International Journal of Development Mathematics Vol 3 Issue 1 Page No. 156 - 171

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

A Stacked ARIMA-GRU meta-model for mortality modelling: An Ensemble Learning Approach

Hauwa'u M. Inuwa[a], Aliyu U. Shelleng[a,*], and Aliyu U. Kinafa[a]

[a]*Department of Mathematical Sciences, Gombe State University, Gombe State, Nigeria*

ARTICLE INFO

Article history:
Received 20 November 2025
Received in revised form 20 March 2026
Accepted 25 March 2026

Keywords:
Mortality forecasting, time series analysis, skewness, machine learning, age-specific mortality

MSC 2020 Subject classification:
62P20, 91B55

ABSTRACT

Accurate mortality forecasting is essential for effective public health planning and demographic analysis, yet it is challenged by nonlinear and age-dependent patterns in mortality data. This study proposes a stacked ARIMA-GRU ensemble model for age-specific mortality forecasting in Nigeria. The model combines the linear modelling strength of the Autoregressive Integrated Moving Average (ARIMA) model and the nonlinear learning capability of the Gated Recurrent Unit (GRU) network, with Extreme Gradient Boosting (XGBoost) used as a meta-learner. Annual age-specific mortality data for Nigeria covering the period 1950-2023 were obtained from the United Nations World Population Prospects database. Model performance was evaluated using out-of-sample forecasts across all age groups based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The proposed stacked ensemble model achieved the lowest average errors (MAE =0.000929, RMSE =0.001344, MAPE = 2.42%) outperforming ARIMA (MAE =0.002096, RMSE =0.003426, MAPE = 4.95%) and GRU (MAE =0.006837, RMSE =0.008005, MAPE = 10.76%). The results demonstrate the effectiveness of the proposed stacked ensemble learning method for improving mortality forecasting accuracy in Nigeria.

1. Introduction

Accurate modelling and forecasting of future mortality rates are crucial for numerous applications, especially in life insurance pricing, demographic statistics, and pension planning, as they enable companies and policymakers to plan more effectively for future uncertainties (Roshini *et al*, 2020). Many countries have experienced a rapid increase in life expectancy in recent decades, which has further compounded the difficulties in modelling and predicting future mortality trends (Chen and khaliq, 2022). Given the significant impact of mortality trends on the population size, structure, social security system, life insurance and pension schemes, it is therefore critical to develop reliable models capable of capturing the evolving patterns of mortality over time (Umar and Chukwudi, 2019).

Consequently, mortality rates prediction has gained more research interest in recent times, with numerous studies being proposed to enhanced mortality rate forecasting accuracy. Traditional approaches, such as the Lee-Carter model , Lee and Carter (1992) and its numerous variants Gao and Shi (2021), model mortality trends using factor models, which explain the variations of mortality rates from the perspective of ages, gender, regions, and other factors. However, it has been observed that the LC model is a simple highly structured model, where there is little or no allowance for accommodating uncertainty and abrupt pattern changes. It primarily captures linear patterns and struggles with complex, nonlinear dependencies inherent in real-world mortality data Wang *et al* (2023).

Motivated by the success of machine learning and deep learning techniques, particularly recurrent neural network

* Corresponding author. Tel.: +2348033517236
E-mail address: umssie"gmail.com (Shelleng A. U)
https://doi.org/10.62054/ijdm/0301.13

(RNN) and its variants, a few recent studies have applied RNNs in mortality rate prediction. For instance, Petnehazi and Gall (2019) applied LSTM to forecast the mortality rates of 35 countries in 111 age group, and achieved better results than the classical Lee Carter model. Nigri *et al (*2019) used LSTM to forecast the time index of the Lee-Carter model. Chen and khaliq (2022) compared the performance of three recurrent neural networks and Lee Carter model for the task of mortality rate prediction in the US. Wang *et al* (2023) proposed a modified Transformer model for forecasting the mortality rates in the major countries around the world. Using a multi-head attention mechanism and positional encoding, their proposed model was able to extract the key features effectively and thus achieves enhanced performance in mortality rate prediction. Shelleng and Dikko (2024) extended the Lee-Carter model by integrating GRU for overall mortality modelling and achieved better results.

Although the Statistical and deep learning techniques have improved on the state of the arts, they still have their own limitations. The statistical methods are effective at managing linear relationships in time series, however, they are incapable of handling nonlinear relationships Martinez *et al* (2018). The deep learning methods on the other hand can model nonlinear relationships but often require large volume of data, long training time and sometimes, miss the inherent linear patterns that the classical statistical approaches can handle more effectively Adeyeye and Nkemnole (2023).

Therefore, this research proposed stacking ensemble approach for overall mortality modelling in Nigeria. Unlike the traditional additive hybrid methods, the ARIMA and Gated Recurrent Unit (GRU) models both function as base learners. The predictions from both models are then used as input features to a meta-learner, which learns the best combination of the predictions of the base learners to improve forecasting accuracy.

2. Material and Method

2.1 Data source for the study

The data used in this study were obtained from the World Population Prospects (WPP) database, which is an official platform for population projections and estimates maintained by the United Nations Population Division. The mortality measure used throughout is the central (period) death rate $m(x, n)$, defined as the number of deaths in age group $[x, x + n)$ per person-year of exposure during a given calendar year. This measure is directly available in the WPP output and is distinct from the probability of death $q(x)$. Age-specific time series spanning 1950–2023 were extracted for 21 five-year age groups: 0, 5, 10, …, 95, 100+. Age columns with fewer than 60% non-missing annual observations were excluded before analysis. No further outlier removal was applied; however, the year-over-year change series for each age group is plotted in Figure 2.

To ensure time-respecting evaluation, the observation window is partitioned into a strictly chronological training block and a held-out test block. For each age group with T annual observations, the test-block size is:

$$n\_test = min(15, max(10, floor(T \times 0.20))) \quad (1)$$

yielding n_test = 14 years (2010–2023) and a 60-year training block (1950–2009) for the present dataset. No test-period observation enters any model, scaler, or order-selection procedure during training. Table 1 summarizes the dataset partition used in the proposed study.

Table 1: Summary of dataset Partitioning

Partition	Years	Number of Years
Training	1950–2009	60
Internal val. (GRU only)	10% of training sequences drawn randomly from the pooled training pool	6 sequences
Test (held-out)	2010–2023	14
Future forecast	2024–2033	10

2.2 Model Specification

To capture the complex temporal dynamics inherent in mortality time series, this study proposes a Stacked Ensemble Learning model (SEM) that integrates ARIMA and GRU models within a meta-learning framework. The ARIMA models are efficient in capturing linear dependencies, and GRUs are excellent at modelling nonlinear time dependencies and long-term patterns.

