



AI-Based Unemployment Forecasting in Fragile Economies: Evidence from Palestine Using LSTM and Machine Learning Models

Haitham Zeidan^{1*}, Rabeh Morrar², Adel Karaa³

Type: Full Article. Received: 06 July 2026, Accepted: 17 September 2026

Accepted Manuscript, First Online

Abstract: Purpose: This study assesses whether machine-learning and deep-learning models can improve one-quarter-ahead forecasts of unemployment in Palestine, an economy exposed to recurrent shocks. **Methodology:** Official quarterly Labour Force Survey data for 2015–2024 were used to compare SARIMA, Seasonal Naïve, LightGBM, LSTM, and an equal-weight LightGBM–LSTM ensemble. Models followed a common expanding-window rolling-origin design, with strict evaluation over 2022Q1–2024Q4. Forecast errors, Diebold–Mariano tests, directional accuracy, and turning-point detection were examined. **Results:** The ensemble had the lowest descriptive errors (RMSE 3.399 percentage points, MAE 2.039, sMAPE 7.16%). LightGBM and LSTM were also competitive, but tests did not establish statistically significant differences between these models, the ensemble, and SARIMA. No model led on every directional or turning-point measure. **Conclusion:** Combining complementary forecasting methods may support short-term labour-market monitoring, but the small evaluation window and structural shocks require cautious interpretation. **Recommendations:** Statistical offices should retain transparent benchmarks, validate AI-assisted forecasts over time, report uncertainty, and reassess models after major shocks.

Keywords: Unemployment forecasting, Palestine; Labour Force Survey, LightGBM; LSTM, Ensemble forecasting

التنبؤ بالبطالة باستخدام الذكاء الاصطناعي في الاقتصادات الهشة: أدلة من فلسطين باستخدام شبكات LSTM ونماذج التعلم الآلي

هيثم زيدان^{1*}، رابع مرار²، عادل قرعة³
تاريخ التسليم: 2026/07/06، تاريخ القبول: 2026/09/17

الملخص : الهدف: تقم الدراسة إمكان تحسين التنبؤ بمعدل البطالة في فلسطين لربع سنة مقبل باستخدام نماذج التعلم الآلي والتعلم العميق في اقتصاد يتعرض لصددمات متكررة. **المنهجية:** استُخدمت بيانات مسح القوى العاملة الرسمية الفصلية للفترة 2015–2024 لمقارنة نماذج SARIMA والتنبؤ الموسمي الساذج و LightGBM و LSTM ومتوسط تنبؤات النموذجين الأخيرين. اتبعت النماذج تصميم تقييم متدرجاً بنافذة تدريب متوسعة، مع اختبار صارم للفترة من الربع الأول 2022 إلى الربع الرابع 2024 وشمل التقييم أخطاء التنبؤ واختبارات دييولدماريانو ودقة الاتجاه ورصد نقاط التحول. **النتائج:** سجل النموذج التجميعي أدنى أخطاء وصفية جذر متوسط مربع الخطأ 3.399 نقطة مئوية، ومتوسط الخطأ المطلق 2.039، و sMAPE بنسبة 7.16% وقدم LightGBM و LSTM نتائج تنافسية أيضاً، لكن الاختبارات لم تثبت فروقاً ذات دلالة إحصائية بينها وبين النموذج التجميعي و SARIMA ولم يتفوق نموذج واحد في جميع مقاييس الاتجاه ونقاط التحول. **الاستنتاجات:** قد يدعم الجمع بين مناهج تنبؤية متكاملة المتابعة قصيرة الأجل لسوق العمل، مع ضرورة الحذر في تفسير النتائج بسبب قصر فترة التقييم والصددمات البنيوية. **التوصيات:** يُوصى بالإبقاء على النماذج المرجعية الواضحة، والتحقق المستمر من التنبؤات المدعومة بالذكاء الاصطناعي، وإظهار عدم اليقين، وإعادة تقييم النماذج بعد الصدمات الكبرى. **الكلمات المفتاحية:** التنبؤ بالبطالة، فلسطين، مسح القوى العاملة، LightGBM، LSTM، التنبؤ التجميعي.

1 General Directorate of Information Systems, Palestinian Central Bureau of Statistics, Ramallah, Palestine. ORCID: 0009-0009-1803-398X

* Corresponding author: Haitham Zeidan, haitham@pcbs.gov.ps

2 Department of Economics, Faculty of Business and Communication, An-Najah National University, Nablus, Palestine

3 Department of Computer Science, Higher Institute of Management of Tunis, University of Tunis, Tunis, Tunisia.

1 الإدارة العامة لأنظمة المعلومات، الجهاز المركزي للإحصاء الفلسطيني، رام الله، فلسطين. ORCID: 0009-0009-1803-398X

* الباحث المراسل: هيثم زيدان، haitham@pcbs.gov.ps

2 قسم الاقتصاد، كلية الأعمال والاتصال، جامعة النجاح الوطنية، نابلس، فلسطين

3 قسم علم الحاسوب، المعهد العالي للتصرف بتونس، جامعة تونس، تونس

1. Introduction

Accurate forecasts of unemployment rates are essential to support the process of monitoring economic trends and designing appropriate policies, even more so in times of high uncertainty and structural fragility (Chakraborty et al., 2021). Unemployment rates do not follow stable linear processes, but rather they depend on a variety of structural factors such as territorial fragmentation, sudden shocks and events both from the economic and political sphere which can undermine the validity of the assumption upon which many traditional forecasting models rely (World Bank, 2023; Goulet Coulombe et al., 2022).

The Palestinian labour market constitutes a relevant example of these complexities. Characterized by repeated crises and major disparities between the West Bank and Gaza, it also presents important vulnerabilities related to youth and women (World Bank, 2023; PCBS, 2025), creating a very unstable context where most traditional forecasting methods will find it difficult to provide precise and timely predictions.

Traditional time-series models such as ARIMA and SARIMA remain widely used because of their transparency and modest data requirements (Box et al., 2015). However, their linear structure may be less suitable for labour markets affected by nonlinear dynamics, abrupt shocks, and structural change. Recent forecasting research has therefore increasingly explored machine-learning, deep-learning, and ensemble approaches as complementary tools for short-term labour-market forecasting.

In addition to individual model performance, recent unemployment-forecasting studies have increasingly examined hybrid and advanced machine-learning approaches to accommodate nonlinear and changing labour-market dynamics. These include combinations of statistical and neural-network models, recurrent-network hybrids, and machine-learning frameworks incorporating macroeconomic predictors (Chakraborty et al., 2021; Yurtsever, 2023; Mero et al., 2024; Annastasya et al., 2025). Building on this literature, the present study develops and

evaluates an AI-based framework for one-quarter-ahead forecasting of the national unemployment rate in Palestine using quarterly PLFS data covering 2015–2024. The study compares a SARIMA benchmark, a Seasonal Naïve benchmark, LightGBM, LSTM, and a simple ensemble that averages LightGBM and LSTM forecasts.

Recent studies further demonstrate the growing importance of data-driven approaches in understanding labour-market and broader economic behavior. Waqas et al. (2025), for example, examined determinants of self-employment intention and behavior, highlighting the importance of individual and structural factors in employment-related outcomes. More broadly, recent economic research has continued to apply advanced quantitative approaches across different economic settings (Warganegara et al., 2026). In parallel, the increasing application of machine-learning and artificial-intelligence methods to unemployment forecasting further supports the need for updated and context-sensitive forecasting frameworks, particularly in volatile and data-constrained economies.

This study provides empirical evidence from a fragile and economically vulnerable country, which extends the evidence base beyond the European and other relatively advanced-economy contexts that dominate much of the existing unemployment forecasting literature (Katris, 2020; Gogas et al., 2022). This study provides a fair comparative assessment of statistical, machine-learning, and deep-learning approaches under a common forecasting horizon and evaluation protocol. While, most of the previous relevant studies used a rolling-origin expanding-window procedure to generate one-quarter-ahead forecast using only information available at the relevant forecasting origin, this study develops an operational forecasting evaluation framework designed to preserve temporal integrity. Finally, the study evaluates a simple ensemble forecasting strategy that combines the out-of-sample forecasts of LightGBM and LSTM through arithmetic averaging. Rather than introducing additional predictors, the ensemble combines two models with complementary

forecasting behaviour: the different forecasting behaviours of LightGBM and LSTM.

2. Literature Review and Conceptual Framework

2.1 Unemployment Forecasting in Fragile and Shock-Prone Economies

Forecasting unemployment is particularly challenging in fragile and shock-prone economies, where labour-market conditions may be affected by political uncertainty, mobility restrictions, economic disruptions, and abrupt changes in labour-force participation. In such contexts, historical unemployment patterns may not evolve smoothly, and forecasting models must account for persistence, seasonality, and possible nonlinear changes in labour-market dynamics. Palestine represents a relevant case because its labour market is characterized by structural vulnerability, geographical disparities, and repeated external shocks that may affect both employment opportunities and labour-force behaviour (World Bank, 2023; PCBS, 2025).

