Swinburne University of Technology
Sarawak Campus

COS40007
Artificial Intelligence for Engineering

Design Project Theme 3

Malaysian Labour Force Analytics and Forecasting

Abdullahi Hussein DAHIR (102778118)
Rafid AL JAWAD (102776293)
Shifaz Ahamed RIFAN DEEN (102789657)
Tadiwanashe Terrence FUNGURANI (101232778)
Yassir BENGUARHOU ANGUILET (101220809)

Table of Contents

1. Introduction

1.1. Background and Motivation

As final year students approaching graduation, our group is aware of the challenges facing jobseekers in Malaysia. Our project on forecasting Malaysia's unemployment rate using time series models directly addresses our own imminent need for reliable information about employment prospects and labor market conditions. By developing this forecasting tool, we're equipping ourselves and our fellow graduates with valuable insights to assess market adoption rates and employment opportunities in the country.

Accurate forecasting of labor force indicators is essential for policymakers, economists and social planners to anticipate employment trends, respond to crises and allocate resources effectively. Our motivation stems from recent economic shifts that have highlighted the need for proactive data-driven policy planning. Prior studies, such as by Ismail et al. (2022), have shown that ARIMA and ARFIMA models provide promising accuracy when applied to Malaysia's unemployment data. Similarly, Aziz & Foo (2024) emphasized the predictive strength of ARIMA models using DOSM data, while Kamarudin & Vizie (2024) demonstrated how artificial neural networks outperform traditional models in certain forecasting tasks.

We selected Theme 3 (open topic) to deviate from common image classification projects, gain experience with time series data and address research questions like "what will happen?" and "why did this happen?" - providing predictive insights with direct policy and personal relevance.

1.2. Project Objectives

Through this project we aim to develop and evaluate a comprehensive suite of time series forecasting models to predict Malaysia's monthly unemployment rate using publicly available labour force data from the Department of Statistics Malaysia (DOSM). As graduating students, we recognize both the technical value and practical applications of this research as we address 4 key critical research questions.

- **Research Question 1** – What are the most effective models for forecasting Malaysia's unemployment trends among traditional statistical approaches (ARIMA, SARIMA) and modern machine learning methods (LSTM neural networks)?
- **Research Question 2** – How do different forecasting methods compare in terms of forecast accuracy using established metrics such as Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE)?
- **Research Question 3** – Can we produce reliable short-term forecasts spanning 3-6 months that could effectively aid workforce planning and policy formulation?
- **Research Question 4** – What underlying patterns, trends or seasonal variations in Malaysia's labor market data might reveal insights about economic shifts and structural changes in employment dynamics?

With this, the team aims to implement advanced time series analysis, utilize neural network frameworks and lastly develop end-to-end models with practical forecasting capabilities.

1.3. Summary of Outcomes

Our comprehensive analysis yielded significant findings in Malaysian unemployment forecasting capabilities. The LSTM neural network achieved superior performance with 1.84% MAPE for general unemployment forecasting and 2.72% MAPE for seasonally adjusted data, substantially outperforming traditional approaches. SARIMA demonstrated strong performance among statistical models with 3.20% MAPE for general unemployment and 5.39% for seasonally adjusted data. Basic ARIMA models showed limitations with MAPE values exceeding 12%.

Extended youth unemployment modeling using CEEMDAN-LSTM methodology achieved exceptional results: 4.59% MAPE for 15-24 age group and 2.7% MAPE for 15-30 age group through the advanced decomposition techniques. Youth-specific SARIMA models achieved 3.43% MAPE (15-24) and 3.70% MAPE (15-30), demonstrating the effectiveness of demographic-targeted forecasting approaches.

Technical achievements include successful multi-step forecasting across 1, 3, 6 and 12-month horizons, comprehensive stationarity analysis confirming first-order differencing requirements, advanced CEEMDAN decomposition revealing intrinsic mode functions and a strong validation using temporal splitting techniques. An interactive web-based demonstrator was developed using Plotly and Dash frameworks, providing real-time forecasting capabilities and comprehensive data analysis tools.

2. Dataset

2.1. Data Source

For this project we used data from OpenDOSM (Open Department of Statistics Malaysia), Malaysia's official statistical repository and National Statistics Organization. The dataset encompasses five comprehensive CSV files containing monthly labor force statistics with differentiated temporal coverage: General Labor Force Indicators and Seasonally Adjusted Labor Force data span from 2010 to 2025 (providing 15 years of longitudinal data), while Youth Unemployment, Unemployment Duration, and Employment Status datasets cover 2016 to 2025 (providing 10 years of data).

This temporal structure enables comprehensive employment pattern analysis across multiple economic cycles, including the post-2008 financial crisis recovery period (captured in the longer series), sustained economic growth phases, COVID-19 pandemic impact and post-pandemic recovery trends. The authoritative government source ensures data credibility and reliability for forecasting applications

2.2. Data Overview

TABLE I. Dataset Overview

Dataset	Timeframe	Data points	Key Variables
General Labour Force Indicators	2010 – 2025	183 months	Labour force size, employment, unemployment rate and participation rate
Seasonally Adjusted Labour Force	2010 – 2025	183 months	Seasonally adjusted unemployment rate, labor force metrics
Youth Unemployment	2016 – 2025	111 months	Unemployment rates for ages 15-24 and 15-30
Unemployment Duration	2016 – 2025	111 months	Active job seekers by duration (3, 6, 12+ months)
Employment Status	2016 – 2025	111 months	Worker categories (employers, employees, self-employed, unpaid family)

Comprehensive data quality check and analysis showed good dataset integrity with zero missing values across all core unemployment rate series. The General Labor Force dataset (182 observations till February 2025) shows unemployment rates ranging from 2.6% to 5.3% with a mean of 3.44% and a standard deviation of 0.52%. Youth unemployment data (110 observations till February 2025) demonstrates higher volatility with 15-24 age group rates ranging from 9.0% to 14.2% (mean 11.5%, $\sigma = 1.26\%$) and 15-30 age group rates from 5.7% to 10.0% (mean 7.25%, $\sigma = 0.94\%$).