The ARIMA model acts as the first-level base learner, responsible for modelling and forecasting the linear components of the mortality time series, while the GRU model serves as the second-level base learner, focusing on capturing the nonlinear patterns and residual errors not modelled by ARIMA. In the proposed stacking approach, a single model is trained jointly across all age groups for each component of the ensemble. One globally-selected ARIMA lag structure, one GRU with age as input features, and one XGBoost meta-learner. This approach shares statistical strength across age groups and produces age-coherent forecasts that are all generated by the same learned mapping.

2.2.1 Base Learner 1: Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model combines three key components: (i) **Auto regression term of order (p) that** uses past values (lags) of the time series to predict future values; (ii) an **integration term of order (d) that** differences the time series to make is stationary (i.e. removing trends and seasonality); (iii) a moving-average term (q) uses past forecast errors to improve predictions.

The selected ARIMA (p,d,q) model takes the form:

$$\Delta^d y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (2)$$

where y_t represents the mortality rate at time t, Δ^d denotes the differencing operator of order d, μ denote the mean term or constant of the model, ϕ_i represent the coefficient of the i-th autoregressive term (AR), θ_j represent the coefficient of the j-th moving average term (MA), and ε_t represents random error at time t.

In the proposed framework, a single global (p, d, q) order is selected that minimizes the mean Akaike Information Criterion (AIC) across all age groups simultaneously. The search space is defined as $p \in \{0, 1, 2\}$, $d \in \{0, 1\}$, $q \in \{0, 1, 2\}$, yielding 18 candidate specifications. For each candidate order, ARIMA is independently fitted to the training slice of every age group, the AIC is recorded, and the order that minimizes the mean AIC across all age groups is retained. The selected order is subsequently used to re-fit ARIMA on each age's training series before generating test-period and future forecasts. This joint structure shares the lag specification across age groups while preserving age-specific coefficient estimates, reducing the total number of model selection decisions from 21 to 1. The global ARIMA order: (2,0,2) with a mean AIC of -731.35 was selected for the experiment. Table 1 presents the top-5 ARIMA orders ranked by mean AIC across all age groups.

Table 2: Top -5 ARIMA orders by mean AIC across all age groups

P	D	Q	Mean AIC
2	0	2	-731.35
1	1	2	-727.41
2	1	0	-726.43
2	1	2	-726.19
2	0	0	-724.00

2.2.2 Base Learner 2: Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a recurrent architecture that employs reset and update gates to selectively retain long-range temporal dependencies. A single GRU network is trained jointly across all 21 age groups in the proposed stacking ensemble. To enable the shared network to discriminate between age-group dynamics, age is appended as a second input feature at each time step. The input vector at step *t* takes the form:

$$x_t = [\ \log \tilde{m}_{x,t}\ ,\ \tilde{a}_x\ ] \quad (3)$$

where $\tilde{m}_{x,t}$ is the scaled log-mortality rate for age x at time t, and $\tilde{a}_x$ is the normalized age feature. The log transform is applied before scaling to stabilise variance across age groups whose raw rates span approximately three orders of magnitude ($0.003 \leq m(x,n) \leq 0.50$). Both scalers are fitted exclusively on the training slice to prevent data leakage. The GRU network architecture used in this study is summarized in Table 3. The 64/32 unit configuration is enlarged relative to a per-age network to accommodate the richer pooled input without underfitting. The input shape is (5, 2): a lookback window of w = 5 consecutive years, each with 2 features. The network is trained on the pooled-sequence dataset constructed from the training slices across all 21 age groups. The Adam optimizer, with learning rate $\eta = 10^{-3}$, minimizes the mean squared error (MSE) using mini-batches of 32 sequences. Training proceeds for a maximum of 200 epochs with early stopping applied if the validation loss on a 10% hold-out of the pooled *training* sequences does not improve for 20 consecutive epochs (restore_best_weights = True). This internal validation split involves only training-period sequences; no test-year observations enter this phase.

Table 3: Summary of GRU Network

Layer	Configuration	Specification
GRU Layer 1	Hidden Units	64
	Return Sequences	True
	Input Shape	(5, 2)
Dropout	Dropout Rate	0.20
GRU Layer 2	Hidden Units	32
	Return Sequences	False
Dropout	Dropout Rate	0.20
Dense Output Layer	Units	1

2.2.3 Stacked ARIMA-GRU Meta-Model

To exploit the complementary strengths of ARIMA and GRU, their predictions are integrated via a meta-learner based on the Extreme Gradient Boosting (XGBoost) algorithm. Unlike the traditional additive hybrid methods that sequentially model residuals, the proposed method trains both base learners independently and combines their outputs through a learned non-linear mapping.

A single XGBoost meta-model is trained jointly across all age groups**.** Age is included as a third meta-feature, allowing the meta-learner to learn age-specific blending weights. The final ensemble forecast is given as:

$$\hat{y}_t^{(S)} = f_m(\hat{y}_t^{(A)}, \hat{y}_t^{(G)}, \text{X}; \theta^*) \quad (4)$$

where $\hat{y}_t^{(A)}$ and $\hat{y}_t^{(G)}$ are the ARIMA and GRU predictions for age group *x* at time *t*; *_m* is the XGBoost regression function; and θ^* denotes the optimized meta-learner parameters.

The XGBoost objective function minimizes the regularized MSE:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \Omega(\theta)) \quad (5)$$

where $\Omega(\theta)$ is an L1/L2 regularization penalty on tree complexity that prevents overfitting. The meta-learner is configured using the following parameters: `n_estimators` = 200, `learning_rate` = 0.05, `max_depth` = 4, `subsample` = 0.8, `colsample_bytree` = 0.8. Shallow trees (depth 4) are appropriate for the three-feature meta-problem, while the low learning rate reduces sensitivity to noisy base predictions.

2.2. 4 Multi-Step Forecast Generation and Model Evaluation

After model training, forecasts were generated for both the experimental observation period and the future projection horizon (2024–2033). Model performance is assessed using out-of-sample predictions on tests for all age groups. For the test period evaluation, the ARIMA model can generate h-step forecasts directly using the model forecast (step = h) procedure. In contrast, the GRU model generates forecast autoregressively, in which each predicted value is used as input for the next step. The final stacked prediction is obtained by using the trained XGBoost meta-model to combine the ARIMA and GRU outputs at each prediction step. The forecast accuracy is evaluated using three quantitative metrics. These are mean absolute error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

These metrics capture both the absolute and relative errors of the predictions, where lower values indicate better prediction accuracy. For the future projection period (2024–2033), each baseline model is re-estimated using the observed data set (1950–2023). ARIMA directly generates 10-level forecasts, and The GRU model uses the last five observations (2019–2023) as seed windows to generate iterative forecasts. The trained XGBoost meta-model then combines the output of the base model to produce a final unit mortality estimate.