Recent research also emphasizes that labour-market outcomes are shaped by a combination of structural, behavioral, and individual factors. For example, Waqas et al. (2025) show that employment-related behavior is influenced by multiple determinants, reinforcing the importance of considering labour-market dynamics within their broader economic and behavioral context.

Traditional forecasting approaches remain useful in this context because they provide transparent benchmarks and can be implemented with limited data. However, when labour-market dynamics are affected by structural breaks or abrupt changes, models based primarily on fixed linear relationships may not fully capture the complexity of the underlying process. This has encouraged growing interest in machine-learning and deep-learning methods as complementary forecasting tools, particularly when the objective is to improve short-term monitoring under uncertainty.

2.2 Traditional, Machine-Learning, and Deep-Learning Approaches

Time-series models such as ARIMA and SARIMA remain widely used in economic and labour-market forecasting because of their interpretability, relatively modest data requirements, and ability to represent autoregressive and seasonal patterns (Box et al., 2015; Hyndman & Athanasopoulos, 2021). These models provide an important baseline for evaluating whether more complex forecasting approaches add practical value. Nevertheless, their linear structure may be less flexible when unemployment dynamics are shaped by nonlinear relationships, abrupt shocks, or changing interactions among labour-market indicators.

Recent studies have therefore explored machine-learning techniques for unemployment forecasting. Katris (2020) compared time-series and machine-learning approaches for predicting unemployment rates, while Gogas et al. (2022) examined machine-learning forecasting for unemployment in the euro area. Tree-based models are particularly relevant because they can model nonlinear relationships and interactions among lagged indicators without imposing a predefined functional form. Such methods may be useful when recent unemployment movements, seasonal patterns, and structural labour-market characteristics interact in complex ways.

Deep-learning approaches have gained attention in unemployment forecasting because they can learn temporal dependencies from sequential data. Long Short-Term Memory networks were developed to retain relevant information across sequences and address limitations associated with conventional recurrent neural networks (Hochreiter & Schmidhuber, 1997). More recent studies have applied recurrent and hybrid deep-learning architectures to unemployment forecasting. Yurtsever (2023) used an LSTM–GRU hybrid framework to forecast unemployment rates across several countries, while Mero et al. (2024) combined LSTM networks with genetic algorithms to improve unemployment-rate prediction. However, deep-learning models may be sensitive to sample size, model

specification, and the volatility of the underlying series, especially when quarterly observations are limited.

2.3 Hybrid and Ensemble Forecasting Approaches

A further stream of research considers hybrid and ensemble forecasting methods. Rather than relying on a single modelling approach, hybrid models combine methods that capture different aspects of the forecasting problem. For example, statistical models may represent persistence and seasonality clearly, machine-learning models may capture nonlinear relationships among structured predictors, and recurrent neural networks may learn sequential temporal patterns.

Chakraborty et al. (2021) proposed a hybrid approach to unemployment-rate forecasting, while Yurtsever (2023) examined a hybrid deep-learning framework. More recent work has continued to investigate hybrid approaches, including recurrent neural networks combined with optimization techniques and machine-learning models incorporating macroeconomic indicators (Mero et al., 2024; Annastasya et al., 2025). These studies suggest that combining models may be useful when individual approaches produce different forecasting errors or respond differently to changes in the underlying series.

Recent studies published in 2025 and 2026 provide further evidence of the growing use of artificial intelligence in unemployment forecasting. Magazzino et al. (2025) applied an advanced artificial neural-network methodology to unemployment forecasting across 23 developed and technologically advanced economies and demonstrated the capacity of neural models to represent nonlinear relationships between unemployment and macroeconomic indicators. Annastasya et al. (2025) compared several machine-learning algorithms for forecasting unemployment in Indonesia and found that gradient boosting provided competitive predictive performance, while also emphasizing the importance of expanding-window validation and robustness assessment. More recently, Mohamed and Ibrahim (2026) compared Random Forest,

XGBoost, Support Vector Regression, and ARIMA in forecasting unemployment in Somalia. Their findings indicated that machine-learning methods can provide useful forecasting gains in a fragile and data-constrained economic context, although such results remain dependent on the selected data, predictors, and evaluation design. Collectively, these recent studies reinforce the value of comparing statistical and AI-based approaches while also highlighting the need for strict out-of-sample evaluation, particularly in volatile economies.

Nevertheless, an ensemble does not automatically guarantee superior performance. Its value depends on whether the component models provide complementary information and whether the combination is evaluated under a fair out-of-sample design. For this reason, ensemble performance should be assessed against both traditional benchmarks and individual advanced models using the same forecasting horizon, target variable, and information constraints.

2.4 Research Gap and Conceptual Framework

Although the literature on machine learning, deep learning, and hybrid forecasting is expanding, comparatively less attention has been given to unemployment forecasting in fragile economies using official quarterly labour-force data. Many existing studies focus on relatively stable economies, longer time series, or richer sets of macroeconomic predictors. The Palestinian case therefore provides an opportunity to examine whether advanced forecasting approaches can offer practical value under conditions of volatility, limited quarterly observations, and recurrent structural disruption.

The present study addresses this gap by comparing SARIMA, Seasonal Naïve, LightGBM, LSTM, and a simple LightGBM–LSTM ensemble within a common one-quarter-ahead forecasting setting. The conceptual framework assumes that unemployment dynamics contain several potentially relevant elements: autoregressive persistence, recurring seasonal patterns,

nonlinear relationships, and changing temporal behaviour during volatile periods. SARIMA and Seasonal Naïve provide transparent baseline forecasts; LightGBM captures nonlinear relationships in lagged and seasonal predictors; LSTM models sequential temporal patterns; and the ensemble combines the forecasts produced by LightGBM and LSTM.

Accordingly, the study evaluates whether combining these complementary approaches can improve practical short-term unemployment forecasting while preserving temporal validity and comparability across models. The methodological implementation of this framework, including data preparation, model specifications, feature selection, and rolling-origin evaluation, is presented in the next section.

3. Methodology

3.1 Overall Forecasting Framework

The overall workflow of the proposed forecasting framework is presented in Figure 1. The framework begins with the preparation and quarterly aggregation of PLFS microdata, followed by the construction and pre-test assessment of candidate predictors. It then estimates the benchmark, machine-learning, and deep-learning models under a common rolling-origin expanding-window design. Finally, LightGBM and LSTM forecasts are combined through simple averaging, and model performance and forecast uncertainty are evaluated using out-of-sample forecasts.

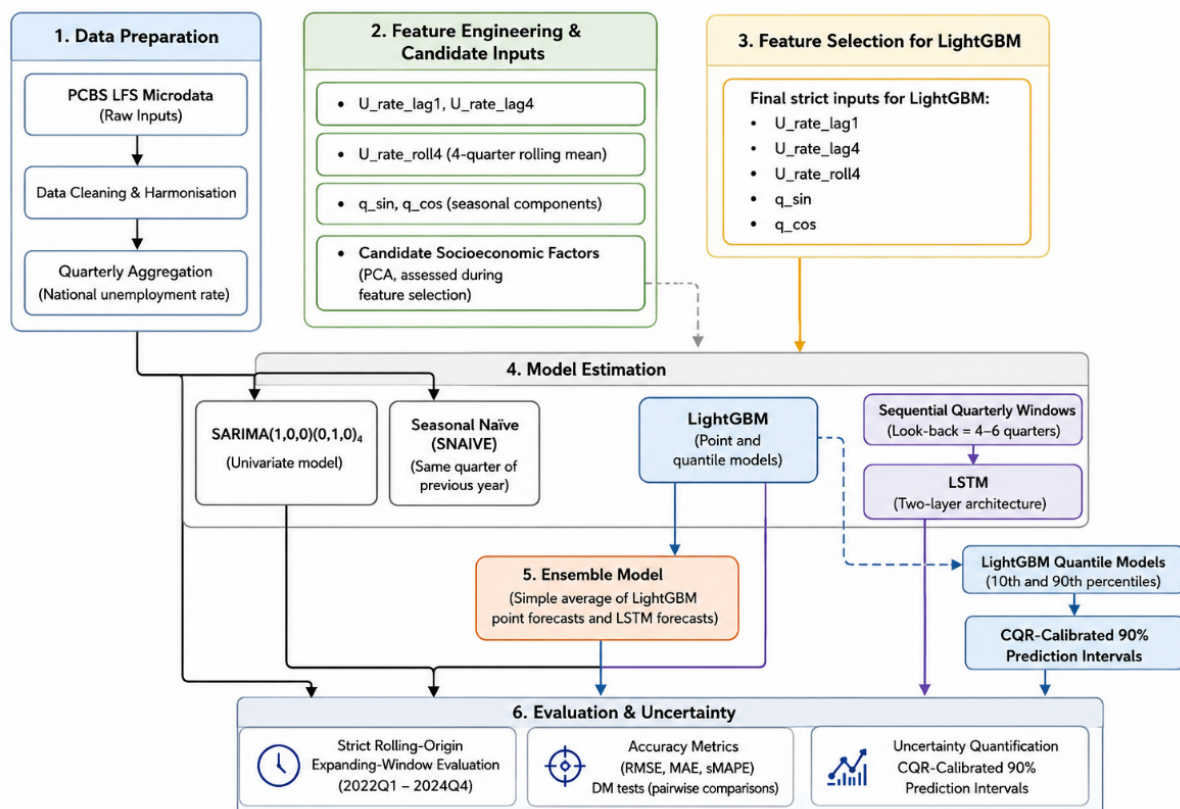


Figure 1: AI-Based Framework for Quarterly Unemployment Forecasting in Palestine

The proposed methodology integrates econometric, machine-learning, and deep-learning approaches within a unified forecasting framework. A hybrid ensemble approach is also included by averaging the forecasts generated by LightGBM and LSTM.

