architectures within a unified experimental framework. Most studies analyse a single model type (typically LSTM) and apply it to one national or subnational dataset. As a result, little is known about how architectures such as GRU, CNN, BiLSTM, or attention-based networks perform relative to one another, or whether their learned temporal structures generalise across different demographic contexts. Similarly, few works examine how data processing choices, such as country-specific rescaling, influence cross-region model transferability.

This study addresses these gaps by comparing seven deep learning models (Dense, LSTM, Auto-LSTM, A-LSTM, GRU, BiLSTM, and CNN) using a harmonised multivariate dataset sourced from the World Bank Data [25]. Spain is used to identify the best-performing architecture, which is then evaluated across global regions under strict and per-country scaling scenarios. By doing so, the research contributes a scalable, multi-architecture benchmarking framework for fertility forecasting and provides new evidence on the conditions under which deep learning models generalise across diverse socioeconomic settings.

3 Research Methodology

This section outlines the research design, data acquisition, preprocessing, modelling techniques, and evaluation procedures to address the study's central question: *Which deep learning architecture provides the most accurate and generalisable fertility rate forecast across different global regions, and what factors influence model performance?*

3.1 Data Sources and Indicators

This study uses harmonised socioeconomic indicators sourced from the World Bank World Development Indicators (WDI) database [25], which provides annually comparable data across more than 200 countries. To ensure consistency and complete feature coverage across all regions, the analysis focuses on the 2004-2023 period. Earlier decades were excluded due to data sparsity and inconsistent reporting.

The target variable is Total Fertility Rate. Predictors were selected through an exploratory correlation analysis conducted by the author using the full WDI panel, combined with domain knowledge about plausible links to fertility behaviour. Indicators with weak or unstable correlations across regions were removed. The final predictor set includes:

- Contributing Family Workers (% of total employment)
- Female Employment in Agriculture (% of female employment)
- Female Employment in Services (% of female employment)
- Health Expenditure per Capita
- GDP per Capita
- Rural Population (% of total population)
- Youth Not in Education, Employment, or Training

- Female Population Aged 15-49

Countries meeting full data availability across all variables were retained and grouped into five global regions: Africa, Asia, Europe, the Americas, and Oceania.

3.2 Data Preprocessing

The raw data underwent several preprocessing steps prior to modelling:

- Missing value treatment: Forward-fill and linear interpolation were applied to address time series gaps in annual reporting.
- Normalisation: Features were scaled using Min-max normalisation to enhance neural network training stability.
- Correlation analysis:
  - Heatmaps were generated per country and region levels.
  - Spain served as the detailed case study for initial exploratory correlation analysis.
- Feature reduction: Indicators exhibiting weak or inconsistent relationships with fertility were excluded to improve model efficiency and interpretability.

3.3 Model Architectures

Seven deep learning architectures were evaluated to determine the most accurate and generalisable model. A simple Dense Neural Network was included as a baseline to benchmark performance against non-sequential deep learning. The architectures evaluated include:

- Dense Neural Network: a fully connected baseline model without temporal recurrence.
- LSTM (Long Short-Term Memory): captures long-range temporal dependencies using gated memory cells.
- Auto-LSTM: Autoregressive LSTM, incorporating automated lag-window selection.
- A-LSTM (Attention-LSTM): augments LSTM outputs with attention weights to highlight influential time steps.
- GRU (Gated Recurrent Unit): streamlined recurrent architecture designed for efficient modelling of sequential patterns.
- BiLSTM (Bidirectional LSTM): processes sequences forward and backward to capture richer temporal structure.
- CNN (1D Convolutional Neural Network): extracts short-term temporal patterns using convolutional filters.

All models were implemented in Jupypter Notebook, using Python 3.10, with libraries including TensorFlow and Keras for model development; Numpy and Pandas for data handling; Seaborn and Matplotlib for visualisation; and Scikit-learn for preprocessing and metrics.

3.4 Baseline Statistical Model: ARIMA

A univariate ARIMA (p,d,q) model was implemented as the statistical baseline. Univariate ARIMA was selected rather than ARIMAX or VAR models due to the short (20-year) annual series and incomplete coverage of exogeneous predictors across countries.

For Spain, the development country, candidate ARIMA orders were evaluated using an AIC-minimising grid search with $p \in \{0,1,2\},\ d \in \{0,1\}, q \in \{0,1,2\}$.

The same grid-search procedure was applied independently to each country, resulting in country-specific ARIMA models with their own optimal (p,d,q) values. Fitting ARIMA separately for each country reflects the fact that ARIMA models do not support cross-country transfer learning.

All ARIMA models were evaluated using a chronological 80/20 train-test split and a one-step-ahead rolling forecast, ensuring methodological consistency with the deep learning evaluation.

3.5 Hyperparameter Tuning

Hyperparameter tuning was carried out using Spain as the development country. For each deep learning model, a shared grid of sequence lengths and hidden units was explored, while other training settings were kept fixed for stability and comparability.

The following configuration was evaluated:

- Architectures tested: Dense, LSTM, GRU, BiLSTM, CNN, A-LSTM, Auto-LSTM
- Sequence lengths (time steps): 3, 5, 7, 10
- Hidden units: 64, 128
- Dropout rate: fixed at 0.2
- Optimiser: Adam
- Loss function: Mean Squared Error (MSE)
- Batch size: 32
- Maximum epochs: 100
- Validation split: 0.2 of the training data
- Early stopping: patience = 10 epochs, with best weights restored

Each combination of model type, time steps, and units, a model was trained and evaluated using Spain's chronological 80/20 train-test split.