2.2.5 Implementation Details

The proposed ARIMA–GRU ensemble model was implemented in Python 3.9 using several scientific computing libraries. Data pre-processing and numerical manipulations were performed using Pandas and NumPy. The ARIMA component was implemented using Statsmodels, the GRU network was built using TensorFlow, and the stacking meta-model was implemented using XGBoost library. Model testing was conducted using Scikit-learn, with MAE, RMSE, and MAPE as the performance metrics. Visualization was performed using Matplotlib and Seaborn. The experiments were performed on a 64-bit Windows 10 system with an Intel Core i3 processor and 8 GB of RAM.

3. Result and Discussion

3.1 Descriptive Statistical Analysis of Mortality by Age

A descriptive statistical analysis was conducted on mortality rate data across 21 age groups (0, 5, 10… 95, 100+ years) to characterize the underlying distributional properties and age-specific patterns of the dataset. The summary statistics for some selected age groups are presented in Table 4, while the visual representations are shown in Figure 1. As shown in Table 4, the descriptive statistics clearly reveal age-dependent patterns in the central tendency, dispersion, and the distributional properties of mortality across ages. Mortality is highest at birth (mean = 0.1315) and declines sharply by age 5 (mean = 0.0097), after which it remains low and stable as children grow older (ages 10 -35). The mortality rate then rises progressively from 0.0119 at age 40 to 0.6205 at age 100. The dispersion indicators (Standard deviation and interquartile range) also revealed a similar trend, showing minimal change during childhood and early adulthood and wider changes at older ages. The skewness results show that most age groups show positive skewness, indicating that the mortality distributions are right-skewed, with values concentrated at the lower end and a prolonged tail extending toward higher mortality. The kurtosis results show that most age groups have a negative kurtosis value, implying that the mortality distributions are platykurtic with tails that are flatter than those of a normal distribution. The descriptive statistical analysis results demonstrate that mortality is lowest and most uniform in adolescent and early adulthood, but substantially increases and becomes more heterogeneous with increasing age, consistent with established demographic evidence.

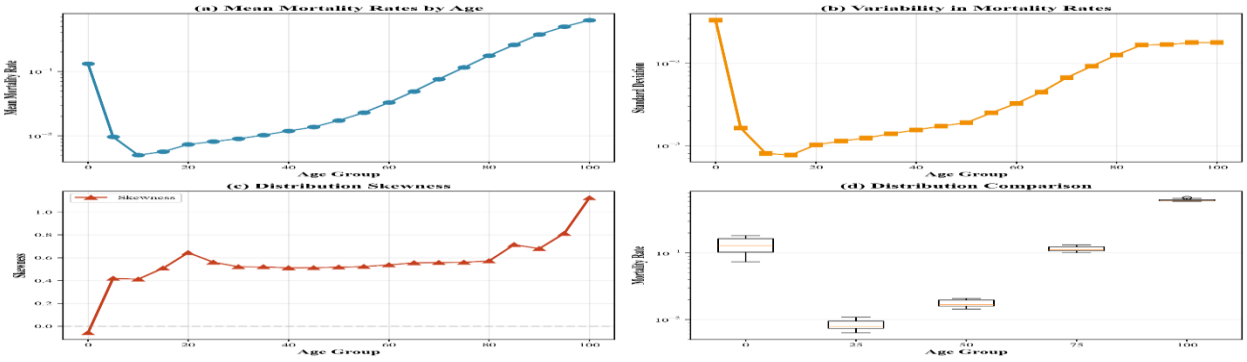

Figure 1: Descriptive Statistical Analysis of Nigeria's Mortality by Age

Table 4: Descriptive Statistics of Mortality Rate by Age (1950-2023)

Age	Mean	Std_Dev	Min	Q25	Median	Q75	Max	Skewness	Kurtosis
0	0.131498	0.033308	0.07333	0.103135	0.129515	0.16386	0.18253	-0.05732	-1.1857
5	0.009652	0.001638	0.00691	0.008565	0.00926	0.011118	0.01256	0.419472	-1.03921
10	0.005029	0.000805	0.0037	0.004488	0.004835	0.00578	0.00644	0.411907	-1.06132
15	0.005689	0.000767	0.00449	0.00515	0.00547	0.006595	0.00718	0.508911	-1.09383
20	0.00738	0.001022	0.00587	0.00668	0.007065	0.00855	0.00993	0.644565	-0.7927
25	0.008156	0.001137	0.00631	0.00738	0.00783	0.00946	0.01086	0.56107	-0.85726
30	0.009034	0.001237	0.0071	0.008173	0.008675	0.010475	0.01167	0.52131	-1.01847
35	0.010266	0.001398	0.00811	0.00929	0.009855	0.01191	0.01301	0.520311	-1.0911
40	0.011923	0.001553	0.00949	0.010853	0.01146	0.013775	0.01484	0.511388	-1.117
45	0.013722	0.001726	0.01101	0.012495	0.0132	0.015813	0.0169	0.512489	-1.12668
50	0.017284	0.0019	0.0143	0.015913	0.0167	0.019588	0.02072	0.517887	-1.12603
55	0.022928	0.002498	0.019	0.021123	0.02215	0.025863	0.02727	0.522718	-1.11803
60	0.032885	0.003256	0.0278	0.03053	0.03188	0.036423	0.0386	0.537319	-1.10283
65	0.048988	0.00447	0.04209	0.045765	0.04761	0.05365	0.05692	0.555307	-1.09062
70	0.075852	0.006677	0.06569	0.071033	0.073825	0.08257	0.08777	0.557573	-1.08744
75	0.115105	0.009202	0.10139	0.1081	0.11238	0.124125	0.13171	0.560253	-1.08144
80	0.175491	0.012636	0.15703	0.165685	0.171805	0.187925	0.19826	0.572213	-1.08594
85	0.257637	0.016693	0.23463	0.244587	0.252285	0.271275	0.29086	0.714517	-0.96448

90	0.373599	0.016873	0.35011	0.35886	0.3687	0.388785	0.40957	0.679643	-0.88141
95	0.49328	0.017836	0.47211	0.478125	0.487035	0.508153	0.53524	0.813207	-0.61734
100	0.620511	0.017854	0.59956	0.607775	0.613685	0.63249	0.67231	1.123764	0.355558

3.2. **Temporal Trend Patterns across Age Groups**

The temporal evolution of age-specific mortality rates in Nigeria between 1950 and 2020 is visualized in Figure 2**.** Figure 2 (a) depicts the mortality trends across different age groups (0 – 100 years), showing a consistent decline in mortality across all ages with steeper declines among younger and middle-aged groups (0 - 60) and slower decline at advanced ages (80 – 100 years). This suggests that health improvements have disproportionately benefited lower age groups, whereas mortality among the elderly remains relatively high.