The framework combines rolling-origin expanding-window evaluation with forecast-accuracy measures and formal statistical comparisons to assess model performance under common operational conditions. The use

of machine learning in official statistics is discussed by UNECE (2022).

At each forecasting origin, every model was trained using only the information available up to that point. This design ensured a temporally valid out-of-sample comparison and prevented the use of future information in model estimation or forecasting (Hyndman & Athanasopoulos, 2021; Cerqueira et al., 2020).

3.2 Data and Feature Engineering

3.2.1 Data Source

This study utilizes official quarterly labour market indicators derived from the Labour Force Survey (LFS) conducted by the Palestinian Central Bureau of Statistics (PCBS). The LFS represents the primary and most authoritative source of labour market statistics in Palestine and is fully aligned with International Labour Organization (ILO) standards for measuring labour-force participation, employment, and unemployment (PCBS, 2025; ILO, n.d.).

The analysis covers the period from 2015 to 2024, providing sufficient variation to capture both stable economic conditions and periods characterized by structural shocks and volatility (World Bank, 2023).

Individual-level LFS microdata were first harmonized across survey waves and then aggregated into quarterly indicators using official sampling weights. Official sampling weights were used to account for the survey design when constructing quarterly analytical features (PCBS, 2025).

3.2.2 Target Variable

The target variable is the official quarterly headline unemployment rate in Palestine. It is calculated from the Labour Force Survey microdata using the official sampling weights as the weighted number of unemployed persons divided by the weighted labour force, multiplied by 100. All models generate one-quarter-ahead forecasts.

$$U_t = \left(\frac{Unemployed_t}{LabourForce_t} \right) * 100$$

In this regard it corresponds to international standards of the ILO, also with regards to the official methods used by PCBS (PCBS, 2025).

All forecasts were generated one quarter ahead, denoted as $t+1$. This forecasting horizon supports short-term labour-market monitoring, early-warning analysis, and timely evidence-informed decision-making. Models can be updated as new quarterly data become available.

The same target variable and forecasting horizon are consistently applied across all models, including the hybrid ensemble model, to ensure comparability of results.

3.2.3 Feature Construction

To support predictive performance while preserving interpretability and operational feasibility, candidate predictors were constructed from harmonized quarterly Labour Force Survey indicators. The use of lagged unemployment measures and seasonal Fourier terms is well established in time-series forecasting for representing persistence and recurring seasonal patterns (Hyndman & Athanasopoulos, 2021). The candidate set included four broad groups: autoregressive unemployment measures, seasonal Fourier terms, socioeconomic and demographic indicators, and lagged structural-shock indicators. At each forecasting origin, all predictors were constructed using only information available up to that point, thereby preserving temporal validity and preventing information leakage.

For LightGBM, feature selection and model development were completed exclusively within the pre-test development period. Candidate predictor sets were assessed using an inner rolling-origin expanding-window validation procedure covering 2019Q1–2021Q4, consistent with time-series evaluation approaches that preserve the temporal ordering of observations (Cerqueira et al., 2020; Hyndman & Athanasopoulos, 2021). The strict

outer test period, 2022Q1–2024Q4, was held out and was not used for feature selection, hyperparameter tuning, or model specification.

Input representations were tailored to the requirements of each forecasting architecture. SARIMA and Seasonal Naïve were estimated as univariate benchmarks using only the quarterly unemployment-rate series. LightGBM used the final selected lagged unemployment and seasonal Fourier features.

The LSTM model learned temporal patterns from sequential quarterly input windows, with the final look-back specification selected during the pre-test development stage. The ensemble model introduced no additional predictors; it combined the out-of-sample point forecasts generated by LightGBM and LSTM.

Table 1 summarizes the model-specific input structures used in the strict one-quarter-ahead forecasting evaluation.

Table 1: Model-Specific Input Structures Used in Strict Unemployment Forecasting

Model	Input structure in strict evaluation	Purpose / role
SARIMA(1,0,0)(0,1,0) ₄	Quarterly unemployment-rate series only	Univariate statistical benchmark capturing autoregressive and seasonal dynamics
Seasonal Naïve	Unemployment rate from the same quarter of the previous year	Simple seasonal benchmark
LightGBM	(U _{t-1}), (U _{t-4}), four-quarter rolling mean, (q_sin), (q_cos)	Captures nonlinear relationships in lagged unemployment dynamics and seasonal patterns
LSTM	Sequential quarterly input windows; candidate look-back lengths of 4–6 quarters were assessed during pre-test development.	Learns temporal dependencies and changing patterns over time
Ensemble	Arithmetic average of out-of-sample LightGBM and LSTM forecasts	Combines the complementary predictive behaviour of both models

Note: PCA components, socioeconomic indicators, and lagged shock indicators were assessed as candidate predictors but were not retained in the final strict LightGBM specification.

(i) Autoregressive Features

Labour-market time series often exhibit short-run persistence and seasonal dependence. Accordingly, lagged unemployment values and rolling summary measures are commonly used to represent recent conditions and recurring quarterly patterns in short-horizon forecasting (Box et al., 2015; Hyndman & Athanasopoulos, 2021). The following features were included:

- First-order lag of unemployment (U_{t-1})
- Seasonal lag (U_{t-4})
- Four-quarter rolling mean

These features summarize information available at each forecasting origin and allow the model to represent recent unemployment dynamics and seasonal recurrence without using future observations.

(ii) Seasonal Components

Seasonality is modeled using Fourier terms, allowing both linear and nonlinear models to capture smooth cyclical patterns without over-parameterization. This approach has been shown to improve forecasting performance in macroeconomic time series (Hyndman & Athanasopoulos, 2021).

$$\sin\left(\frac{2\pi q}{4}\right), \quad \cos\left(\frac{2\pi q}{4}\right)$$

Where q is the quarterly index. It has been demonstrated that using Fourier terms can lead to improved accuracy in forecasting short macroeconomic time series by allowing the models to capture recurring cycles in a compact way (Hyndman & Athanasopoulos, 2021).

(iii) Structural and Demographic Indicators

To account for labour market heterogeneity, a set of structural indicators derived from LFS microdata is included, such as:

- Youth and female unemployment rates
- Employment-to-labour-force ratio
- Labour force participation characteristics
- Educational attainment
- Refugee population share

These variables provide additional context beyond aggregate trends and improve the ability of models to capture structural differences across population groups (Katris, 2020; Gogas et al., 2022).

Given the volatility of the Palestinian economy, lagged binary indicators representing major periods of disruption, including the COVID-19 pandemic and conflict-related shocks, were assessed as candidate predictors. All shock indicators were lagged to ensure that only information available at the relevant forecasting origin was used.

The feature engineering process used to construct the model inputs is illustrated in Figure 2.