The exploratory data analysis (EDA) was conducted during preliminary stages using data available through February 2025, establishing the foundational understanding of unemployment patterns and trends. Subsequently, March 2025 data became available and was incorporated for model training and forecasting. The inclusion of this additional month did not materially alter the observed patterns, confirming the effectiveness of the initial analytical insights.

For primary unemployment forecasting, models were trained using the complete 2010-2025 General and Seasonally Adjusted datasets (183 observations till March 2025). Additional specialized models were developed for youth unemployment analysis, with separate training conducted for 15-24 and 15-30 age group unemployment rates using the 2016-2025 youth dataset (111 observations till March 2025). This approach enables both comprehensive national unemployment forecasting and targeted youth-specific predictions addressing demographic-specific labor market challenges.

2.3. Exploratory Data Analysis

Analysis of the primary unemployment series reveals distinct temporal patterns essential for forecasting model development. The general unemployment rate showed relative stability from 2010-2019 (3.0-4.0% range) with notable COVID-19 impact causing unprecedented spikes to 5.2% in 2020, followed by gradual recovery to current levels around 3.1%. Seasonal decomposition confirmed strong 12-month cyclical components with unemployment typically peaking during specific months, aligning with Malaysia's economic calendar and monsoon-influenced employment patterns. This is consolidated by Figure 1 below.

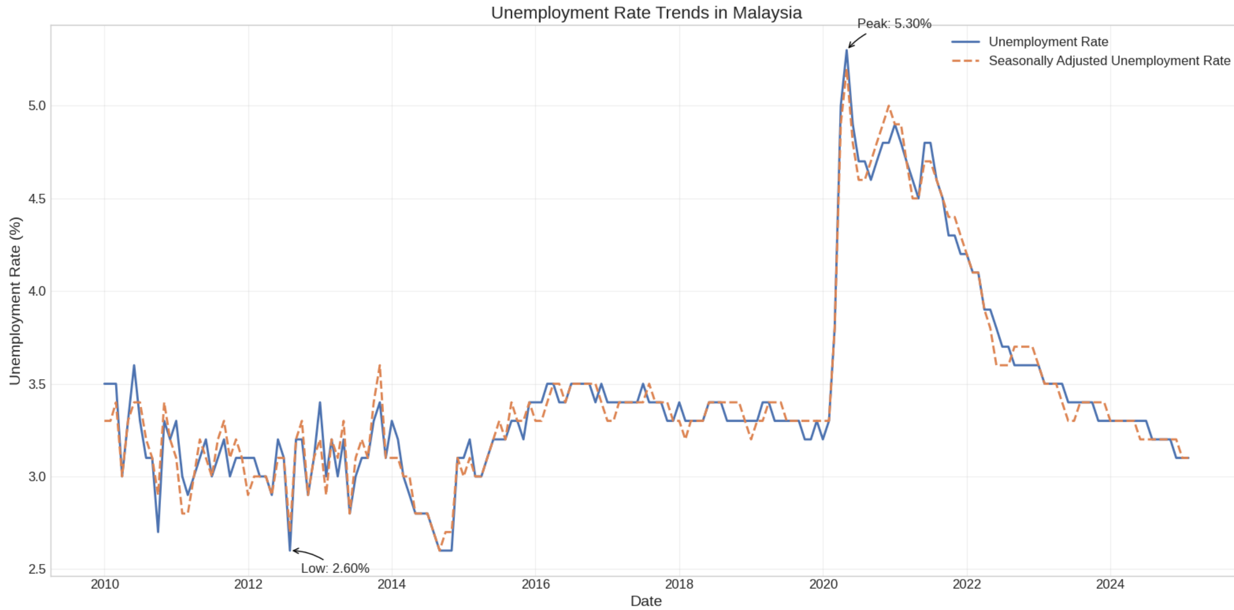

Figure 1. *Unemployment Rate Trends.*

Youth unemployment on the other hand demonstrates significantly higher volatility and elevated baseline rates compared to the general population. The 15-24 age group shows consistently higher unemployment (mean of 11.5%) with pronounced vulnerability periods, particularly during 2020-2021 when rates exceeded 14%. Analysis reveals substantial unemployment gaps: youth (15-24) rates average 8.1 percentage points above general rates, while young adults (15-30) maintain 3.8 percentage points above general rates. As shown in figure 2 below, Critical vulnerability periods were identified through rolling statistical analysis, with high-risk periods occurring during economic transitions and external shocks.