Figure 2(b), depicts the trend direction and magnitude across all ages. The uniformly negative slopes confirm a general downward trend in mortality for all age groups. The slope magnitude of the trend becomes more negative with increasing age, indicating that while mortality is decreasing for all age groups, the rate of decline is more pronounced in older adults. Figure 2(c) illustrates the overall percentage change in mortality rates from the first to the last observation year. The analysis reveals that younger age groups experienced the largest proportional reductions, exceeding 60% over the study period, while the change for the older age group is smaller but still significant. This pattern aligns with global demographic transitions, wherein infant and middle-age mortality declines tend to outpace those observed at older ages.

Figure 2(d) presents the linear trend fit quality (R^2 values), which measures how well linear models capture historical mortality patterns. High R^2 values (>0.8) for younger and middle-aged groups indicate that mortality changes in these cohorts follow a relatively stable and predictable trend. However, the lower R^2 values at extreme ages (especially 80 – 100 years) suggest increased nonlinearity and variability, highlighting the need for non-linear models such as GRU or ensemble models to effectively capture mortality behaviour at older ages.

The trend analysis suggests a consistent negative slope across all age cohorts, indicating an overall decrease in mortality rates over time. The high R^2 values (R^2>0.8) indicate that the linear model explains a great portion of the changes in mortality rates. However, despite the declining trend, the Augmented Dickey-Fuller (ADF) test reveals that none of the age-specific series are stationary (ADF P-values > 0.05). This implies the presence of a non-stationary behaviour across age groups, which can be addressed by differencing or some form of transformation. The Augmented Dickey-Fuller (ADF) test reveals that none of the age-specific series are stationary (ADF P-values > 0.05). This implies the presence of a non-stationary behaviour across age groups, which can be addressed by differencing or some form of transformation before modelling.

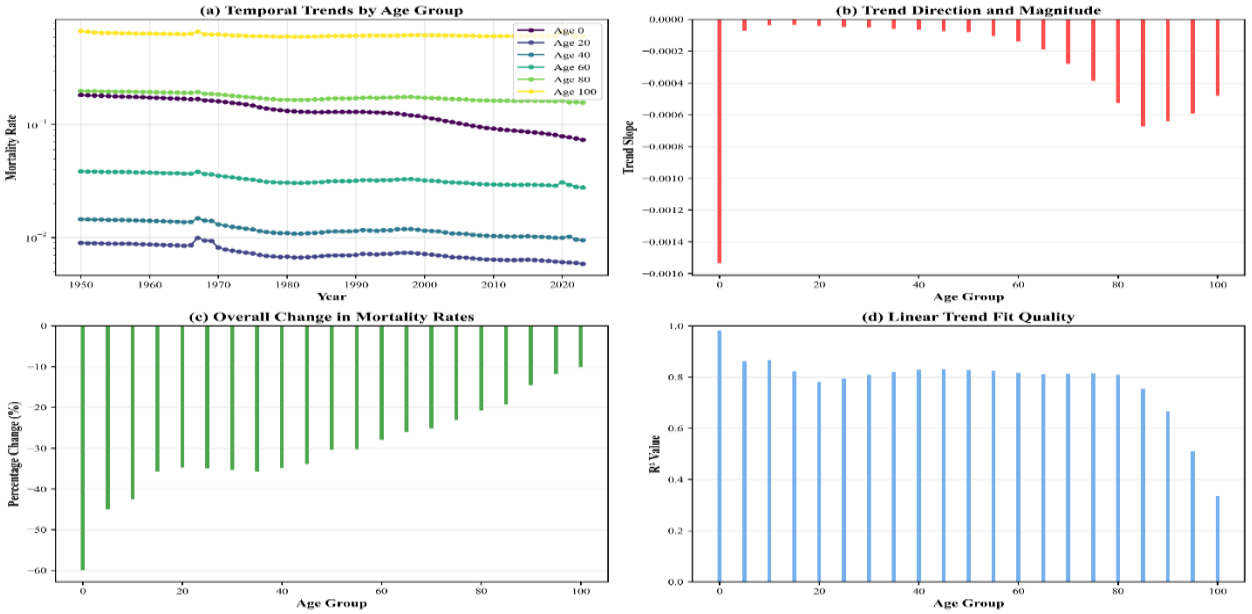

Figure 2: Temporal evolution of age-specific mortality rates in Nigeria between 1950 and 2020

3.3 Model Evaluation Results

The prediction performance of the proposed Stacked Ensemble Model (SEM) and that of its constituent base learners was evaluated using three standard performance evaluation metrics: MAE, RMSE and MAPE on the out-of-sample test dataset. To fully explore the generalization capability of the SEM and its base learners, the models were evaluated across all age groups (0 – 100 years) at 5-year intervals. Table 5 presents the comparative results for representative age groups, while Table 6 summarizes the average performance across all age categories. The detailed performance of all models across all ages is provided in the Appendix.

As presented in Table 5, at age 0, the proposed SEM achieved the lowest error across all three metrics (MAE = 0.000562, RMSE =0.00075, MAPE = 0.70%), outperforming both the ARIMA (MAE =0.002561, RMSE =0.004082, MAPE = 3.32%) and GRU(MAE =0.001598, RMSE =0.001857, MAPE = 1.93%) models. At age 10, the GRU model is the best-performing model in terms of the three evaluation metrics. Conversely, at age 25, the SEM demonstrated clear dominance again, recording a significant reduction in prediction errors (MAE =0.000271, RMSE =0.000383, MAPE = 4.11%) compared to the ARIMA (MAE=0.000757, RMSE=0.000891, MAPE = 11.35%) and GRU(MAE=0.000661, RMSE=0.00076, MAPE = 9.90%).

At age 50, where mortality rates exhibit relative stability, the SEM remained the best, recording MAE =0.000295, RMSE =0.000369, MAPE = 1.97%, followed by GRU (MAE =0.002424, RMSE =0.00267, MAPE = 16.18%) and ARIMA (MAE =0.000677, RMSE =.00088, MAPE = 4.56%). Although the performance margin was narrower at this mid-age group, the SEM still showed improved robustness over the GRU and ARIMA models. The performance analysis at older ages, such as 75 and 100, where mortality is greater, shows that the SEM still maintained superior performance. The SEM yielded the lowest MAPE values (1.26% and 0.26%, respectively) when compared to ARIMA and GRU models.