The Spain results were used to identify a shortlist of high-performing candidate architectures. Because sequential models are theoretically better suited for temporal transfer across countries, a recurrent model from this high-performing group was selected for cross-regional generalisations (see Section 5).

3.6 Evaluation metrics

To assess model performance and generalisability, the following metrics were considered: Mean Absolute Error (MAE); Root Mean Squared Error (RMSE). These metrics were selected based on their prevalence in forecasting literature and their interpretability for policy applications [3], [4], [12].

A pilot analysis was first conducted using Spain as a case study to identify the most accurate and stable architecture across time and feature sets. The top-performing model was then applied to other countries and regions (Europe, Asia, Africa, Americas, Oceania) without re-tuning, to evaluate its generalisability and transferability. Models were evaluated separately by country and by across five regions to assess generalisability.

3.7 Statistical analysis

Post-training results were analysed using descriptive statistics, such as mean, standard deviation of MAE and RMSE across models and regions. Visualisations such as line plots, box plots, and heatmaps were used to illustrate temporal trends, regional variation and model performance.

3.8 Ethical considerations

All data used were publicly available. No personal or sensitive information was processed. The study complies with ethical standard for secondary data analysis and open data usage under CC BY 4.0.

4 Design Specification

This section presents the architectural and technical design underpinning the fertility rate forecasting. It identifies the frameworks, modelling techniques, and neural architectures used in the study, along with the associated design requirements.

4.1 Framework and tools

The implementation was developed in Python using frameworks and libraries tailored for time series forecasting and deep learning. These include TensorFlow and Keras for model development and training; Scikit-learn for preprocessing, scaling, and metric evaluation; Pandas and Numpy for data wrangling, transformation, and feature engineering; and Matplotlib and Seaborn for visualisation of trends and model outputs.

These tools were selected for their compatibility with sequential modelling, scalability across large datasets, and support for custom architecture design. In addition to deep learning models, the study incorporates ARIMA as a traditional statistical benchmark, enabling comparative evaluation between classical and neural forecasting approaches.

4.2 Model architectures

To evaluate forecasting performance across global regions, seven model architectures were designed and implemented. These include one statistical baseline (ARIMA), one non-sequential neural baseline (Dense), and five sequential deep learning models. This structure enables a layered comparison, assessing both the added value of neural methods over traditional techniques, and the impact of temporal modelling within deep learning itself.

4.2.1 ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is included as a traditional statistical benchmark for fertility forecasting. It models time series by combining autoregressive terms, differencing to achieve stationarity, and moving averages to smooth residuals [26]. ARIMA is well-suited for univariate forecasting and has been widely used in demographic and economic time series analysis.

The general ARIMA model is denoted [26] as ARIMA (p,d,q), where:

- p is the number of autoregressive terms,
- d is the number of nonseasonal differences needed for stationarity
- q is the number of lagged forecast error in the prediction equation.

The model is expressed as:

$$y_t = c + \phi_{1t_{t-1}} + \phi_{2t_{t-2}} + \cdots + \phi_{pt_{t-p}} + \theta_{1\varepsilon_{t-1}} + \theta_{2\varepsilon_{t-2}} + \cdots + \theta_{q\varepsilon_{t-q}} + \epsilon_t$$

Where y_t is the value at time t, ϕ are autoregressive coefficients, θ are moving averages coefficients, and ε_t is the white noise.

In this study, ARIMA was implemented using the *statsmodels* library in Python. Models were fitted independently for each country, with (p,d,q) selected via AIC-minimising grid search. Despite its simplicity and inability to model multivariate or nonlinear relationships, ARIMA provides a valuable interpretative baseline.

4.2.2 Dense (Feedforward Neural Network)

The Dense or feedforward neural network (FNN) serves as a non-sequential deep learning baseline. Unlike recurrent models, Dense networks do not encode temporal order; instead, they learn nonlinear relationships across predictors. Each neuron in a Dense layer receives input from all neurons in the previous layer, enabling complex mapping between socioeconomic predictors and fertility outcomes.

Inspired by the ForecastNet framework proposed by Dabrowski et al. [27], this model adopts a time-variant feedforward design suitable for multi-step forecasting. Dense networks help isolate the value added by deep learning independently of sequential modelling and serve as a reference point for evaluating recurrent and convolutional architectures.

4.2.3 LSTM (Long Short-Term Memory Networks)

LSTM networks are a type of recurrent neural network that captures long-range dependencies using gated memory mechanisms [28]. They address the vanishing-gradient limitations of traditional RNNs through internal stages regulated by forget, input, and output gates [28], [29], [30].