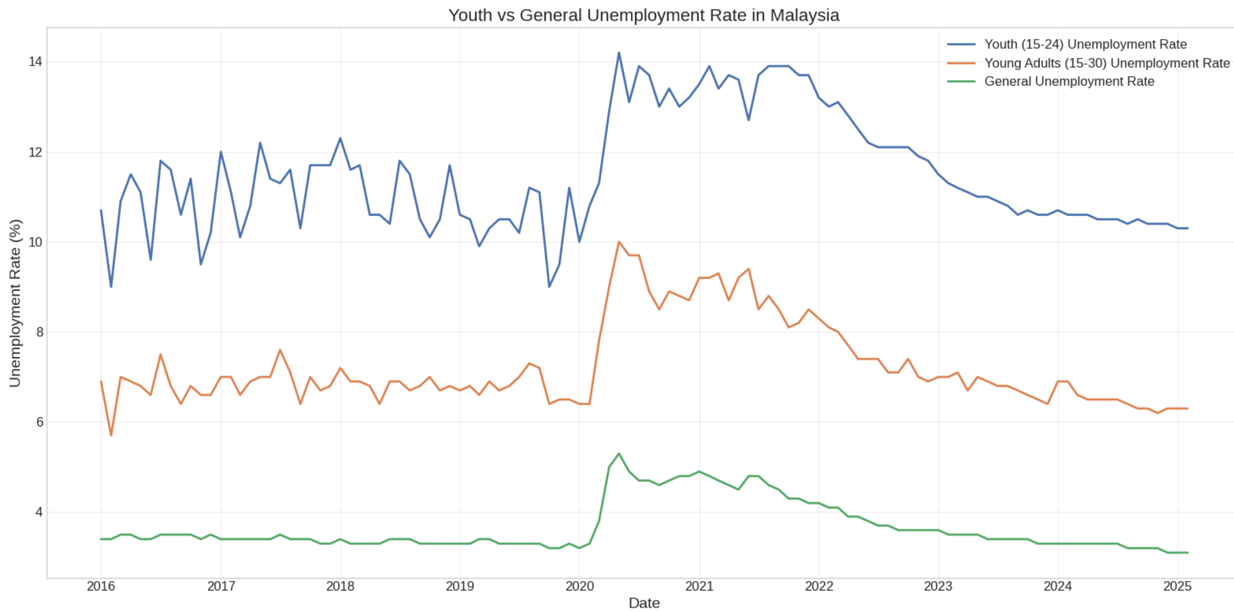

Figure 2. *Youth Unemployment Analysis*

In Figure 3 below, Labor force participation analysis reveals steady growth from 64.0% (2010) to current levels of 70.7%, indicating expanding workforce engagement. Employment-to-population ratios demonstrate a similar upward trajectory, reaching current levels around 68%. The "Unemployed Participation Gap" analysis shows dramatic structural changes during COVID-19, with unprecedented increases in both active and inactive unemployment categories.

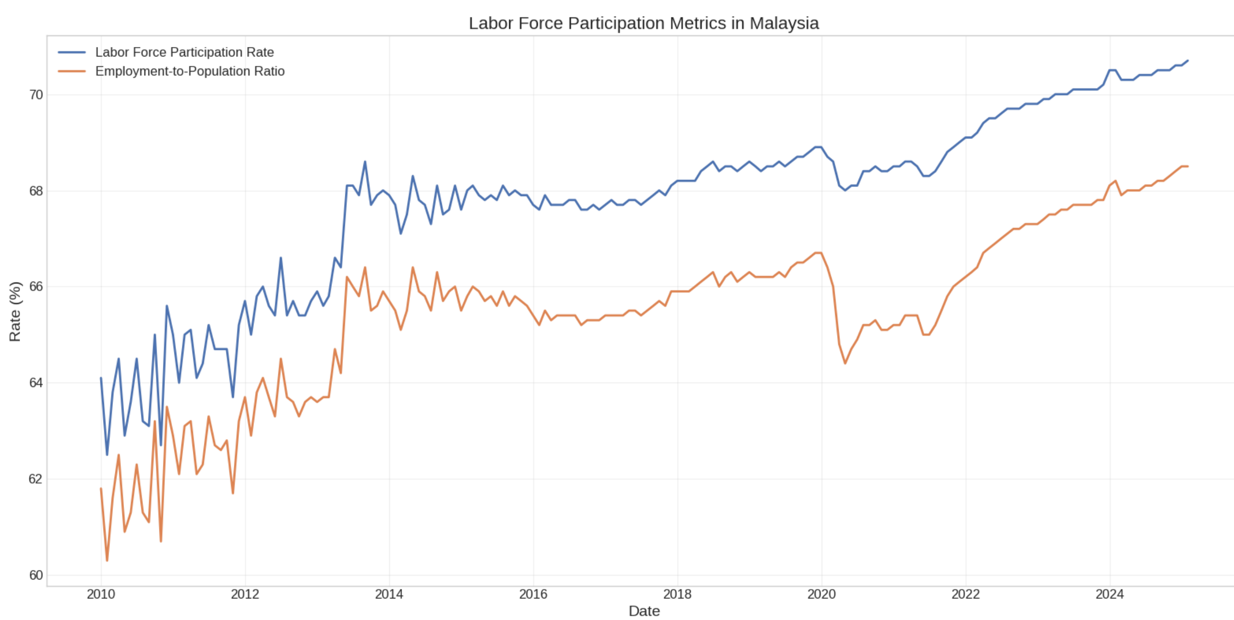

Figure 3. *Labor Force Participation Trends*

Lastly, Unemployment duration analysis reveals structural changes with increased prevalence of long-term unemployment (12+ months) during economic downturns. Active-to-inactive unemployment ratios show dramatic COVID-19 impact with ratios exceeding 5.0 during peak pandemic periods, subsequently declining to current levels around 3.8. This analysis provides critical insights into labor market flexibility and job-matching efficiency. The analysis output is visually supported by Figure 4 below.