(iv) Structural Shock Indicators

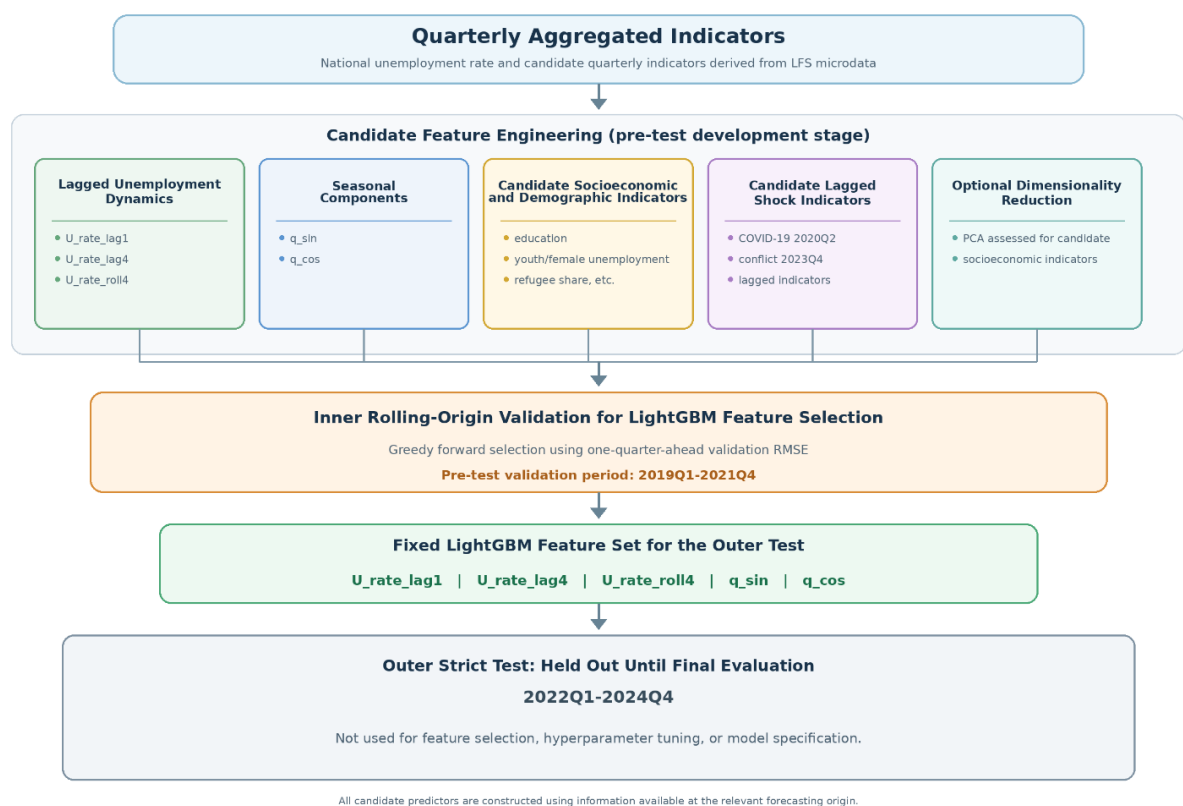


Figure 2: Feature Engineering and Inner Validation-Based Predictor Selection for LightGBM

Note: Candidate predictors were assessed through inner rolling-origin expanding-window validation during 2019Q1–2021Q4. The outer test period, 2022Q1–2024Q4, was held out entirely for final strict evaluation. Socioeconomic indicators, demographic indicators, PCA components, and lagged shock indicators were assessed as candidate predictors but were not retained in the final strict LightGBM specification.

The pipeline illustrates how raw LFS microdata were cleaned, harmonized, and aggregated into quarterly indicators for unemployment forecasting. Candidate predictors included lagged unemployment dynamics, seasonal Fourier terms, socioeconomic and demographic indicators, lagged shock indicators, and optional PCA-derived components. These predictors were assessed exclusively during the pre-test development period through an inner rolling-origin

validation procedure. Based on this process, the final strict LightGBM specification retained only lagged unemployment features and seasonal Fourier terms. The selected feature set was fixed before the outer test period and applied unchanged throughout the strict evaluation from 2022Q1 to 2024Q4.

The forecasting models used input representations tailored to their respective architectures. SARIMA and Seasonal Naïve were estimated as univariate benchmarks using the quarterly unemployment-rate series, whereas LightGBM used the selected lagged unemployment and seasonal Fourier features. LSTM was trained using sequential quarterly input windows. Candidate look-back lengths of 4, 5, and 6 quarters were assessed during the pre-test development stage, and the selected configuration was then used throughout the strict out-of-sample evaluation. The ensemble model introduced no additional predictors; it combined the out-of-sample point forecasts generated by LightGBM and LSTM. All models were evaluated under the same one-quarter-ahead rolling-origin expanding-window protocol to ensure temporal validity and comparability.

3.3 Benchmark Models: SARIMA and Seasonal Naïve

This study employs a univariate Seasonal Autoregressive Integrated Moving Average model and a Seasonal Naïve model. These benchmarks rely exclusively on the quarterly national unemployment-rate series and provide transparent reference points for evaluating the added value of the AI-based models.

The final SARIMA specification was SARIMA(1,0,0)(0,1,0)₄, where the seasonal period of four corresponds to quarterly observations. The model captures short-run autoregressive dependence together with seasonal differencing, while preserving a relatively parsimonious and interpretable structure. SARIMA models remain widely used as benchmark methods in economic and official-statistics forecasting because they provide a clear representation of temporal persistence and seasonal behaviour (Box et al., 2015).

At each rolling forecast origin, the SARIMA model was re-estimated using only the unemployment observations available up to that point. A one-quarter-ahead forecast was then generated, after which the training window was expanded to include the newly observed quarter. This procedure ensured that the SARIMA forecasts were produced under the same time-aware evaluation protocol applied to the competing models and that no future information was used.

A Seasonal Naïve model was also included as a simple baseline. For each one-quarter-ahead forecast, the model repeated the unemployment rate observed in the corresponding quarter of the previous year. Although simple, the Seasonal Naïve benchmark provides an important minimum reference for determining whether more complex statistical and AI-based methods deliver meaningful improvements in predictive accuracy (Hyndman & Athanasopoulos, 2021).

3.4 Machine Learning Model: LightGBM

Light Gradient Boosting Machine (LightGBM) was employed as the machine-learning model in the comparative forecasting framework. LightGBM is a gradient-boosted decision-tree method that can capture nonlinear relationships and interaction effects in structured data without requiring a predefined functional form (Goulet Coulombe et al., 2022; Buckmann & Joseph, 2023).

Candidate predictors included lagged unemployment measures, seasonal Fourier terms, socioeconomic and demographic indicators, lagged structural-shock indicators, and optional PCA-derived components. These variables were assessed during the feature-engineering and selection stages. However, the final strict LightGBM specification retained only five predictors: the first-order unemployment lag, the seasonal unemployment lag, the four-quarter rolling mean, and the sine and cosine seasonal Fourier terms.

Using these selected inputs, LightGBM generated one-quarter-ahead point forecasts of the national unemployment rate. Its tree-based

structure enabled the model to capture nonlinear relationships among recent unemployment dynamics and seasonal patterns while maintaining a parsimonious specification suitable for a short quarterly time series.

Hyperparameter settings were chosen conservatively to reduce the risk of overfitting and support stable out-of-sample performance. In addition to the point-forecast model, LightGBM quantile models were estimated at the 10th and 90th conditional percentiles to support the construction of CQR-calibrated 90% prediction intervals.

3.5 Deep Learning Model: Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks were used to model nonlinear temporal dependencies in quarterly unemployment dynamics. LSTM is a recurrent neural-network architecture designed to retain relevant information over extended sequences through gated memory cells, thereby helping to address the vanishing-gradient problem in sequential learning (Hochreiter & Schmidhuber, 1997).

The LSTM hyperparameters were selected during the pre-test development stage without using any observations from the strict 2022Q1–2024Q4 evaluation period. Candidate look-back lengths of 4, 5, and 6 quarters were compared using an inner rolling-origin expanding-window validation procedure covering 2019Q1–2021Q4, with validation RMSE used as the primary selection criterion. Given the limited sample size, a deliberately bounded and conservative architecture was adopted rather than conducting an extensive hyperparameter search, which could increase the risk of overfitting. The final architecture consisted of an initial LSTM layer with 32 units and `return_sequences=True`, followed by a dropout layer with a rate of 0.25, a second LSTM layer with 16 units, another dropout layer with a rate of 0.25, and a single-unit dense output layer. The Adam optimizer was used because of its stable performance in recurrent-network training, while early stopping was applied based on validation loss. Once selected, the architecture and training settings were fixed

and applied unchanged throughout the strict out-of-sample evaluation.

The model was optimized using the Adam optimizer and trained with early stopping to reduce the risk of overfitting, particularly given the limited length of the quarterly time series. At each rolling forecast origin, the model was trained only on the information available up to that point, after which a one-quarter-ahead forecast was generated. This procedure ensured temporal validity and consistency with the evaluation framework applied to all competing models.

The use of LSTM in this study requires careful justification because the quarterly dataset covers 2015–2024 and contains only 40 observations. The model was therefore included as an exploratory comparator to examine whether a sequence-based architecture could capture nonlinear temporal patterns that may not be fully represented by conventional statistical or tree-based models. It was not intended to demonstrate the general superiority of deep learning in small-sample forecasting settings. To limit the risk of overfitting, the model used short input sequences, a deliberately constrained two-layer architecture, dropout regularization, early stopping, and rolling-origin estimation using only the observations available at each forecast origin. The model specification was fixed before the strict 2022Q1–2024Q4 evaluation period. Nevertheless, the limited number of training sequences remains an important constraint, and the LSTM results should be interpreted as exploratory evidence that requires validation using longer or higher-frequency time series.

Figure 3 provides a visual representation of the LSTM architecture.

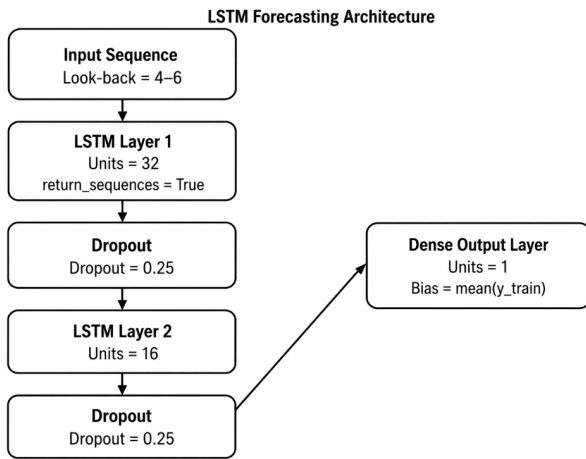


Figure 3: LSTM Architecture

3.6 A Hybrid Approach Using an Ensemble Forecasting Model

To be able to better represent uncertainty of prediction in relation to model reliability in terms of structural stability, an ensemble-based forecasting method was developed. The ensemble model will combine the forecasts from the LightGBM and LSTM models. This allows for the combination of their respective advantages of representing non-linearities and temporal dependencies.