The gating mechanism is defined the following equations [29] :

$f_t = \sigma\left(W_f x_t + U_f h_{t-1}\right)$ (Forget gate)

$i_t = \sigma(W_i x_t + U_i h_{t-1})$ (Input gate)

$C' = tanh(W_C x_t + U_C h_{t-1})$ (Candidate state)

$C_t = f_t \odot C_{t-1} + i_t \odot C'_t$ (Cell state update)

$o_t = \sigma(W_o x_t + U_o h_{t-1})$ (Output gate)
$h_t = o_t \odot \tanh(C_t)$ (Hidden state)

Where x_t is the input at time t, h_t is the hidden state, C_t is the cell state, σ is the sigmoid activation function, W and b are learnable weights and biases, $tanh$ is the hyperbolic tangent function, $\odot$ denotes the element-wise multiplication, W and U are learnable weight matrices for input and hidden state respectively.

These equations follow the standard formulation used in deep learning methods as TensorFlow and PyTorch [29], and are consistent with the conceptual breakdown presented by Staudemeyer and Morris [30], who trace the evolution of LSTM from traditional RNNs and highlight its capacity for long-term memory retention.

LSTMs were implemented using TensorFlow and Keras. A uniform grid search explored time steps (3,5,7, and 10) and hidden units (64 and 128). Each model was trained on a country-specific multivariate time series data.

4.2.4 Auto-LSTM (Autoregressive LSTM)

Auto-LSTM augments the LSTM by including lagged values of the target variable, combining autoregressive structure with nonlinear recurrence [31], [32]. At each time step, lagged fertility values (e.g., $x_{t-1}, x_{t-2}, x_{t-3}$), are concatenated with exogenous features, following the supervised-learning formulation described by Brownlee [33].

The model uses the same LSTM formulation as Section 4.2.3, with the main difference being the enriched input vectors. Hyperparameter tuning followed same grid search as other deep learning models.

4.2.5 A-LSTM (Attention-based LSTM)

The A-LSTM integrates an attention mechanism on top of an LSTM encoder, allowing the model to focus on the most relevant time steps when generating predictions [8], [34]. Each A-LSTM cell maintains the same structure as the standard LSTM, and uses three gates to regulate information flow [28], [29], [30]. These gates are defined by the same equations presented in Section 4.2.3.

After the hidden states $h_1, h_2, \dots, h_T,$ is generated by the LSTM layer, the attention mechanism computes a set of attention weights α_t for each time step, determining its relative importance to the final output. The attention process is formally defined as follows [8]:

$$e_t = v^t tanh(W_h h_t + b_h)$$

$$\alpha_t = \frac{exp(e_t)}{\sum_{i=1}^{T} exp(e_i)}$$

$$c_t = \sum_{t=1}^{T} \alpha_t h_t$$

Where h_t is the hidden state of the LSTM at time t, W_h, v, and b_h are learnable parameters, e_t represents the alignment score of each time step, α_t are the normalised attention weights, and c_t is the context vector computed as weighted sum of all hidden states.

The context vector c_t is passed to a Dense layer for prediction. A-LSTM was implemented using a custom TensorFlow attention layer and trained with the same hyperparameter grid as other architectures.

4.2.6 GRU (Gated Recurrent Unit)

GRUs simplify structures by combining gates and eliminating the separate cell state [35]. This reduces computational cost while maintaining the ability to capture long-range dependencies. The update and reset gates are defined as:

$$Z_t = \sigma(W_Z x_t + U_Z h_{t-1}) \qquad \text{(Update gate)}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \qquad \text{(Reset gate)}$$

$$h'_t = tanh\big(W_h x_t + \ U_r\ (r_t \odot h_{t-1})\big) \qquad \text{(Candidate activation)}$$

$$h_t = \ (1 - Z_t) \odot h_{t-1} + Z_t \odot h'_t \qquad \text{(Final hidden state)}$$

Where x_t is the input at time t, h_t is the hidden state, σ is the sigmoid activation function, $\odot$ denotes the element-wise multiplication, W and b are learnable weight matrices for input and hidden state respectively

GRU models were implemented using TensorFlow/Keras and trained using same grid as LSTM, Auto-LSTM, and BiLSTM.

4.2.7 BiLSTM (Bidirectional LSTM)

BiLSTMs process sequences in both backward and forward directions, combining the hidden states at each time step:

$$h_t^{Bi} = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

Where $h_t^{\rightarrow}$ is the hidden state from the forward LSTM, $h_t^{\leftarrow}$ is the hidden state from the backward LSTM, $[.\,;.\,]$ denotes vector concatenation.

This structure provides richer temporal context and has been applied in demographic and macroeconomic forecasting [31], [32]. BiLSTMs were implemented in TensorFlow/Keras and evaluated using the same hyperparameter grid as other recurrent models.

4.2.8 CNN (Convolutional Neural Network)

CNNs for 1D time series extract local temporal features using convolutional filters that slide across the input sequence [36]. This allows efficient parallel computation and makes CNNs well-suited for capturing short-term fluctuations or abrupt local changes in fertility patterns.

Although CNNs do not maintain hidden-state memory, they offer strong performance for short-term pattern extraction [26], [36]. CNNs were implemented using TensorFlow/Keras and evaluated alongside all other models to assess the role of non-recurrent architectures in fertility forecasting.

5 Implementation

This section outlines the practical implementation of the machine learning framework developed to forecast fertility rates across countries and regions. All work was conducted in Python using TensorFlow and Keras for deep learning models, and the *statsmodels* library for the traditional ARIMA baseline. Data wrangling, model training, and evaluation were performed in Jupyter Notebook.

The implementation followed a staged workflow, beginning with data preparation and exploratory analysis, the Spain pilot study, model configuration, and finally global generalisation testing.