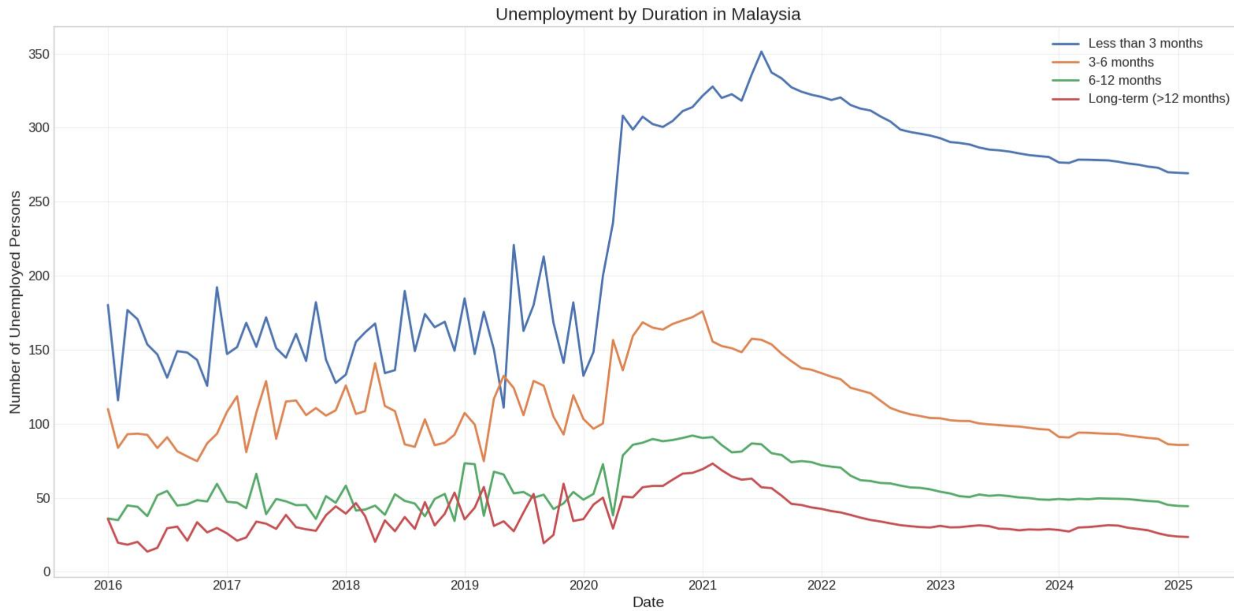

Figure 4. *Duration-Based (long-term) Unemployment Patterns*

3. AI model developments

3.1. Statistical analysis

For stationarity testing we employed dual complementary approaches using Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests (Said & Dickey, 1984). For general unemployment series, ADF testing revealed non-stationarity with test statistics of -2.1790 (p-value: 0.2139) for general unemployment and -2.0568 (p-value: 0.2622) for seasonally adjusted series. KPSS tests confirmed non-stationarity with statistics exceeding critical values (1.0271 and 1.0545 respectively, p-values: 0.01).

Youth unemployment series demonstrated similar non-stationary behavior. The 15-24 age group showed ADF statistic of -1.1637 (p-value: 0.6890) while the 15-30 group exhibited -1.8569 (p-value: 0.3526). KPSS statistics of 0.8711 and 0.8411 respectively (p-values: 0.01) confirmed unit root presence.

First-order differencing successfully achieved stationarity across all series. Post-differencing ADF statistics reached -10.0529 and -10.1660 for general unemployment series (p-values < 0.0001), while youth series achieved -9.9561 and -5.6388 respectively. KPSS statistics fell below critical thresholds which confirmed successful stationarity transformation.

On the other hand, comprehensive normality assessment utilized Jarque-Bera tests revealing significant deviation from normal distribution across variables. Labor force data showed JB statistics of 7.8700 (p-value: 0.0195) and participation rate 21.3809 (p-value < 0.0001). Youth unemployment series also demonstrated JB statistics of 6.6650 for 15-24 group (p-value: 0.0357) and 28.6028 for 15-30 group (p-value < 0.0001). These findings guided our subsequent model specification and residual analysis.

3.2. Time Series Data Differencing

Having identified that our datasets were non-stationary, differencing transformation constituted a critical preprocessing step for achieving stationarity required by traditional time series models (Box et al., 2015). First-order differencing was applied across all unemployment series based on these stationarity test results. This transformation effectively eliminated non-stationarity behavior while preserving essential unemployment dynamics.

The economic interpretation of our first-order differencing focused on month-to-month changes in unemployment rates, capturing dynamic adjustment processes that characterize labor market responses to economic changes. Post-differencing analysis confirmed successful stationarity achievement with substantially improved ADF and KPSS statistics across all series. Importantly, differencing preserved essential seasonal patterns while eliminating trend components, enabling effective SARIMA model application.

Furthermore, autocorrelation analysis revealed significant patterns essential for model specification. For original unemployment series, ACF analysis showed persistent positive autocorrelations extending across numerous lags, characteristic of non-stationary behavior. Despite our earlier assumption, Figure 5 shows that there were no seasonal patterns rather just smooth consistent decay from the first lag.

Post-differencing ACF analysis revealed improved correlation structure with rapid decay after initial lags, confirming successful stationarity achievement. Several spikes became visible approximately at 12-month intervals, supporting SARIMA specification with seasonal components. PACF analysis of differenced series showed clear cutoff patterns enabling ARIMA parameter identification.