For notation purposes we can define the one-step ahead forecast from LightGBM (i.e., $U^t + 1_{LGBM}$), and the one-step ahead forecast from LSTM (i.e., $U^t + 1_{LSTM}$). Therefore, the one-step ahead forecast from the ensemble model (i.e., $U^t + 1_{ens}$) would be given by:

$$\hat{U}_{\{t+1\}}^{\{ens\}} = \frac{1}{2} \left(\hat{U}_{\{t+1\}}^{\{LGBM\}} + \hat{U}_{\{t+1\}}^{\{LSTM\}} \right)$$

Averaging of the individual model forecasts is often employed due to its ability to reduce variability in the combined model's output, and therefore, mitigate bias introduced through the use of differing structures within each of the component models. Through this method of integration, the ensemble model seeks to achieve a balance between feature learning through structure and pattern identification through sequence analysis.

Equal weights of 0.50 for LightGBM and 0.50 for LSTM were specified before the strict

evaluation period. This simple averaging rule was selected because it is transparent, reproducible, and does not require estimating an additional weighting parameter from the limited pre-test sample. Estimating optimized weights from a short validation period could produce unstable weights and increase the risk of overfitting. Equal weighting also allows both component models to contribute symmetrically despite their different learning structures. The ensemble should therefore be interpreted as a simple benchmark combination rather than an optimized forecast-combination model.

It is important to note that the ensemble model did not use any additional predictors. Rather, it was constructed by averaging the out-of-sample point forecasts generated by LightGBM and LSTM at each rolling forecast origin. Although LightGBM and LSTM used input representations tailored to their respective architectures, both forecasts were produced for the same target variable, one-quarter-ahead horizon, rolling-origin expanding-window protocol, and information-availability constraints. The ensemble therefore preserved temporal consistency and did not incorporate future information or additional explanatory variables beyond those used by its component models.

As such, the ensemble model is tested and compared to SARIMA, LightGBM and LSTM methods, under equivalent experimental conditions; which allows for a valid assessment of whether hybrid forecasting methodologies offer improvements over other methods in terms of forecast accuracy and robustness in highly dynamic/volatile/and/or shock-prone environments.

3.7 Rolling-Origin Evaluation and Temporal Integrity

All models were evaluated using a one-quarter-ahead rolling-origin expanding-window design. At each forecasting origin t , each model was re-estimated using only the information available through quarter t , and a forecast was generated for quarter $t+1$. The newly observed value was then added to the training set before the next forecasting origin was evaluated. This procedure was repeated across the strict out-of-

sample evaluation period, from 2022Q1 to 2024Q4.

Feature selection and model specifications were finalized during the pre-test development stage. The strict outer test period was not used for feature selection, hyperparameter tuning, or model specification. This design preserved temporal ordering, prevented the use of future

information, and ensured that all models were compared under the same operational forecasting conditions (Cerqueira et al., 2020; Hyndman & Athanasopoulos, 2021).

Figure 4 illustrates the expanding training window and the iterative one-quarter-ahead forecasting process. Table 2 summarizes the corresponding evaluation settings.

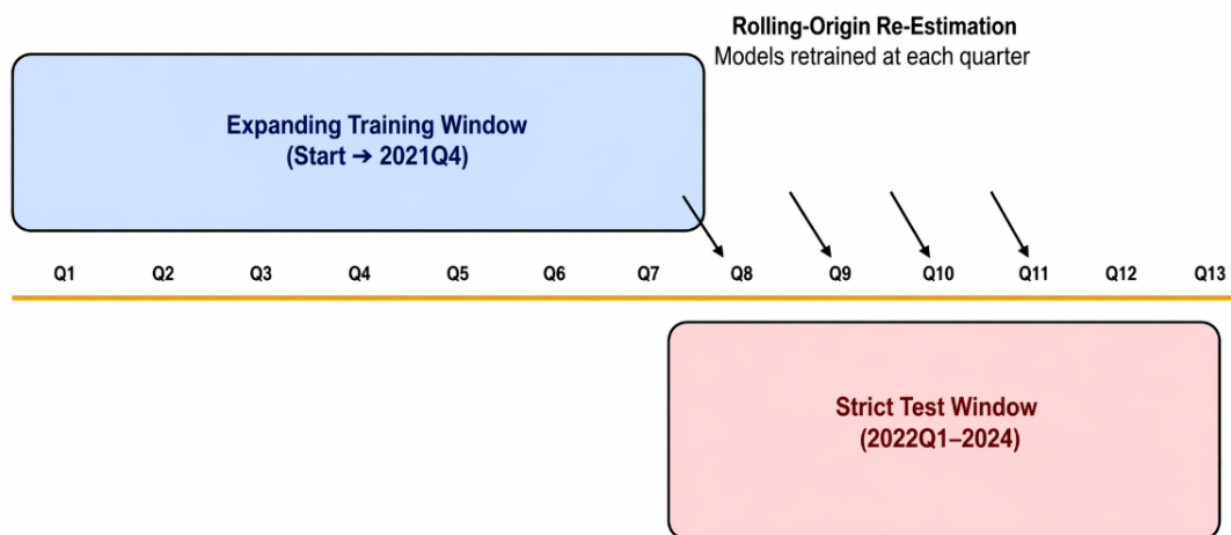


Figure 4: Rolling-Origin Expanding-Window Evaluation

Table 2: Summary of Rolling-Origin Evaluation Parameters

Parameter	Specification
Training period	From first available quarter (2015) to 2021Q4
Strict testing period	2022Q1–2024Q4
Forecast horizon	1 quarter ahead ($h = 1$)
Evaluation framework	Expanding-window, rolling-origin
Minimum training length	12 quarters
Re-estimation frequency	At every forecast origin
Total number of strict forecasts	12
Leakage prevention	Models trained only on historical data available at each origin
Applicable models	SARIMA, SNAIVE, LGBM (mean & quantile), LSTM, Ensemble
Performance metrics	RMSE, MAE, sMAPE, out-of-sample R^2 , and Diebold–Mariano tests, directional accuracy, and turning-point precision, recall, and F1 score

In addition to point forecasts, forecast uncertainty was quantified using Conformalized Quantile Regression (CQR). Two LightGBM quantile models were fitted to estimate the 10th and 90th conditional percentiles of the unemployment rate. At each rolling forecast origin, the initial quantile bounds were calibrated using conformity scores computed from the tail of the corresponding training window to obtain 90% prediction intervals. The resulting intervals were adjusted where necessary to prevent quantile crossing and are presented to illustrate forecast uncertainty during the strict out-of-sample period.

3.8 Evaluating the Quality of Forecasts

Forecast accuracy was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and symmetric Mean Absolute

Percentage Error (sMAPE). RMSE gives greater weight to large forecast errors, MAE provides an easily interpretable measure of average absolute deviation, and sMAPE offers a scale-independent percentage-based measure of accuracy (Hyndman & Athanasopoulos, 2021).

The RMSE is formulated as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (U_t - \hat{U}_t)^2}$$

The MAE is formulated as follows:

$$MAE = \frac{1}{N} \sum_{t=1}^N |U_t - \hat{U}_t|$$

where U_t is the observed level of unemployment at time t ; $\hat{U}_t$ denotes the forecasted unemployment level at time t ; and N is the number of forecasts made. These measures allow an objective evaluation of each model's forecasting capabilities. The RMSE is a metric which emphasizes differences among large errors because of the square of differences, but the RMSE is much more susceptible to outliers than MAE. On the other hand, MAE is a stable and more understandable metric to evaluate average error levels when there is considerable volatility.

The out-of-sample coefficient of determination (R^2) was also calculated as a supplementary measure of predictive fit. It compares the squared forecast errors with the variation of the observed unemployment rate around its mean during the strict evaluation period. The out-of-sample (R^2) is formulated as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^N (U_t - \hat{U}_t)^2}{\sum_{t=1}^N (U_t - \bar{U})^2}$$

where (U_t) is the observed unemployment rate at time t ; $\hat{U}_t$ is the corresponding forecast, $\bar{U}$ is the mean observed unemployment rate during the strict evaluation period, and N is the number of forecasts. Unlike the conventional in-sample coefficient of determination, the out-of-sample

(R^2) may be negative when a model performs worse than the test-period mean benchmark. It was therefore interpreted alongside RMSE, MAE, and sMAPE rather than as a standalone measure of forecasting quality.