5.1 Data preparation

The dataset, obtained from the World Bank API [25], included annual fertility rates and selected socio-economic indicators for all available countries. Preprocessing steps included interpolation of missing values, min-max normalisation of numerical features, and chronological data splitting (80% training, 20% testing) to preserve temporal dependencies.

The modelling pipeline used a cleaned and pre-processed dataset restricted to 2004-2023. The following variables were retained after correlation analysis:

- Target variable: Fertility Rate (births per woman)
- Predictors: Female population ages 15-49; GDP per capita; health expenditure per capita; female employment in agriculture; female employment in services; contributing family workers; rural population; and youth not in education.

Missing values were imputed using forward/backward filling, and features were normalised using Min-Max scaler fitted within each training set.

5.2 Experimental workflow

To ensure methodological consistency, the implementation followed a staged process:

1. Data filtering and scaling: Input data were normalised between 0 and 1, ensuring feature comparability.
2. Time-series windowing: Each model received sequences of past fertility and socio-economic indicators as input, using a sliding window approach.
3. Model building: Multiple architectures were defined programmatically to ensure consistent structure and parameter control.
4. Model training and validation: Each model was trained using an 80/20 chronological split, maintaining temporal consistency. Early stopping was used to prevent overfitting.
5. Evaluation and selection: Models were compared based on MAE, RMSE, and MAPE

5.3 Spain pilot study

Spain was randomly selected as the reference country for model prototyping and hyperparameter tuning. All models were trained using Spain's complete 2004-2023 series to establish baseline performance, compare architecture performance, and to identify a suitable sequential model for subsequent generalisation experiments.

5.4 Experimental configuration and hyperparameter design

To ensure consistent and systemic model evaluation, a shared configuration grid was applied across all deep learning architectures.

- Time-step windows: 3, 5, 7, and 10 years.
  Capture short-, medium-, and long-term temporal dependencies.
- Hidden units: 64 and 128
  Evaluate model capacity and generalisation ability
- Batch size: 16
- Epoch: Up to 60, with early stopping (patience = 8).
- Optimiser: Adam, with MSE as the loss function.
- Validation split: chronological 80% training, 20% validation.

Each model-configuration combination was evaluated using Spain's data based on MAE, RMSE, and MAPE. The evaluation on Spain informed the selection of suitable recurrent architecture for transfer testing, without drawing conclusions about final performance.

5.5 Model implementation

The eight models outlined Section 4.2 were implemented to test a full range of temporal and non-temporal learning capabilities:

Table 5.1: Models' description

Model	**Description**
ARIMA	Classical autoregressive statistical baseline
Dense	Feedforward network serving as deep learning baseline
LSTM	Standard Long-Short Term Memory network capturing temporal dependencies
Auto-LSTM	LSTM with autoregressive connection for recursive fertility prediction
A-LSTM	Attention-based LSTM that dynamically weights temporal importance
BiLSTM	Bidirectional LSTM allowing both past and future temporal context.
GRU	Gated Recurrent Unit, a computationally efficient LSTM variant.
CNN	Convolutional network capturing short-term dependencies in sequential data.

5.6 Generalisation workflow

Following the Spain-based evaluation, a single recurrent architecture was selected for cross-regional generalisation to assess model transferability across diverse fertility contexts.

The cross-regional experiment tested how well a model trained on one country could forecast fertility patterns in other regions. Two evaluation settings were applied:

1. Strict generalisation: The original Spain-trained model as its scaler were applied to all countries without adaptation.
2. Per-country generalisation: The same model weights were used, but a new Min-Max scaler was fitted per country to adjust for national feature distributions.

Performance was evaluated both at the country level and region level, with regional scores weighted by female population aged 15-49 to ensure demographic representativeness.

6 Evaluation and discussion

6.1 Model evaluation using Spain

The first stage of the evaluation assessed all model architectures using Spain's data (2004-2023). To ensure transparency in model comparison, the top 20 configurations ranked by MAE are reported in Table 6.1. Results show that although Dense network achieved the lowest error for Spain, several recurrent models including GRU, LSTM and BiLSTM also appeared within the upper tier of performers, with relatively small differences in absolute values.

This reflects the limited size and smooth behaviour of Spain's fertility series, which narrows the performance gap between sequential and non-sequential models. Importantly, sequence-based models demonstrated stable forecasting accuracy and remained competitive across a range of time-step and unit configurations.

Table 6.1 – Spain experiments: Top 20 results

Seq	Model	Time Steps	Units	MAE	RMSE	MAPE
1	Dense	3	128	0.0036	0.0053	0.3099
2	Dense	7	64	0.0078	0.0123	0.6948
3	Dense	3	64	0.0117	0.0136	1.0087
4	LSTM	3	128	0.0127	0.0182	1.1188
5	LSTM	3	64	0.0133	0.0179	1.1656
6	ALSTM	3	128	0.0137	0.0155	1.1925
7	LSTM	7	128	0.0141	0.0185	1.2427
8	BiLSTM	3	128	0.0147	0.0195	1.2861
9	CNN	5	64	0.0158	0.0165	1.3616
10	LSTM	10	128	0.0164	0.023	1.4597
11	**GRU**	**5**	**64**	**0.0164**	**0.021**	**1.444**
12	BiLSTM	10	128	0.0167	0.0231	1.4942
13	GRU	3	64	0.0169	0.0203	1.4697
14	ALSTM	5	128	0.0175	0.0207	1.5343
15	CNN	3	64	0.0177	0.0184	1.5311
16	GRU	10	128	0.0184	0.0231	1.6337
17	BiLSTM	7	64	0.0188	0.0249	1.6543
18	BiLSTM	5	128	0.0192	0.0206	1.6605