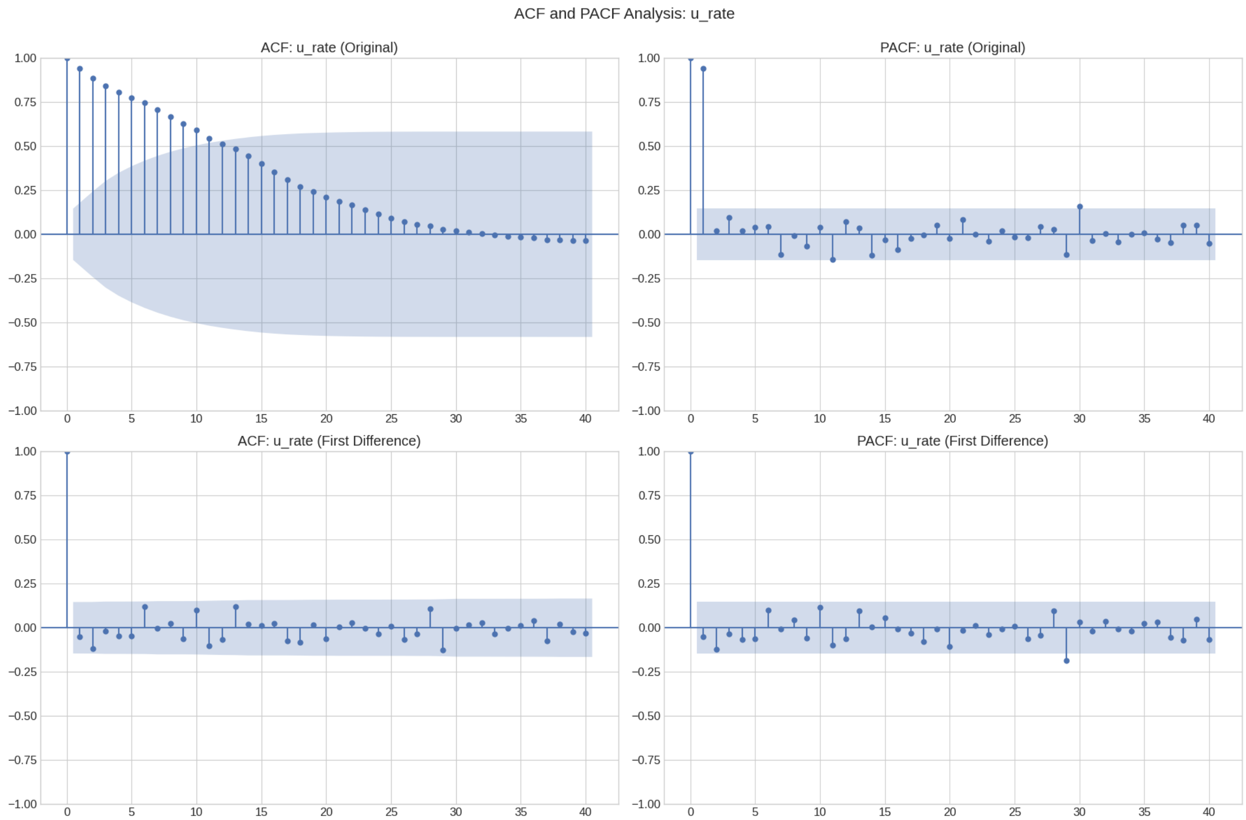

Figure 5. *ACF and PACF plots of unemployment rate*

3.3. Time Series Data Decomposition

Following successful stationarity transformation, decomposition was done to further understand the underlying structure of unemployment dynamics and inform appropriate modeling strategies. Time series decomposition enabled identification of systematic patterns including trends, seasonal variations and irregular components that require different modeling approaches (Hyndman & Athanasopoulos, 2021).

3.3.1 Normal Decomposition

Classical seasonal decomposition employed additive methodology decomposing unemployment series into trend, seasonal and residual components using 12-month periodicity. This approach provided interpretable components essential for understanding Malaysia's labor market dynamics. For general unemployment series, decomposition showed clear seasonal patterns with unemployment fluctuations aligning with economic cycles. The trend component showed three distinct phases: pre-pandemic stability (2010-2019) with gradual decline from ~3.5% to ~3.0%, pandemic disruption (2020-2021) with sharp increases to 5.3% peak, and recovery normalization (2022-2025) with gradual return toward historical norms. The seasonal component maintained consistent patterns across these phases, indicating strong seasonal dynamics independent of trend evolution.

Residual analysis showed random uncorrelated patterns confirming the decomposition was successful. The residuals had higher volatility during COVID-19 (2020-2021) but remained random, indicating that the systematic components were properly separated.