In addition to conventional forecast-error measures, directional accuracy was calculated to assess whether each model correctly predicted the quarter-to-quarter direction of change in unemployment. For each forecast quarter, the realized direction was determined by whether unemployment increased, decreased, or remained unchanged relative to the previous quarter. The forecast direction was defined analogously by comparing the one-quarter-ahead forecast with the unemployment rate observed in the preceding quarter. Directional accuracy was calculated as the percentage of out-of-sample forecasts for which the predicted and realized directions coincided.

Turning-point performance was evaluated using a directional-reversal criterion. A realized turning point was identified when the direction of unemployment between the preceding quarter and the forecast quarter reversed relative to the movement observed over the previous two quarters. A model was considered to predict a turning point when its one-quarter-ahead forecast implied such a reversal using only information available at the forecast origin. Turning-point performance was summarized using precision, recall, and the F1 score, allowing missed reversals to be distinguished from false turning-point signals.

As an additional assessment of model behaviour under structural disruption, forecast performance was also examined separately for the conflict-related period from 2023Q4 to 2024Q4 using RMSE, MAE, sMAPE, and directional accuracy. Because this period contains only five quarterly forecasts, these results are interpreted descriptively. The directional, turning-point, and shock-period measures are treated as complementary diagnostics and do not replace the primary forecast-accuracy measures or the Diebold-Mariano comparisons.

3.9 Statistical Comparison of Forecasting Models

The Diebold-Mariano (DM) Test for Comparing Forecasts of Different Models Statistical comparison of different forecasting methods can be carried out by applying the Diebold-Mariano (DM) test (Diebold & Mariano, 1995). The DM test is a statistical method to evaluate the null hypothesis which assumes that two alternative forecasting methods will have equal predictive power in terms of their loss function during the prediction period.

In order to compare forecast error measures the DM test uses commonly used measures of error such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). Therefore, the DM test is able to determine whether there are statistically meaningful differences in model performance.

4. Results

This section presents the empirical findings for the one quarter ahead unemployment predictions produced from each of the four

models (SARIMA, LightGBM, LSTM and the ensemble model). Our analysis uses a new evaluation method called "rolling-origin expanding window". We measure the quality of our forecasts, i.e. how accurately they predict unemployment, using standard measures of forecasting quality, specifically root mean squared error (RMSE) and mean absolute error (MAE).

Overall, the results indicate that model rankings based on average error measures should be interpreted cautiously, particularly given the short and volatile evaluation period. The ensemble, LightGBM, and LSTM displayed different descriptive forecasting behaviours, but the observed differences should not be interpreted as conclusive evidence of statistically significant superiority among the advanced models.

4.1 Comparative Forecast Accuracy Across Models

The average values for one quarter ahead forecasted accuracy are reported in Table 3 using a rolling origin expanding window method.

Table 3: One-Quarter-Ahead Forecast Accuracy (Out-of-Sample, Expanding-Window Evaluation)

Model	RMSE (pp)	MAE (pp)	sMAPE (%)	Out-of-sample R ²
Ensemble (LGBM + LSTM)	3.399	2.039	7.162	0.197
LightGBM	3.503	2.103	7.363	0.148
LSTM	3.514	2.412	8.494	0.142
SARIMA	4.281	2.901	9.869	-0.273
Seasonal Naïve	5.108	3.927	13.809	-0.813

The ensemble recorded the lowest average forecast errors across RMSE, MAE, and sMAPE, followed closely by LightGBM and LSTM. LightGBM produced a slightly lower RMSE and MAE than LSTM, although the differences between the two models were small. These results describe the average out-of-sample error rankings during the strict evaluation period but do not, by themselves, establish statistically significant superiority.

The Diebold–Mariano tests did not identify statistically significant differences between the ensemble, LightGBM, LSTM, and SARIMA within the available evaluation window. By

contrast, LightGBM, LSTM, and the ensemble showed statistically significant improvements over the Seasonal Naïve benchmark. Therefore, the ensemble's lower average errors should be interpreted as descriptive evidence of competitive forecasting performance rather than conclusive evidence of superiority over SARIMA or the individual advanced models.

4.2 Model Performance During Periods of Elevated Volatility

Model behaviour was examined during periods of elevated labour-market volatility to assess how reliably the forecasting approaches

responded to rapid changes in unemployment dynamics. In such periods, point-forecast accuracy alone may not provide a complete assessment of model usefulness; the associated uncertainty around the forecast is also important for interpretation and decision-making.

In addition to point forecasts, forecast uncertainty was illustrated using CQR-calibrated 90% prediction intervals generated from the LightGBM quantile models. As illustrated in Figure 5, the prediction intervals widen during periods of substantial variation in the unemployment rate, indicating greater uncertainty in the expected short-term labour-market trajectory. This result highlights the value of reporting interval forecasts alongside point estimates, particularly in fragile and shock-prone economic environments.

To complement the interval-based assessment, forecast performance was examined separately during the conflict-related period from 2023Q4 to 2024Q4. Across these five quarters, LSTM recorded the lowest RMSE at 5.103 percentage points, while the ensemble recorded the lowest MAE at 3.687 percentage points and the lowest sMAPE at 12.441%. LightGBM, LSTM, and the ensemble each achieved directional accuracy of 60.0% during this period, compared with 40.0% for SARIMA and the Seasonal Naïve benchmark.

The shock-period results therefore indicate differences in model behaviour across evaluation criteria rather than clear superiority of a single forecasting approach. Given the small number of shock-period forecasts, these findings are interpreted as supplementary descriptive evidence and do not alter the statistical comparisons based on the full evaluation window.

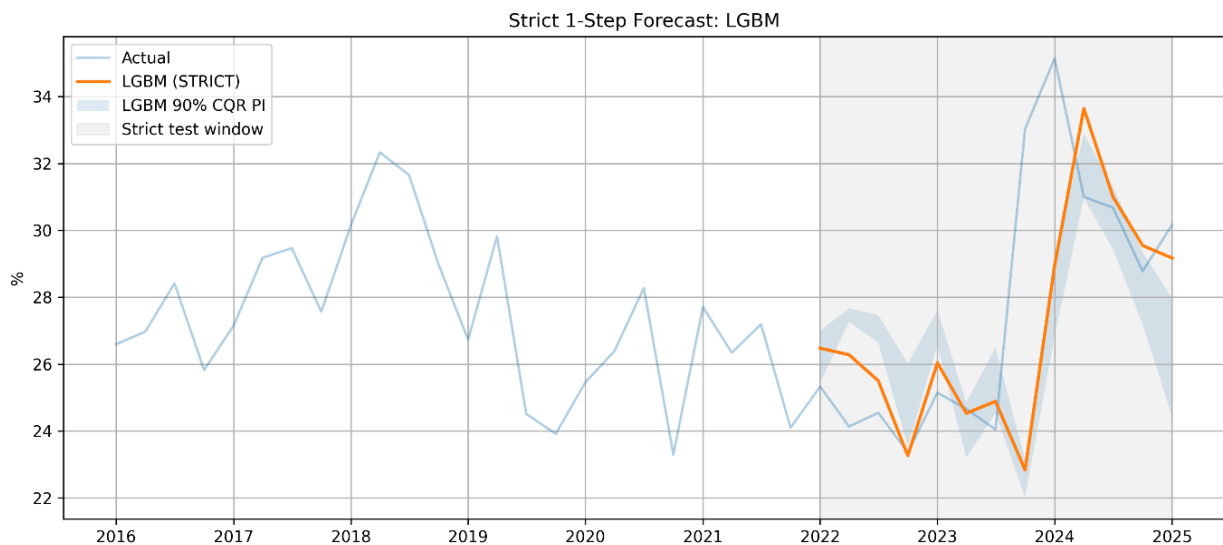


Figure 5: LightGBM Forecasts and CQR-Calibrated 90% Prediction Intervals During the Strict Evaluation Period

4.3 Statistical Analysis of Forecast Performance Differences

Diebold-Mariano (DM) testing was performed on each model pair as a means to determine if

differences in their respective forecasting accuracies were statistically significant. The DM tests were constructed using the squared error of forecasts and the p-values from the DM tests for each model pair are provided in Table 4.

Table 4: Summary of Diebold-Mariano Tests Conducted on Each Model Pair's Forecasts

Model_A	Model_B	DM	p_value
LGBM	SARIMA	-0.96218	0.35495
LGBM	SNAIVE	-2.45409	0.030365
LGBM	LSTM	-0.04312	0.966311
LGBM	ENS_LGBM_LSTM	0.72701	0.481163
SARIMA	SNAIVE	-0.90773	0.381889
SARIMA	LSTM	0.932541	0.369441
SARIMA	ENS_LGBM_LSTM	1.088538	0.297737
SNAIVE	LSTM	2.425903	0.031971
SNAIVE	ENS_LGBM_LSTM	2.570477	0.024524
LSTM	ENS_LGBM_LSTM	0.790901	0.444352

In conclusion, the DM tests indicate that the hierarchy of performance observed in the descriptive metrics is only partially supported by statistical evidence. LightGBM, LSTM, and the ensemble model significantly outperform the Seasonal Naïve benchmark, while differences among LightGBM, LSTM, the ensemble, and SARIMA are not statistically significant within the available evaluation window.