19	ALSTM	10	64	0.0197	0.0207	1.7396
20	CNN	7	64	0.0198	0.0213	1.7283

Because the purpose of the pilot study was to identify a robust architecture for cross-regional generalisation, the selection of the GRU was based not only on Spain's ranking but also on its strong theoretical suitability for multivariate temporal forecasting, computational efficiency relative to LSTM and BiLSTM, consistently competitive performance across multiple configurations, and suitability for transfer experiments involving different fertility regimes.

Thus, while Dense achieved the best single-country performance, the GRU was chosen as the primary architecture for generalisation testing due to its balanced accuracy, stability, and capacity to model temporal dependencies across diverse socio-demographic contexts.

6.2 Cross-regional generalisation

After completing the Spain pilot study, the GRU architecture, identified as a strong and stable performer among the top-ranked models, was selected for cross-regional generalisation analysis.

To evaluate generalisability, the Spain-trained GRU was tested on all remaining countries using two strategies:

1. Strict generalisation: Spain's scaler and weights were applied to all countries without adaptation.
2. Per-country generalisation: Model weights remained fixed, but a new Min-Max scaler was fitted independently for each country to adjust for differences in feature magnitude.

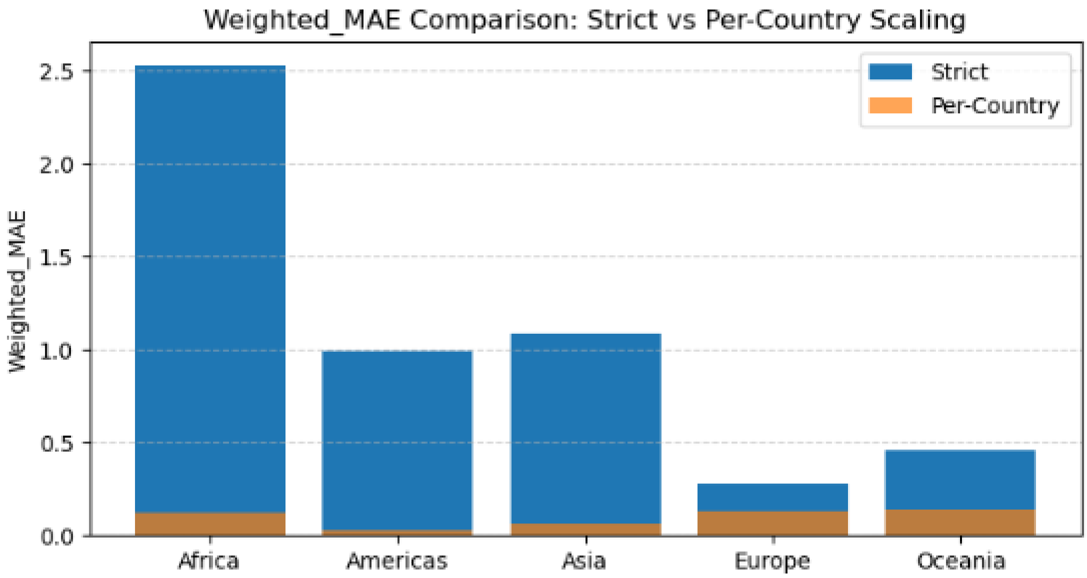
Both country-level and region-level performance were computed. Regional summaries were weighted by the female populations aged 15-49 to reflect demographic representativeness.

6.3 Weighted regional performance

To evaluate the extent to which the Spain-trained model generalises across global regions, weighted regional error scores were computed using female population aged 15-49 as a demographic weight. This ensures that regions with larger populations contribute proportionally to the aggregated metrics.

Table 6.2 – Weighted average performance by region

Region	Weighted MAE (Strict)	Weighted MAE (Per-country)	Delta (%)
Africa	2.5287	0.1203	-95%
Americas	0.994	0.0264	-97%
Asia	1.0927	0.0591	-95%
Europe	0.2798	0.1307	-53%
Oceania	0.4622	0.1402	-70%



Across all five regions, per-country scaling substantially reduced error, confirming that local differences in feature magnitudes and value distributions affect transferability. Several consistent patterns emerge:

- Africa and Asia show the largest absolute improvements (95%), reflecting substantial heterogeneity in socio-economic indicators and broader fertility ranges that differ from Spain's.
- The Americas achieved the highest improvement (97%), indicating that the model's learned temporal patterns were highly compatible once feature scaling was aligned.
- Europe shows the smallest improvement (53%), suggesting that its overall feature distributions are more similar to Spain's even under strict transfer.
- Oceania improved by 70% though results are influenced by the region's small population and limited country sample.

Overall, the results demonstrate that the learned temporal dependencies generalise well across regions, but accurate transfer required adapting the input scaling to each country's numerical range. Without this adjustment, prediction errors increase substantially, particularly in regions with higher data variability.