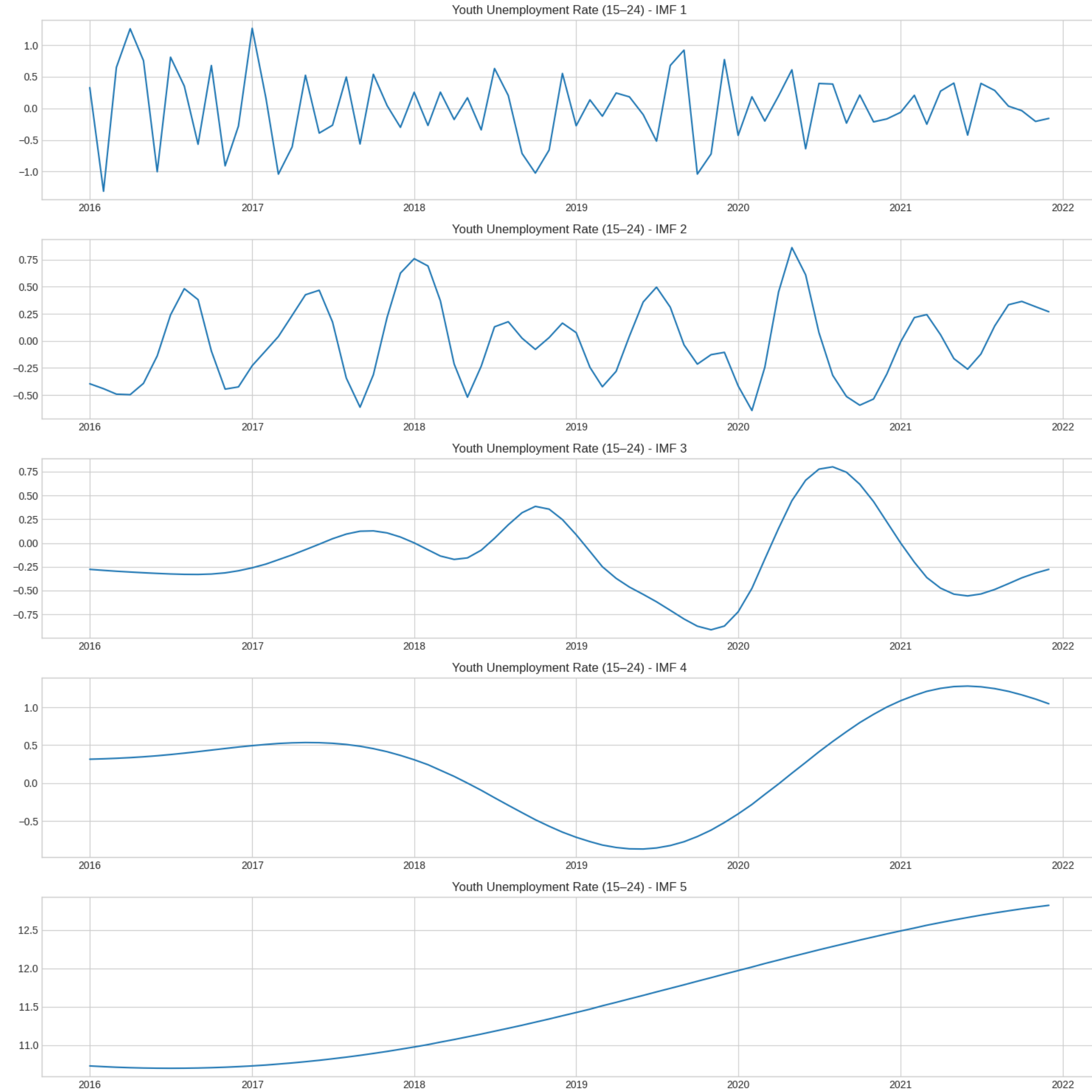

Figure 6. *Youth unemployment rate (15-24) decomposition*

As shown in Figure 6 above, Youth unemployment decomposition revealed more pronounced seasonal effects reflecting educational calendar influences and graduate entry patterns. The 15-24 age group demonstrated stronger seasonal components compared to the 15-30 group, indicating different labor market integration patterns. Peak seasonal unemployment coincides with graduation periods and potential summer employment searches thus creating predictable cyclical patterns valuable for forecasting.

3.3.2 CEEMDAN Decomposition

While traditional decomposition proved adequate for general unemployment modeling, youth unemployment's seasonal complexity required advanced decomposition techniques. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) provided a more apt non-linear decomposition capability for youth unemployment LSTM modeling, inspired by successful applications in financial time series prediction (Li et al., 2022).

CEEMDAN offered advantages over traditional decomposition by adaptively identifying frequency components without assuming predetermined periodicity thus enabling for the capture of complex non-linear relationships across multiple temporal scales. The methodology iteratively decomposed time series into Intrinsic Mode Functions (IMFs) with progressively lower frequencies, providing comprehensive representation of unemployment dynamics.

For youth unemployment 15-24 series, CEEMDAN extracted 5 distinct IMFs revealing different frequency components. IMF1 captured high-frequency fluctuations (1-3 months), IMF2 identified short-term patterns (3-6 months), IMF3 revealed medium-term cycles (6-12 months), IMF4 captured long-term trends (multi-year periods), and IMF5 represented residual components. Each IMF enabled specialized LSTM modeling optimized for specific frequency characteristics.

Similarly 15-30 age group showed similar IMF structure however with lower-frequency components, indicating more stable employment patterns for older youth. This decomposition approach proved particularly valuable where complex interactions between educational cycles, economic conditions and demographic transitions create multi-scale temporal dependencies difficult to capture with traditional methods.

CEEMDAN decomposition enabled individual modeling of each IMF component using specialized LSTM architectures before reconstruction, contributing to a better forecasting performance by separately capturing short-term volatility and long-term structural patterns.

3.4. Train/Test Split

We used a chronological data splitting approach to maintain the natural time sequence while ensuring good coverage of different economic periods. For general unemployment data spanning 2010-2025, we allocated the training period from 2010-2022 covering 156 months, followed by a validation period from February to December 2023 with 11 months, and a test period from February 2024 to March 2025 containing 14 months, as illustrated in Figure 7.