The average performance across all ages presented in Table 3 further demonstrates the superiority of the SEM across all age categories. On average, the SEM achieved the lowest MAE (0.000929), RMSE (0.001344), and MAPE (2.42%),), outperforming both ARIMA MAE (0.002096), RMSE (0.003426), and MAPE (4.94%), and GRU MAE (0.006837), RMSE (0.008005), and MAPE (10.76%).

Overall, the superior performance of the proposed SEM across all ages demonstrates its enhanced capacity to capture both linear trends and nonlinear temporal dependencies inherent in mortality data. The SEM's consistent accuracy reflects the advantage of integrating ARIMA and GRU through meta-learning with XGBoost, which effectively optimizes the contribution of each base learner. The framework enables balanced learning, improved generalization,

and lower prediction error across all age categories.

Table 5: Performance Metrics Comparison of ARIMA, GRU, and SEM by Age on Out-of-Sample Data

Age	Model	MAE	RMSE	MAPE (%)
0	ARIMA	0.002561	0.004082	3.321491
	GRU	0.001598	0.001857	1.92661
	Stacked	0.000562	0.00075	0.702542
5	ARIMA	0.000378	0.000604	5.25826
	GRU	0.000197	0.00027	2.692572
	Stacked	0.000149	0.00017	1.963445
25	ARIMA	0.000757	0.000891	11.354616
	GRU	0.000661	0.00076	9.902366
	Stacked	0.000271	0.000383	4.109445
50	ARIMA	0.000677	0.00088	4.561755
	GRU	0.002424	0.00267	16.179558
	Stacked	0.000295	0.000369	1.968768
75	ARIMA	0.002311	0.003725	2.24987
	GRU	0.016651	0.018328	15.997623
	Stacked	0.001314	0.00197	1.264746
100	ARIMA	0.009578	0.01642	1.507679
	GRU	0.021679	0.029812	3.449532
	Stacked	0.001644	0.00281	0.263134

Table 6: Average Performance Metrics of ARIMA, GRU, and SEM Across All Age Groups

Model	MAE	RMSE	MAPE
ARIMA	0.002096	0.003426	4.949736
GRU	0.006837	0.008005	10.760042
SEM	0.000929	0.001344	2.420166

3.4 Forecasting Result

Figures 3 - 6 illustrate the short-term and long-term mortality rate forecasts produced by the ARIMA, GRU, and the SEM for selected representative age groups over the period 1950-2033. In each figure, the upper panel depicts the model fit and short-term forecasts relative to the historical data (1950-2023), while the lower panel illustrates the long-term forecasts for the period 2023-2033.

Across all the age categories investigated, the SEM consistently produces forecast that more closely track the observed data and yield smoother long-term predictions than either ARIMA or GRU. While ARIMA adequately models linear historical patterns, it fails to capture the nonlinear fluctuations inherent in mortality trends. On the other hand, the GRU shows stronger capability in learning temporal dependencies but exhibits mild instability and occasional drift in long-term forecast.

For instance, at age 25 (Figure 4), the short-term forecast shows that both SEM model closely tracked the actual mortality trend, whereas ARIMA and GRU exhibits relatively higher deviation. In the long-term projection, the SEM model maintain a stable downward trend, reflecting a sustained improvement in survival rates, while the GRU and ARIMA models predicts a slightly increasing (upward) trend. This observation highlights the superior ability of the SEM to capture the non-linear temporal dependencies and to produce more plausible long-range mortality forecasts.

Similarly, at age 50 (Figure 5), the three models show different predictive behaviours in the short-term forecasts. The ARIMA model produces stable and moderate forecasts, closely following the historical data trends. The GRU model shows an upward trend in mortality rates, highlighting a stronger sensitivity to recent changes in the data. The SEM model outperform both models by having more stable and balanced prediction, providing predictions that align closely with the actual test data. Over the long-term forecast period, the ARIMA model predicts a stable mortality rate, while the GRU model forecast an upward trend. In contrast, the SEM provided a more stable and conservative forecast, demonstrating its ability to mitigate extreme variations from the base models.

The short-term and long-term forecasts highlight that the proposed SEM achieves enhanced generalization and robustness in both short-term and long-term mortality forecasting. Through the integration of the strengths of ARIMA and GRU using meta-learning strategy, the SEM effectively balances bias and variance, resulting in more stable and realistic mortality projections across age groups.

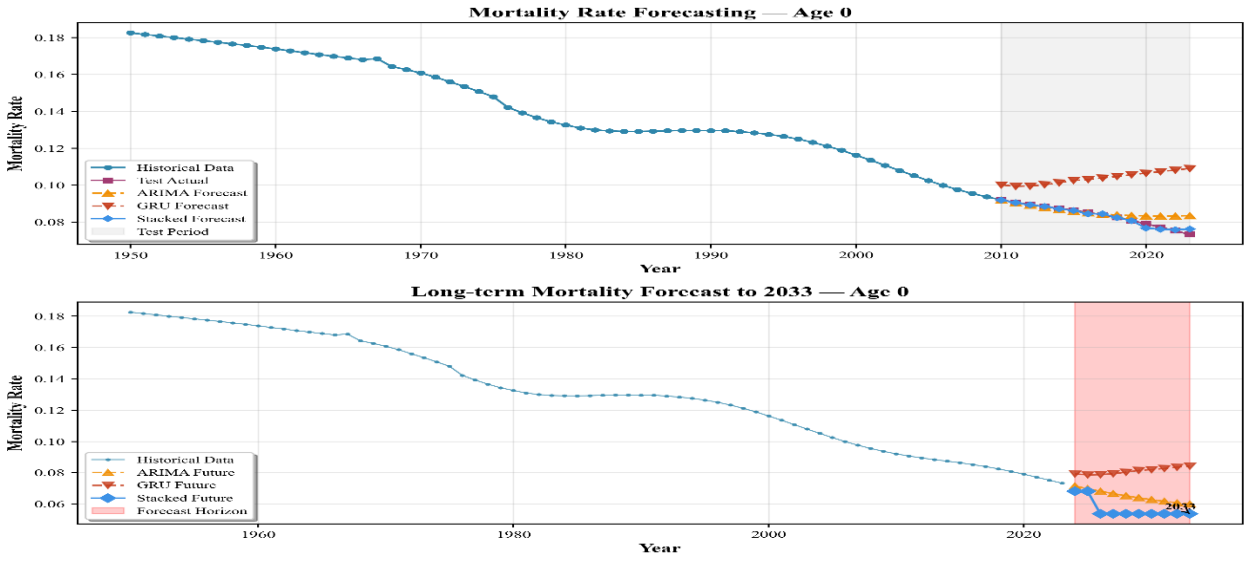

Figure 3: Comparison of ARIMA, GRU and Stacked Ensemble Models for Mortality Rate Forecasting at Age 0 (1950 - 2033)