6.4 Discussion and findings

The key insights emerge from the results:

1. Temporal patterns are transferable across context.

   The Spain-trained model was able to reproduce broad fertility trends in multiple regions, indicating that some underlying temporal structures in fertility change are shared across countries. This suggests that temporal dependencies can transfer across countries.

2. Feature scaling is critical to generalisation

   The large reductions in error under the per-country rescaling condition show that while temporal relationships may transfer, differences in the magnitude of socio-economic indicators can distort predictions unless adjusted.

3. Socio-demographic proximity enhances performance.

   Regions with fertility patterns and data characteristics close to Spain (e.g., Europe) exhibited lower errors, consistent with demographic theory that countries at similar stages of the fertility transition tend to display more comparable dynamics [37].

6.5 Comparison with baselines

When compared with traditional and non-sequential baselines, the sequential deep learning models generally showed stronger capacity to capture temporal dependencies relevant to fertility change. The ARIMA model offered a useful statistical benchmark but was limited by its univariate and linear structure. The Dense network, although competitive in the Spain experiment, lacked explicit temporal modelling and therefore provided a weaker foundation for cross-regional transfer.

Among the sequential architectures, several models such as LSTM, GRU, and BiLSTM demonstrated consistently lower errors than ARIMA and Dense in the Spain pilot study, reflecting the importance of learning temporal structure in fertility series. Because only one model configuration could be carried forward to the generalisation stage, a single sequential architecture was selected for cross-regional testing to ensure methodological clarity and comparability.

Overall, the results support the value of sequential deep learning approaches over classical baselines for modelling fertility trajectories, while recognising that cross-regional performance depends strongly on preprocessing and country-specific data characteristics rather than architecture alone.

6.6 Summary

Overall, the GRU demonstrated forecasting accuracy for Spain and robust generalisation across global regions, particularly when local feature scaling was applied. The results support the feasibility of deep learning for global fertility forecasting and highlight the importance of region-sensitive preprocessing.

These findings directly address the study's central research by identifying GRU as the most accurate and generalisable architecture for fertility forecasting using harmonised international data.

7 Conclusion and Future Work

7.1 Conclusions

This study evaluated the generalisability of deep learning models for global fertility rate forecasting, evaluating whether temporal neural architectures can capture fertility dynamics that transfer across diverse socio-economic contexts. Eight models were implemented (ARIMA, Dense, LSTM, GRU, BiLSTM, Auto-LSTM, A-LSTM, and CNN), using harmonised indicators from the World Bank Open Data [25]. A structured workflow was followed, including a Spain-trained pilot phase to evaluate configuration behaviour, and cross-regional tests under both strict and per-country scaling.

The research achieved its core objectives:

1. Develop and compare multiple architectures: All statistical and deep learning models were implemented using a consistent experimental framework
2. Assess behaviour and stability on national data: Although the Dense (non-temporal) model achieved the lowest error in the Spain experiment, several recurrent architectures (such as LSTM, GRU, BiLSTM) performed competitively and demonstrated the importance of modelling temporal structure for longitudinal fertility dynamics.
3. Examine cross-regional generalisability: When a recurrent model trained on Spain was transferred to other countries, it demonstrated the ability to reproduce broad fertility patterns, particularly when per-country rescaling was applied.
4. Analyse regional variation in performance: Regions with more stable socio-economic indicators (e.g., Europe) showed stronger generalisation performance, while regions with greater heterogeneity exhibited larger error variability.

Overall, the findings demonstrate that temporal deep learning models can capture fertility dynamics that extend beyond national boundaries, and that feature scaling is essential for effective cross-regional transfer.

7.2 Implications and efficacy

The study offers methodological, empirical, and policy-relevancy contributions:

- Methodological: It provides a reproducible pipeline, demonstrating how multiple deep learning architectures can be compared using hyperparameters, harmonised data, and a unified preprocessing workflow.
- Empirical: The cross-regional results show that fertility trajectories contain temporal structures that can be learned from one context and applied to others, supporting the idea of transferable demographic patterns.
- Policy relevance: Improved forecasting accuracy enables better anticipation of demographic change, supporting planning in health, labour markets, education, and family policy.

7.3 Limitations

Several limitations should be acknowledged:

1. Data constraints: The analysis relied on World Bank indicators [25], which contain missing or interpolated values and varying update frequencies across countries.
2. Temporal depth: The 2004-2023 period may limit the model's ability to capture long demographic cycles or structural shifts.
3. Model interpretability: Deep learning models are inherently less transparent, limiting insight into how individual predictors drive forecasts.
4. Uniform configuration grid: All architectures were trained using the same hyperparameter ranges, and region-specific tuning may yield improved accuracy.
5. Socio-cultural factors: Important variables such as family policies, cultural norms, religion, and migration patterns were not included due to limited global data availability.
6. A formal robustness analysis (e.g., sensitivity to random seeds, alternative train-test splits, or noise perturbation) was not conducted, which limits the ability to assess model stability under different experimental conditions.

Recognising these limitations ensures transparency and identifies opportunities to strengthen future demographic forecasting models.