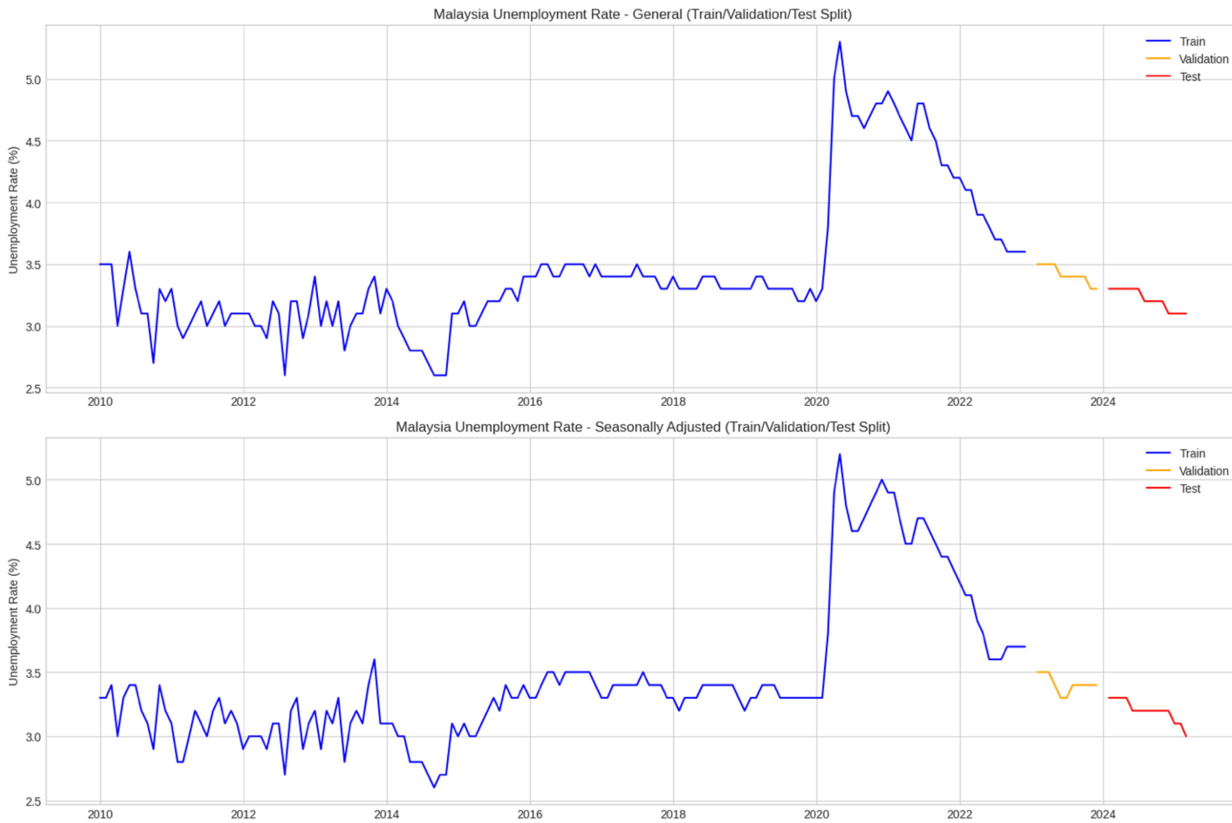

Figure 7. *General Unemployment Data Splitting*

Youth unemployment modeling required a modified approach due to shorter data availability from 2016-2025. The training period covered 2016-2021 with 72 months, validation spanned 2022-2023 with 23 months, and the test period included 2024-2025 with 14 months. This strategy balanced the need for adequate training data with sufficient out-of-sample testing across both age groups. This data splitting ensured that there is preserved representation of different economic cycles while enabling enough validation across both normal and crisis periods, ensuring our models would generalize well for practical forecasting applications.

3.5. Training Model

We developed three distinct modeling approaches to capture different aspects of unemployment dynamics for each of the datasets we used for modelling. Initially, we implemented ARIMA models to establish baseline benchmarks for traditional time series forecasting. Subsequently, SARIMA models were employed to address seasonal patterns and stationarity requirements more comprehensively. Finally, LSTM neural networks were utilized to leverage their capacity for modeling complex non-linear relationships and long-term dependencies prevalent in unemployment data.

3.5.1 ARIMA

The ARIMA implementation process began with automated parameter selection through pmdarima's auto_arima function, which systematically optimized the Akaike Information Criterion (AIC) across a comprehensive parameter space ranging from (0,0,0) to (3,3,3) (Li et al., 2022). This stepwise search approach carefully evaluated various model specifications while prioritizing both model simplicity and forecasting accuracy to ensure optimal performance.

Our analysis consistently identified ARIMA (0,1,0) as the optimal specification across both general and seasonally adjusted unemployment series, which indicates underlying random walk behavior with drift components. This finding suggests that unemployment rates exhibit highly persistent characteristics where current values are primarily determined by previous periods plus random error terms.

For youth unemployment modeling, we conducted a more extensive grid search across 27 different parameter combinations, ultimately identifying ARIMA (2,0,2) as optimal for the 15-24 age group, with similar specifications proving effective for the 15-30 age group too. Comprehensive model Checks showed that the that residuals were random and all coefficients demonstrated statistical significance, thereby proving that we chose an appropriate model
for the unemployment rate forecasting.

3.5.2 SARIMA

On the other hand, SARIMA modeling addressed present seasonal patterns through a comprehensive grid search across 60 parameter combinations combining non-seasonal orders [(0,1,1), (1,1,0), (1,1,1), (2,1,1), (0,1,2), (2,1,2)] with seasonal specifications [(0,1,1,12), (1,1,1,12), (2,1,1,12), (1,1,2,12), (2,1,2,12)] reflecting 12-month seasonality.

Optimization employed TimeSeriesFold cross-validation with 6-step ahead forecasting and Mean Absolute Error (MAE) criteria. Results identified SARIMA(0,1,2)×(0,1,1,12) as optimal for both general and seasonally adjusted unemployment with MAE values of 0.1092 and 0.1293 respectively.

Youth-specific SARIMA modeling achieved superior specifications with SARIMA(1,1,0)×(2,1,2,12) for 15-24 group and SARIMA(2,1,1)×(0,1,1,12) for 15-30 age groups, reflecting different seasonal adjustment patterns across age demographics. Similar to the ARIMA model, we conducted residual analysis that confirmed model's relevance to our forecasting task.