Although the ensemble model achieves the lowest average forecast errors, the DM test results indicate that these improvements are not statistically significant when compared to other advanced models. This may be attributed to the relatively small sample size and the high volatility of the underlying time series, which limit the statistical power of the test.

The Diebold-Mariano tests evaluate differences in average forecast loss but do not directly assess whether models correctly anticipate the direction of unemployment movements or identify reversals in the series. Directional accuracy and turning-point measures were therefore examined as complementary diagnostics, as reported in Section 4.4. These additional measures do not change the interpretation of the Diebold-Mariano results.

Overall, these findings suggest that hybrid approaches may be useful for short-term unemployment forecasting, while emphasizing the need for cautious interpretation when performance differences are small relative to the uncertainty of the evaluation setting.

4.4 Forecast Trajectories and Turning Point Detection

When evaluating an ensemble of models that produce forecasts from different approaches, the trajectory of each model's predictions offers additional insight into performance over time. Each of the five models included in this analysis was evaluated against a single set of actual unemployment rates and one-quarter ahead forecasts for those rates as presented in Figure 6.

Figure 6 provides descriptive context for the forecast trajectories, while formal responsiveness was additionally assessed using directional accuracy and turning-point detection. During the 2022Q1-2024Q4 evaluation window, SARIMA, LSTM, and the ensemble each correctly predicted the direction of quarterly unemployment changes in 75.0% of forecasts, compared with 66.7% for LightGBM and the Seasonal Naïve benchmark.

Under the directional-reversal definition, eight realized turning points were identified during the evaluation period. SARIMA detected seven of these reversals and achieved the highest recall (0.875) and F1 score (0.824). LSTM and the ensemble each detected six of the eight realized turning points and achieved the highest precision (0.857), with recall of 0.750 and F1 scores of 0.800. LightGBM and the Seasonal Naïve benchmark each recorded precision, recall, and F1 values of 0.750.

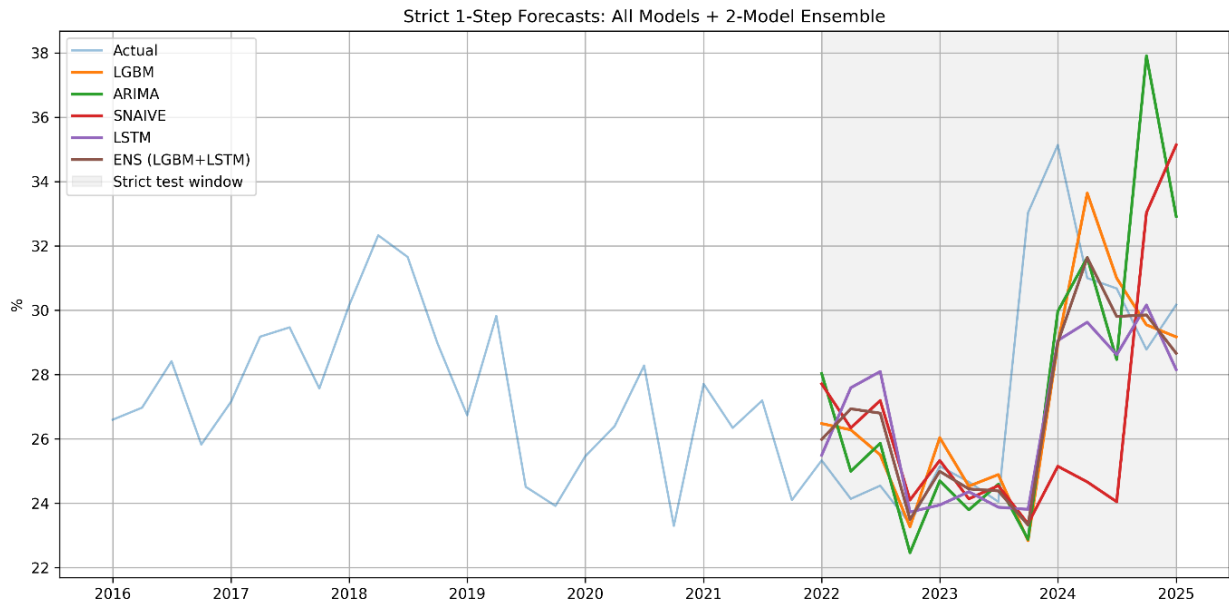


Figure 6: Actual Unemployment Rate and One Quarter Ahead Unemployment Forecasts for All Data Used in Evaluation Window

These results refine, rather than overturn, the descriptive patterns visible in Figure 6. No model dominated all dimensions of turning-point performance. SARIMA identified a larger proportion of realized reversals, whereas LSTM and the ensemble produced fewer false

turning-point signals. The results therefore provide additional evidence of complementary forecasting behaviour across model classes and should be interpreted alongside, rather than in place of, the conventional forecast-error measures and Diebold-Mariano tests.

Table 5: Directional Accuracy and Turning-Point Detection During the Strict Evaluation Period (2022Q1-2024Q4)

Model	Directional Accuracy (%)	Precision	Recall	F1
SARIMA	75.0	0.778	0.875	0.824
Seasonal Naïve	66.7	0.750	0.750	0.750
LightGBM	66.7	0.750	0.750	0.750
LSTM	75.0	0.857	0.750	0.800
Ensemble (LGBM + LSTM)	75.0	0.857	0.750	0.800

Note: A turning point is defined here as a reversal in the direction of quarter-to-quarter unemployment change. Eight such reversals occurred during the evaluation window. Precision measures the proportion of predicted

turning points that were realized, recall measures the proportion of realized turning points detected, and F1 summarizes the balance between precision and recall.

Table 6: Forecast Performance During the Conflict-Related Period, 2023Q4-2024Q4

Model	RMSE (pp)	MAE (pp)	sMAPE (%)	Directional Accuracy (%)
SARIMA	6.615	5.465	17.833	40.0
Seasonal Naïve	7.692	7.380	25.656	40.0
LightGBM	5.479	4.026	13.543	60.0
LSTM	5.103	4.026	13.522	60.0
Ensemble (LGBM + LSTM)	5.184	3.687	12.441	60.0

Note: Because the conflict-related period contains only five forecast observations, these results are descriptive and are not interpreted as evidence of statistically significant model superiority.

5. Discussion

This study compared SARIMA, Seasonal Naïve, LightGBM, LSTM, and a simple LightGBM–LSTM ensemble for one-quarter-ahead unemployment forecasting in Palestine. The ensemble recorded the lowest average forecast errors during the strict evaluation period, while LightGBM also produced competitive descriptive results. However, the Diebold–Mariano tests did not identify statistically significant differences between the advanced models and SARIMA within the available evaluation window. The findings should therefore be interpreted as descriptive evidence of comparative forecasting performance rather than conclusive evidence that the AI-based or ensemble models are statistically superior to SARIMA.

The results broadly complement previous research indicating that machine-learning methods can provide competitive alternatives to conventional time-series approaches in unemployment forecasting. Katris (2020) and Gogas et al. (2022) highlighted the potential value of machine-learning models for representing nonlinear labour-market relationships. In the present study, LightGBM achieved relatively low average forecast errors using a parsimonious feature set based on lagged unemployment measures and seasonal Fourier terms. This suggests that useful nonlinear forecasting performance may be obtained without relying on a large number of external predictors, although the absence of

statistically significant differences relative to SARIMA requires cautious interpretation.

The formal directional and turning-point diagnostics refine the earlier descriptive interpretation of model responsiveness. SARIMA, LSTM, and the ensemble achieved the same directional accuracy of 75.0%, while their turning-point characteristics differed. SARIMA achieved the highest recall and F1 score, whereas LSTM and the ensemble achieved the highest precision. During the conflict-related period, LSTM recorded the lowest RMSE, while the ensemble recorded the lowest MAE and sMAPE. These results indicate that responsiveness to abrupt labour-market changes is multidimensional and depends on the criterion considered. Importantly, they do not alter the broader finding that no single model demonstrated uniform superiority across all performance dimensions.

The ensemble result is consistent with the broader hybrid-forecasting literature, which suggests that combining models with different learning mechanisms may improve average predictive performance (Chakraborty et al., 2021; Yurtsever, 2023; Mero et al., 2024; Annastasya et al., 2025). In the present study, averaging LightGBM and LSTM forecasts produced the lowest average errors, potentially reflecting complementary forecasting behaviour. At the same time, the lack of statistically significant differences between the ensemble, the individual advanced models, and SARIMA qualifies this result. The contribution is therefore not a claim of universal or statistically significant ensemble superiority, but an indication that a simple hybrid approach may provide a practical option for short-term labour-market monitoring in a volatile and data-constrained setting.