7.4 Future work

Building on current findings, several areas are proposed for future research:

1. Develop hybrid global-local transfer models: Adapt pre-trained recurrent models using region-specific data to improve accuracy in data-limited context.
2. Feature relevance and sensitivity analysis: Apply ablation studies, SHAP values, or correlation-based methods to identify the most influential socioeconomic indicators.
3. Explore automated hyperparameter optimisation methods, such as Bayesian optimisation or AutoML frameworks, to replace the fixed grid search and enable region-specific tuning based on data.
4. Conduct a comprehensive robustness evaluation, including repeated experiments with different random seeds, alternative temporal splits, and input perturbations, to assess the stability and reliability of the forecasting models.
5. Integration of additional behavioural or policy indicators: Incorporate data on childcare, education levels, housing costs, or cultural norms to improve contextual modelling.
6. Extend temporal forecasting horizons: Combine neural forecasts with demographic projections or ensemble approaches for longer-term outlooks.

7. Ensemble and hybrid architectures: Explore combinations of recurrent networks and attention-based models to capture both local and long-range temporal dependencies.
8. Interdisciplinary collaboration: Work with demographers and economists to refine feature selection, validate assumptions, and improve domain interpretation.

7.5 Final remarks

This study demonstrates that deep learning provides a promising and adaptable approach to modelling fertility trajectories across countries and regions. Through harmonised data, consistent evaluation, and cross-regional testing, the research strengthens empirical understanding of how fertility patterns can be represented computationally. While not intended to redefine demographic theory, the work offers a methodological contribution and practical foundation for developing scalable, data-driven tools for demographic forecasting.

References

[1] R. J. Aitken, “The Global Decline in Human Fertility: The Post-Transition Trap Hypothesis,” *Life*, vol. 14, no. 3. Multidisciplinary Digital Publishing Institute (MDPI), Mar. 2024. doi: 10.3390/life14030369.

[2] R. J. Aitken, “What is driving the global decline of human fertility? Need for a multidisciplinary approach to the underlying mechanisms,” *Frontiers in Reproductive Health*, vol. 6. Frontiers Media SA, 2024. doi: 10.3389/frph.2024.1364352.

[3] F. Götmark and M. Andersson, “Human fertility in relation to education, economy, religion, contraception, and family planning programs,” *BMC Public Health*, vol. 20, no. 1, Feb. 2020, doi: 10.1186/s12889-020-8331-7.

[4] I. Kashnitsky, J. D. Beer, and L. V. Wissen, “Economic Convergence In Ageing Europe,” *Tijdschr. Voor Econ. En Soc. Geogr.*, vol. 111, no. 1, pp. 28–44, Feb. 2020, doi: 10.1111/tesg.12357.

[5] E. A. Kassaw, B. B. Abate, B. M. Enyew, and A. K. Sendekie, “The application of machine learning approaches to classify and predict fertility rate in Ethiopia,” *Sci. Rep.*, vol. 15, no. 1, p. 2562, Dec. 2025, doi: 10.1038/s41598-025-85695-8.

[6] Y. Xie, J. Guo, and X. Ma, “Digital inclusive financial and household fertility:Discoveries based on dual machine learning algorithm,” *Int. Rev. Econ. Finance*, vol. 97, Jan. 2025, doi: 10.1016/j.iref.2024.103747.

[7] X. Guo *et al.*, “Prediction and unsupervised clustering of fertility intention among migrant workers based on machine learning: a cross-sectional survey from Henan, China,” *BMC Public Health*, vol. 25, no. 1, p. 140, Dec. 2025, doi: 10.1186/s12889-025-21412-4.

[8] M. Xue, J. Zhu, R. Wu, X. Zhang, and Y. Chen, “BRP-Net: A discrete-aware network based on attention mechanisms and LSTM for birth rate prediction in prefecture-level cities,” *PLoS ONE*, vol. 19, no. 9 September, Sept. 2024, doi: 10.1371/journal.pone.0307721.

[9] J. Wu, “Birth Rate Prediction in China Based on Combination Algorithm of Factor Analysis and Neural Network,” in *2024 5th International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 875–879. doi: 10.1109/ICBASE63199.2024.10762744.

[10] X. Tang, X. Cai, R. Zhang, and Y. Jia, “Research and Simulation of Population Forecast Based on BP Neural Network,” in *Proceedings - 2022 2nd International Conference on Electronic Information Engineering and Computer Technology, EIECT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 302–305. doi: 10.1109/EIECT58010.2022.00066.

[11] J. Wu, “Prediction of Birth rate in China under three-child Policy based on neural network,” in *2022 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1652–1655. doi: 10.1109/ICSP54964.2022.9778548.

[12] V. Riiman, A. Wilson, R. Milewicz, and P. Pirkelbauer, “Comparing Artificial Neural Network and Cohort-Component Models for Population Forecasts.”

[13] K. Moulaei, M. Mahboubi, S. G. Kalkhajeh, and H. Kazemi-Arpanahi, “Prediction of childbearing tendency in women on the verge of marriage using machine learning techniques,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-71854-w.

[14] J. Bongaarts, “Human population growth and the demographic transition,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1532, pp. 2985–2990, Oct. 2009, doi: 10.1098/rstb.2009.0137.

[15] J. Fuchs, D. Söhnlein, B. Weber, and E. Weber, “Stochastic Forecasting of Labor Supply and Population: An Integrated Model,” *Popul. Res. Policy Rev.*, vol. 37, no. 1, pp. 33–58, Feb. 2018, doi: 10.1007/s11113-017-9451-3.