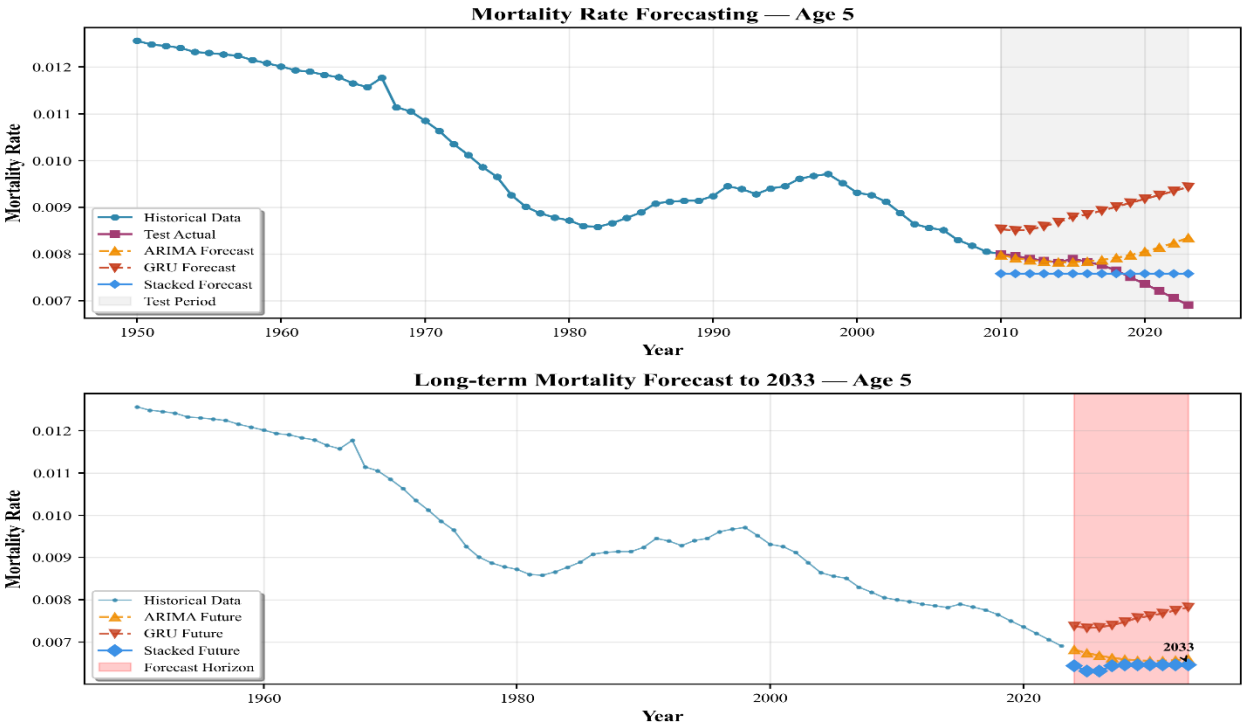

Figure 4: Comparison of ARIMA, GRU and Stacked Ensemble Models for Mortality Rate Forecasting at Age 5 (1950 - 2033)

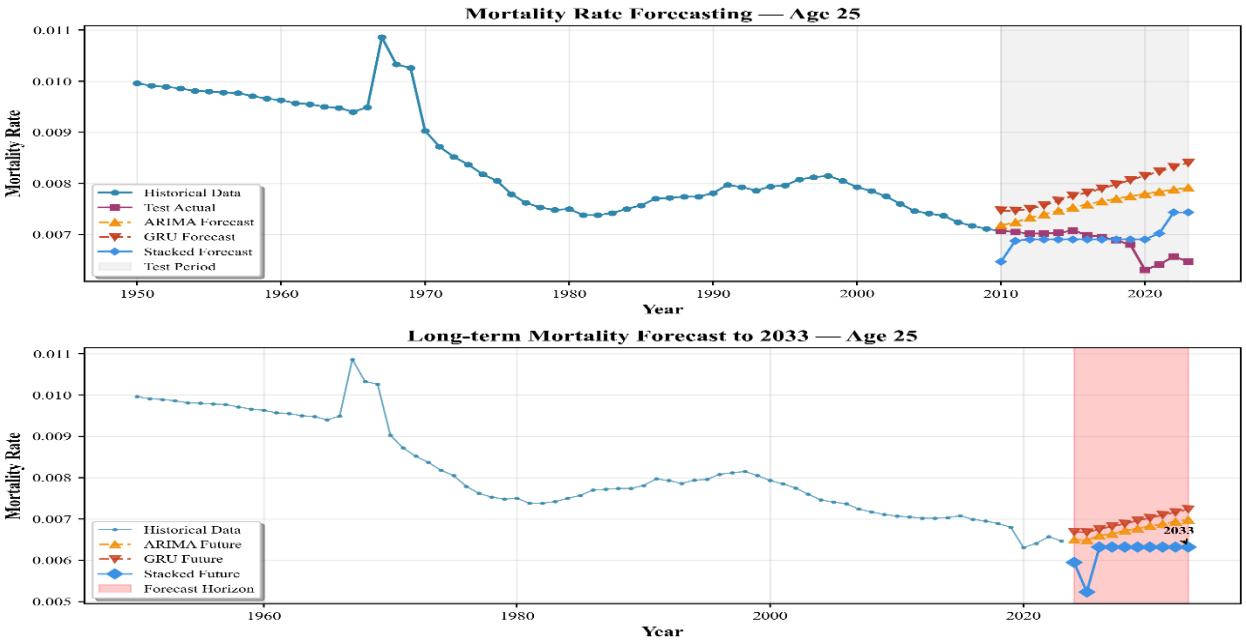

Figure 5: Comparison of ARIMA, GRU and Stacked Ensemble Models for Mortality Rate Forecasting at Age 25 (1950 - 2033)