5.1 Theoretical and Practical Implications

From a theoretical perspective, the findings support the view that unemployment dynamics in fragile economies cannot necessarily be represented adequately by a single modelling structure. Statistical, machine-learning, and deep-learning models capture different aspects of the forecasting problem, including persistence, seasonality, nonlinear relationships, and sequential temporal behaviour. The descriptive performance of the ensemble suggests that forecast combination may help integrate complementary information from models with different learning mechanisms. However, the absence of statistically significant differences between the ensemble, the individual advanced models, and SARIMA indicates that model complexity alone does not guarantee superior predictive performance, particularly when the available time series is short and affected by structural shocks.

From a practical perspective, the proposed framework provides official statistical organizations with a reproducible approach for comparing conventional and AI-based unemployment forecasts under common information constraints. Practitioners may use SARIMA and Seasonal Naïve as transparent benchmarks, LightGBM as a parsimonious nonlinear forecasting option, and LSTM or ensemble forecasts as supplementary tools when sufficient validation data are available. Rather than automatically replacing established statistical methods, AI-based models can be incorporated into a model-comparison and monitoring system in which forecasts are updated quarterly, prediction intervals are reported, and model performance is periodically reassessed. Such an approach can support early labour-market monitoring, resource planning, and evidence-informed policy discussions while maintaining transparency and human oversight.

A further limitation is the small number of quarterly observations available for training the LSTM model. Although regularization, early stopping, and a constrained architecture were used, these measures cannot fully eliminate the risk of overfitting. Accordingly, the LSTM

findings should not be generalized beyond the present dataset without further validation using longer or higher-frequency series.

The use of fixed equal weights in the ensemble represents an additional limitation. Although equal weighting provides a transparent combination and avoids estimating additional parameters from a short validation sample, it does not establish that the 0.50–0.50 allocation is optimal. Future research should compare equal weighting with validation-error-based, inverse-error, and adaptive weighting schemes using longer evaluation periods.

6. Conclusion

This study evaluated the forecasting performance and potential operational value of SARIMA, Seasonal Naïve, LightGBM, LSTM, and a simple LightGBM–LSTM ensemble for one-quarter-ahead unemployment forecasting in Palestine. All models were assessed using a common rolling-origin expanding-window framework during the strict 2022Q1–2024Q4 evaluation period.

The ensemble recorded the lowest average RMSE, MAE, and sMAPE, followed closely by LightGBM and LSTM. However, the Diebold–Mariano tests did not identify statistically significant differences between the ensemble, LightGBM, LSTM, and SARIMA within the available evaluation window. By contrast, LightGBM, LSTM, and the ensemble showed statistically significant improvements over the Seasonal Naïve benchmark. The results should therefore be interpreted as descriptive evidence of competitive forecasting performance rather than proof that AI-based or ensemble models are generally or statistically superior to established statistical models.

The additional directional and turning-point diagnostics lead to a similar conclusion. Different models exhibited different strengths in anticipating changes and reversals in unemployment, but no single approach dominated across all measures. This reinforces the value of evaluating multiple forecasting criteria in volatile and shock-prone environments.

Based on these findings, official statistical organizations are encouraged to adopt a staged and evidence-based approach to AI-assisted labour-market forecasting. This approach should begin with transparent benchmark models, followed by controlled comparison with machine-learning and deep-learning alternatives under a rolling-origin evaluation design. Forecasts should be accompanied by uncertainty intervals, documented model specifications, and regular performance reviews. During structural shocks, model outputs should complement rather than replace expert judgment and newly available administrative or high-frequency indicators. Institutional implementation should also include staff training, reproducible data pipelines, model-governance procedures, and clear communication of forecast uncertainty to policymakers. Governance should also address transparency, robustness, and accountability in accordance with the OECD Recommendation on Artificial Intelligence (OECD, 2024).

Future research should examine whether the findings remain robust when longer or higher-frequency time series become available. Further extensions could incorporate additional macroeconomic indicators, evaluate directional and turning-point performance over longer evaluation windows and additional shock episodes, and compare equal-weight averaging with validation-based or adaptive ensemble weights. Such extensions would help determine whether more advanced combination strategies offer consistent improvements while maintaining transparency, interpretability, and reproducibility.

Disclosure Statement

- **Ethical approval and consent to participate:** This study used anonymized secondary microdata from the Palestinian Labour Force Survey conducted by the Palestinian Central Bureau of Statistics (PCBS). No new data were collected directly from participants, and the researchers had no access to personally identifiable information. Data were processed in accordance with PCBS

confidentiality and data-protection procedures.

- **Availability of data and materials:** The microdata are subject to PCBS access and confidentiality regulations. Access may be requested through PCBS procedures and is subject to institutional approval.
- **Author contributions:** Haitham Zeidan contributed to conceptualization, data preparation, model implementation and analysis, visualization, interpretation, and drafting. Rabeh Morrar contributed to interpretation of labour-market and policy implications and manuscript review. Adel Karaa contributed to supervision, methodological guidance, validation, and manuscript review. All authors reviewed the final manuscript.
- **Conflict of interest:** The authors declare no competing interests.
- **Funding:** The authors report no specific grant for this research.
- **Acknowledgments:** The authors acknowledge the Palestinian Central Bureau of Statistics for access to the Labour Force Survey microdata and the academic guidance of the University of Tunis and An-Najah National University.
- **Thesis derivation:** This article is derived from Haitham Zeidan's doctoral dissertation, "AI-Based Unemployment Forecasting and Nowcasting in Palestine: A Comparative Study of Statistical and Machine Learning Models," University of Tunis.

References

- Annastasya, T., Passarella, R., & Yamani, Z. (2025). Unemployment rate forecasting in Indonesia using macroeconomic indicators with a machine learning approach. *Discover Analytics*, 3, 15. <https://doi.org/10.1007/s44257-025-00044-3>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley. [Publisher's link](#).
- Buckmann, M., & Joseph, A. (2023). An interpretable machine learning workflow with an application to economic forecasting. *International Journal of Central Banking*, 19(4), 449–522. <https://www.ijcb.org/journal/v19n4/interpretable-machine-learning-workflow-application-economic-forecasting>
- Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109, 1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>
- Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment rate forecasting: A hybrid approach. *Computational Economics*, 57(1), 183–201. <https://doi.org/10.1007/s10614-020-10040-2>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
- Gogas, P., Papadimitriou, T., & Sofianos, E. (2022). Forecasting unemployment in the euro area with machine learning. *Journal of Forecasting*, 41(3), 551–566. <https://doi.org/10.1002/for.2824>
- Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5), 920–964. <https://doi.org/10.1002/jae.2910>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>
- International Labour Organization (ILO). (n.d.). Labour force statistics (LFS and RURBAN databases). ILOSTAT. <https://ilostat.ilo.org/methods/concepts-and-definitions/description-labour-force-statistics/>
- Katris, C. (2020). Prediction of unemployment rates with time series and machine learning techniques. *Computational Economics*, 55, 673–706. <https://doi.org/10.1007/s10614-019-09908-9>
- Magazzino, C., Mele, M., & Mutascu, M. (2025). An artificial neural network experiment on the prediction of the unemployment rate. *Journal of Policy Modeling*, 47(3), 471–491. <https://doi.org/10.1016/j.jpolmod.2024.10.004>

- Mero, K., Salgado, N., Meza, J., Pacheco-Delgado, J., & Ventura, S. (2024). Unemployment rate prediction using a hybrid model of recurrent neural networks and genetic algorithms. *Applied Sciences*, 14(8), 3174. <https://doi.org/10.3390/app14083174>
- Mohamed, A. M., & Ibrahim, A. D. (2026). Forecasting unemployment in Somalia using machine learning models. *Discover Sustainability*. <https://doi.org/10.1007/s43621-026-04175-9>
- Organisation for Economic Co-operation and Development. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449; revised May 3, 2024). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Palestinian Central Bureau of Statistics. (2025). Palestinian Labour Force Survey annual report—West Bank: 2024. <https://pcbs.gov.ps/media/5sultexj/book2717.pdf>
- United Nations Economic Commission for Europe (UNECE). (2022). Machine learning for official statistics. <https://unece.org/statistics/publications/machine-learning-official-statistics>
- Waqas, U., Umair, S., Al Shamsi, I. R., Kamran, H., & Bystrov, I. (2025). Self-employment intention and behavior: An investigation of determinants and the moderating role of intrinsic motivation. *Jordan Journal of Business Administration*, 21(4), 609–621. <https://doi.org/10.35516/jjba.v21i4.351>
- Warganegara, D. L., Hutagaol-Martowidjojo, Y., & Noviyanti, E. (2026). The effect of profitability on the accrual quality-cash holding relationship of Indonesian listed firms. *Jordan Journal of Business Administration*, 22(1), 82–101. <https://doi.org/10.35516/jjba.v22i1.179>
- World Bank. (2023). Racing against time: Economic monitoring report to the Ad Hoc Liaison Committee. <https://www.worldbank.org/en/country/westbankandgaza/publication/world-bank-economic-monitoring-report-to-the-ad-hoc-liaison-committee-september-2023>
- Yurtsever, M. (2023). Unemployment rate forecasting: LSTM-GRU hybrid approach. *Journal for Labour Market Research*, 57, Article 18. <https://doi.org/10.1186/s12651-023-00345-8>