[16] H. Zhang, Y. Yang, and S. Sun, “It is not a vacation of leisure: The effect of maternity-leave stigma on fertility intentions in China,” *Acta Psychol. (Amst.)*, vol. 248, Aug. 2024, doi: 10.1016/j.actpsy.2024.104413.

[17] D. Bloome and S. Ang, “Marriage and Union Formation in the United States: Recent Trends Across Racial Groups and Economic Backgrounds,” *Demography*, vol. 57, no. 5, pp. 1753–1786, Oct. 2020, doi: 10.1007/s13524-020-00910-7.

[18] A. J. Cherlin, “Rising nonmarital first childbearing among college-educated women: Evidence from three national studies”, doi: 10.1073/pnas.2109016118/-/DCSupplemental.

[19] Y. Wang, H. Di, and Q. Duan, “Prediction of Preschool Teacher Demand Under China’s Three-Child Policy: A Case Study of Shanghai,” *Early Child. Educ. J.*, Feb. 2023, doi: 10.1007/s10643-023-01604-6.

[20] D. Li, M. Qiu, and Z. Luo, “Huizhou resident population, Guangdong resident population and elderly population forecast based on the NAR neural network Markov model,” *AIMS Math.*, vol. 9, no. 2, pp. 3235–3252, 2024, doi: 10.3934/math.2024157.

[21] R. D. Lee and L. R. Carter, “Modeling and forecasting U.S. mortality.,” *J. Am. Stat. Assoc.*, vol. 87, 419, pp. 659–671, Sept. 1992.
[22] R. J. Hyndman and Md. Shahid Ullah, “Robust forecasting of mortality and fertility rates: A functional data approach,” *Comput. Stat. Data Anal.*, vol. 51, no. 10, pp. 4942–4956, June 2007, doi: 10.1016/j.csda.2006.07.028.
[23] L. Alkema *et al.*, “Probabilistic Projections of the Total Fertility Rate for All Countries,” *Demography*, vol. 48, no. 3, pp. 815–839, Aug. 2011, doi: 10.1007/s13524-011-0040-5.
[24] M. S. Farahani, “Applications of Artificial Intelligence in Social Science Issues: a Case Study on Predicting Population Change,” *J. Knowl. Econ.*, vol. 15, no. 1, pp. 3266–3296, Mar. 2024, doi: 10.1007/s13132-023-01270-4.
[25] W. B. Group, “World Bank Open Data.” [Online]. Available: https://data.worldbank.org/
[26] M. Peixeiro, *Time Series Forecasting in Python*, 1st edition. CA, USA: Manning, 2022.
[27] J. J. Dabrowski, Y. Zhang, and A. Rahman, “ForecastNet: A Time-Variant Deep Feed-Forward Neural Network Architecture for Multi-Step-Ahead Time-Series Forecasting,” vol. 12534, 2020, pp. 579–591. doi: 10.1007/978-3-030-63836-8_48.
[28] Y. Kong *et al.*, “Unlocking the Power of LSTM for Long Term Time Series Forecasting,” Feb. 24, 2025, *arXiv*: arXiv:2408.10006. doi: 10.48550/arXiv.2408.10006.
[29] S. Lee, “The Ultimate LSTM Guide for Time Series Modeling,” The Ultimate LSTM Guide for Time Series Modeling. Accessed: Oct. 04, 2025. [Online]. Available: https://www.numberanalytics.com/blog/ultimate-lstm-guide-time-series-modeling
[30] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks,” Sept. 12, 2019, *arXiv*: arXiv:1909.09586. doi: 10.48550/arXiv.1909.09586.
[31] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “The Performance of LSTM and BiLSTM in Forecasting Time Series,” in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA: IEEE, Dec. 2019, pp. 3285–3292. doi: 10.1109/BigData47090.2019.9005997.
[32] A. Staffini, “A CNN–BiLSTM Architecture for Macroeconomic Time Series Forecasting,” in *ITISE 2023*, MDPI, June 2023, p. 33. doi: 10.3390/engproc2023039033.
[33] J. Brownlee, “How to Develop LSTM Models for Time Series Forecasting,” Machine Learning Mastery. Accessed: Oct. 05, 2025. [Online]. Available: https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/
[34] S. Hu *et al.*, “CNN-BiLSTM-Attention Model for Wind and Solar Power Forecasting and Analysis of Factors Affecting Forecast Accuracy,” in *2025 6th International Conference on Electrical, Electronic Information and Communication Engineering, EEICE 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 347–350. doi: 10.1109/EEICE65049.2025.11033699.
[35] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, “Dive into Deep Learning,” Dive into Deep Learning. Accessed: Oct. 04, 2025. [Online]. Available: https://d2l.ai/index.html
[36] J. Brownlee, “How to Develop Convolutional Neural Network Models for Time Series Forecasting,” Machine Learning Mastery. Accessed: Oct. 05, 2025. [Online]. Available: https://machinelearningmastery.com/how-to-develop-convolutional-neural-network-models-for-time-series-forecasting/
[37] C. Wilson, “On the Scale of Global Demographic Convergence 1950–2000,” *Popul. Dev. Rev.*, vol. 27, no. 1, pp. 155–171, Mar. 2001, doi: 10.1111/j.1728-4457.2001.00155.x.