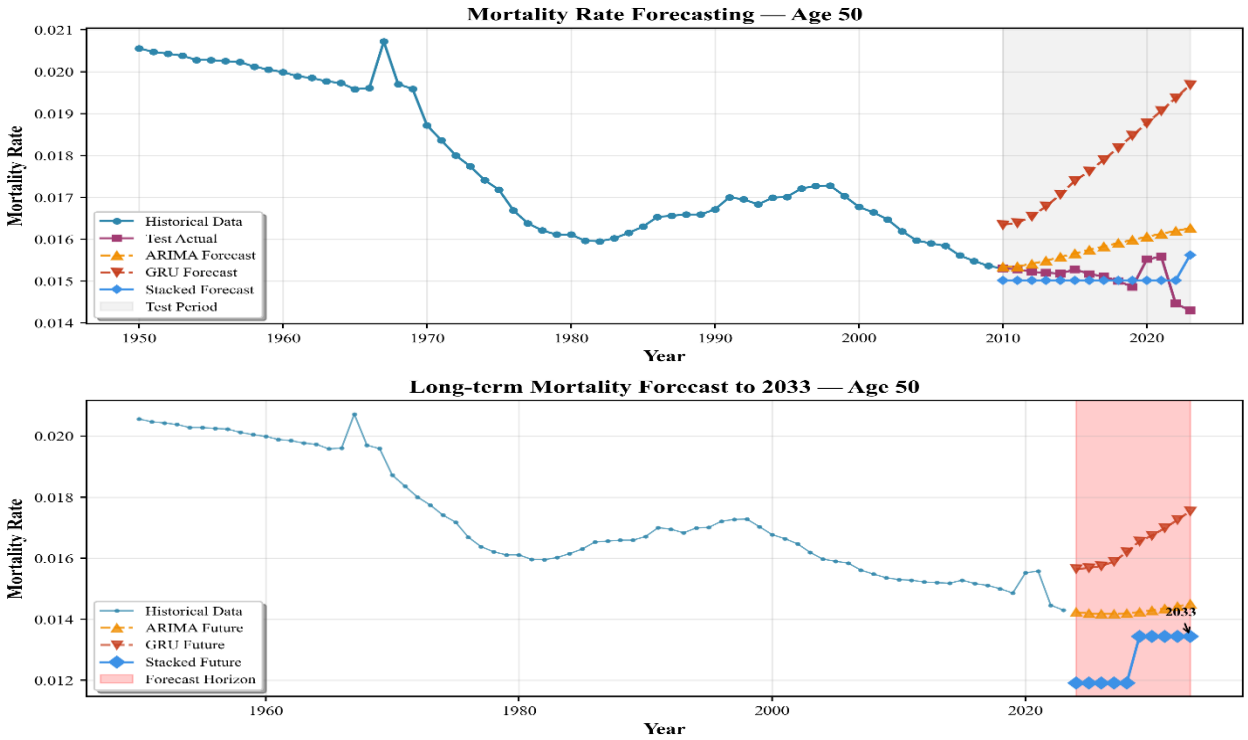

Figure 6: Comparison of ARIMA, GRU and Stacked Ensemble Models for Mortality Rate Forecasting at Age 50 (1950 - 2033).

4. Discussion

This study proposed a novel Stacked Ensemble Model (SEM) that integrates the strength of ARIMA and GRU models through XGBoost meta-learning for the task of age-specific mortality rate forecasting in Nigeria. When tested with its constituent models, the proposed ensemble model performed best in almost all evaluated age groups and forecasting horizons. The superior and consistent performance of SEM demonstrates the effectiveness of integrating linear and nonlinear models to capture both the deterministic and stochastic trends inherent in mortality data.

The findings of this study are consistent with those from previous works, such as the studies by Gyamerah et al. (2023) and Shelleng *et al.* (2023), which underscore the significance of integrating linear and non-linear patterns in mortality forecasting models. Similar to the deep learning-based method proposed by De Mori et al. (2025), this study affirms that the GRU-which is a variant of the recurrent neural network architecture- effectively captures complex temporal dynamics of time series data. However, unlike the single deep learning models, the proposed SEM reduces overfitting and improves stability in long-term forecasting through meta-learning integration. This hybrid learning approach supports the findings of Makridakis et al. (2020) that ensemble and hybrid models generally perform better than individual models in time series forecasting problems, particularly if the underlying dataset exhibits both linear and non-linear behaviours.

In addition to improvement in forecast accuracy, the proposed SEM also provided more stable long-term forecasts across all age categories when compared with those of its constituent base models. The ability of the proposed SEM to produce smoother and more realistic future mortality projections is in conformity with evidence from demographic models such as Raftery *et al.* (2014), which advocate integrating multiple modelling approaches for enhanced population forecasts. The SEM proposed in this study balanced performance across short-term and long-term horizons, underscoring its suitability in demographic planning and mortality risk assessment

The findings from this study carry significant implications from a public health perspective. Timely and accurate forecasting of mortality is key for policy formulation, particularly in guiding healthcare resource allocation, pension planning, and monitoring progress toward national and global health targets. The proposed SEM provides a more evidence-driven tool for mortality monitoring and intervention by capturing both age-specific and temporal dynamics inherent in mortality data.

5. Conclusion

The results of this research indicated that Nigeria's mortality rate forecasting is challenging due to the fact that the patterns involved are dynamic, age-dependent, and above all, exhibit non-linear behaviour over time. The traditional statistical methods, such as ARIMA, are effective at capturing linear patterns but fall short in capturing the nonlinear trends that characterize mortality behaviours at early and old ages. While the deep learning models, such as GRU, overcame some of the limitations of the traditional methods, they tend to neglect linear patterns in the mortality data and may overfit if the training data is not sufficient. To leverage the strength of both approaches, this study successfully integrates the complementary strengths of both methods through a stacked ARIMA-GRU XGBoost Meta-Model. The ARIMA is employed for capturing the linear pattern, while the GRU is utilized for capturing nonlinear trends in the mortality data. The combined feature set is then passed to the XGBoost for the final decision. Experimental results show that the proposed method yielded superior forecasting accuracy, improved robustness across all age groups, and forecasting horizons. Therefore, this research concludes that combining traditional statistical and deep learning approaches via ensemble learning techniques, particularly the stacking ensemble method, provides a promising and more reliable alternative for demographic and mortality forecasting in Nigeria.

Limitations

This study focuses on point forecasting accuracy using the proposed stacked ARIMA–GRU ensemble and its constituent base models. Although ARIMA and GRU provided reliable forecasting metrics, we acknowledge that the study lacked additional standard mortality models, such as Lee-Carter families. These models were excluded due to computational resources and limitations in the scope of this initial investigation. For future work, incorporating these additional baseline parameters will enable more comprehensive comparisons and a more comprehensive approach.

REFERENCES

Adeyeye, J. S., and Nkemnole, E.B. (2023). Predicting malaria incidence using hybrid SARIMA-LSTM model, *International Journal of Mathematical Sciences and Optimization: Theory and Applications*, 9(1), 59–80.

Chen, Y., and Khaliq, A.Q. (2022). Comparative study of mortality rate prediction using data-driven recurrent neural networks and the Lee–Carter model. *Big Data and Cognitive Computing*, 6(4), 134.

De Mori, L., Haberman, S., Millossovich, P., & Zhu, R. (2025). Mortality forecasting via multi-task neural networks. *ASTIN Bulletin: The Journal of the IAA*, *55*(2), 313-331. Gao, G., and Shi, Y. (2021). Age-coherent extensions of the Lee–Carter model. *Scandinavian Actuarial Journal*, 2021(10), 998–1016.

Gao, G., and Shi, Y. (2021). Age-coherent extensions of the Lee–Carter model. *Scandinavian Actuarial Journal*, 2021(10), 998–1016.

Gyamerah, S. A., Mensah, A. A., Asare, C., & Dzupire, N. (2023). Improving mortality forecasting using a hybrid of Lee–Carter and stacking ensemble model. *Bulletin of the National Research Centre*, *47*(1), 158.

Lee, R.D., and Carter, L.R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419), 659–671.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74.

Martínez, F., Frías, M.P., Pérez-Godoy, M.D., and Rivera, A.J. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with kNN, *Expert Systems with Applications*, 103(38–48) .

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated Lee–Carter model. *Risks*, 7 (1), 33.

Petneházi, G., and Gáll, J. (2019). Mortality rate forecasting: can recurrent neural networks beat the Lee-Carter model? *arXiv preprint arXiv:1909.05501*.

Raftery, A. E., Alkema, L., & Gerland, P. (2014). Bayesian population projections for the United Nations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *29*(1), 58.

Roshani, A., Izadi, M., and Khaledi, B.E. (2020). Transformer self-attention network for forecasting mortality rates. *Journal of the Iranian Statistical Society*, 21(1), 81–103.

Shelleng, A.U., and Dikko, H.G. (2024). Gated Recurrent Unit Integrated Lee-Carter Model for Overall Mortality Modelling, in *Proc. 2nd Int. Conf. American University of Nigeria*, 6–9, 3027-0650.

Shelleng, A.U., Dikko, H.G., Garba, J., Abdulkarim, M., and Alhaji, B.B. (2023). A gated recurrent unit-aided lee-carter model for mortality projection. *Journal of pure and Applied Mathematics* ,2 (1), 21-26.

Umar, Y.H., and Chukwudi, U.J. (2019). Modeling mortality rates using Heligman-Pollard and Lee-Carter in Nigeria. *American Journal of Theoretical and Applied Statistics*, 8(6), 221–239.

Wang, J., Wen, L., Xiao, L., and Wang, C. (2023). Time-series forecasting of mortality rates using transformer, *Scandinavian Actuarial Journal*, 2023, (2), 109–123.

APPENDIX

Table A1: Complete Performance Metrics Evaluation Results

Age	Model	MAE	RMSE	MAPE
0	ARIMA	0.002561	0.004082	3.321491
0	GRU	0.001598	0.001857	1.92661
0	Stacked	0.000562	0.00075	0.702542
5	ARIMA	0.000378	0.000604	5.25826
5	GRU	0.000197	0.00027	2.692572

5	Stacked	0.000149	0.00017	1.963445
10	ARIMA	0.000197	0.000309	5.132906
10	GRU	0.000308	0.000348	7.832063
10	Stacked	0.000428	0.000457	10.79889
15	ARIMA	0.000385	0.000461	8.204166
15	GRU	0.000319	0.000361	6.771922
15	Stacked	0.000145	0.000239	3.135445
20	ARIMA	0.000667	0.000767	10.8927
20	GRU	0.000435	0.000491	7.096264
20	Stacked	0.000141	0.000204	2.333238
25	ARIMA	0.000757	0.000891	11.35462
25	GRU	0.000661	0.00076	9.902366
25	Stacked	0.000271	0.000383	4.109445
30	ARIMA	0.000725	0.000861	9.774742
30	GRU	0.000871	0.000984	11.71258
30	Stacked	0.000223	0.000331	3.060422
35	ARIMA	0.000721	0.000855	8.515775
35	GRU	0.00113	0.001257	13.28542
35	Stacked	0.000233	0.000305	2.749027
40	ARIMA	0.000701	0.000842	7.059123
40	GRU	0.001438	0.001587	14.39638
40	Stacked	0.000312	0.000434	3.123912
45	ARIMA	0.000685	0.000851	5.948712
45	GRU	0.00182	0.002006	15.6945
45	Stacked	0.000285	0.000374	2.471897
50	ARIMA	0.000677	0.00088	4.561755
50	GRU	0.002424	0.00267	16.17956
50	Stacked	0.000295	0.000369	1.968768
55	ARIMA	0.00085	0.001156	4.319467
55	GRU	0.00349	0.00385	17.53431
55	Stacked	0.00059	0.000966	2.996585
60	ARIMA	0.001037	0.00151	3.630497
60	GRU	0.005226	0.005763	18.03061
60	Stacked	0.000796	0.001146	2.747671
65	ARIMA	0.001384	0.002044	3.208643
65	GRU	0.007863	0.008679	18.00097
65	Stacked	0.00107	0.001713	2.449018
70	ARIMA	0.001857	0.002898	2.774835

70	GRU	0.011861	0.01308	17.47008
70	Stacked	0.001	0.00145	1.468264
75	ARIMA	0.002311	0.003725	2.24987
75	GRU	0.016651	0.018328	15.99762
75	Stacked	0.001314	0.00197	1.264746
80	ARIMA	0.003256	0.005068	2.048449
80	GRU	0.02128	0.023354	13.26094
80	Stacked	0.001174	0.00153	0.731431
85	ARIMA	0.003519	0.005732	1.467303
85	GRU	0.021347	0.023402	8.892131
85	Stacked	0.001807	0.003033	0.749462
90	ARIMA	0.005032	0.008526	1.370191
90	GRU	0.015084	0.01696	4.237875
90	Stacked	0.003924	0.005309	1.091485
95	ARIMA	0.00674	0.013466	1.343283
95	GRU	0.007893	0.01228	1.596582
95	Stacked	0.003151	0.004292	0.644659
100	ARIMA	0.009578	0.01642	1.507679
100	GRU	0.021679	0.029812	3.449532
100	Stacked	0.001644	0.00281	0.263134