3.5.3 LSTM

For LSTM, we created a two-layer architecture design with 64 and 32 hidden units respectively, incorporating dropout regularization (0.3 probability) for overfitting prevention. Dense layer configuration included 25-unit hidden layer with ReLU activation followed by single-unit output for unemployment rate prediction.

Training methodology utilized Adam optimizer (learning rate: 0.001) with early stopping monitoring validation loss (patience: 15 epochs). Sequence preparation transformed unemployment data into 12-month overlapping windows, aligning with economic theory regarding annual cycles in unemployment dynamics.

For youth unemployment, advanced CEEMDAN-LSTM methodology individually modeled each IMF component using specialized architectures before reconstruction. This approach achieved exceptional performance: combined MAE 2.346 and RMSE 2.495 across age groups, demonstrating superior capability for capturing complex non-linear relationships in youth labor market dynamics (Hochreiter & Schmidhuber, 1997).

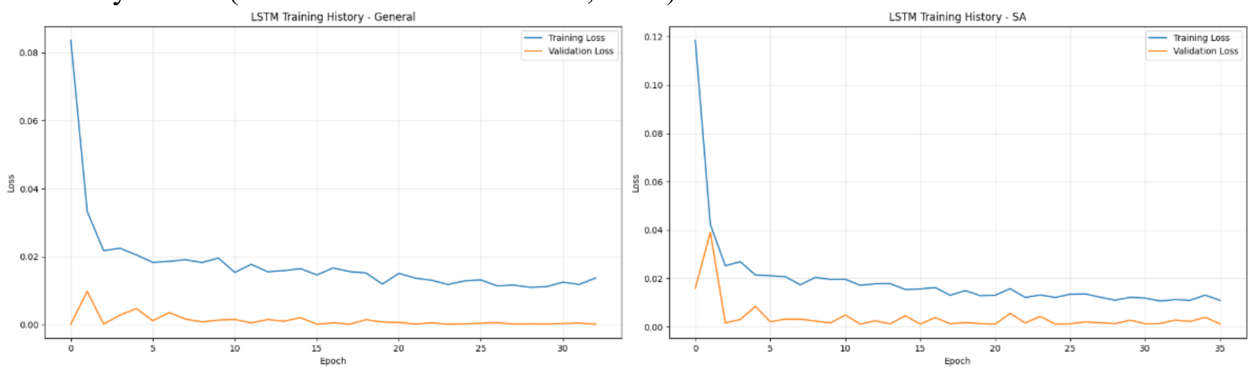

Figure 8. *LSTM Training History Plots*

Training history plots in Figure 8 above show the LSTM models exhibit smooth convergence with minimal overfitting for both general labour force and its corresponding seasonality adjusted version.

3.6. Evaluation of AI model

We evaluated all models using three widely used key metrics that provide clear and relevant insights Kamarudin & Vizie (2024). Mean Absolute Percentage Error (MAPE) served as our primary metric, calculated as shown in equation 1 below

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{Actual}_i - \text{Predicted}_i}{\text{Actual}_i}\right| \times 100 \qquad \text{eq. 1}$$

Additionally, Root Mean Square Error (RMSE) provided additional accuracy assessment with greater sensitivity to large forecasting errors, while Mean Absolute Error (MAE) offered complimentary error measurement in unemployment rate percentage points.

TABLE II. Primary Unemployment Results

Model	Dataset	MAPE	RMSE	MAE
LSTM	General	1.844%	0.074	0.058
LSTM	SA	2.72%	0.106	0.086
SARIMA	General	3.204%	0.12	0.101
SARIMA	SA	5.392%	0.192	0.17
ARIMA	General	12.076%	0.395	0.386
ARIMA	SA	15.707%	0.507	0.5

Note: general – is the raw labour force data, while SA is its seasonally adjusted version

As shown in Table II above, LSTM models achieved exceptional performance with 1.84% MAPE for general unemployment forecasting, establishing excellent benchmark results for Malaysian unemployment prediction. These results demonstrate that unemployment dynamics contain complex non-linear relationships that neural networks capture effectively as earlier investigated by Kamarudin & Vizie (2024).
SARIMA models provided the best traditional statistical approach, achieving 3.20% MAPE while successfully incorporating seasonal patterns in Malaysian labor market data. This performance validates the continued relevance of classical econometric methods when seasonal components are properly included.
ARIMA models showed poor performance with MAPE values exceeding 12%, confirming the necessity of incorporating seasonal components and advanced techniques for accurate unemployment forecasting.

TABLE III. Youth Unemployment Results

Model	Age Group	MAPE	RMSE	MAE
CEEMDAN LSTM	15-24	3.43%	0.786	0.764
CEEMDAN LSTM	15-30	2.72%	0.285	0.267
SARIMA	15-24	3.43%	0.42	0.357
SARIMA	15-30	3.70%	0.287	0.241
ARIMA	15-24	4.59%	0.521	0.478

The CEEMDAN-LSTM methodology outperformed traditional approaches while providing insights into age-specific labor market dynamics. The 15-30 age group showed superior forecasting accuracy, reflecting more stable employment patterns compared to the highly volatile 15-24 age group.
Lastly, as shown in Figure 9 below, Extended evaluation across 1, 3, 6, and 12-month horizons revealed important behavioral differences. LSTM models maintained consistent performance across all horizons with predictions converging to approximately 3.15-3.16% for general unemployment, indicating efficient architecture suitable for both short-term and medium-term forecasting.
SARIMA models showed moderate variation across horizons (3.61-3.72%), reflecting appropriate adaptation to different forecasting contexts while maintaining reasonable accuracy. ARIMA models maintained constant predictions due to their random walk nature, confirming their limitations for practical forecasting applications.

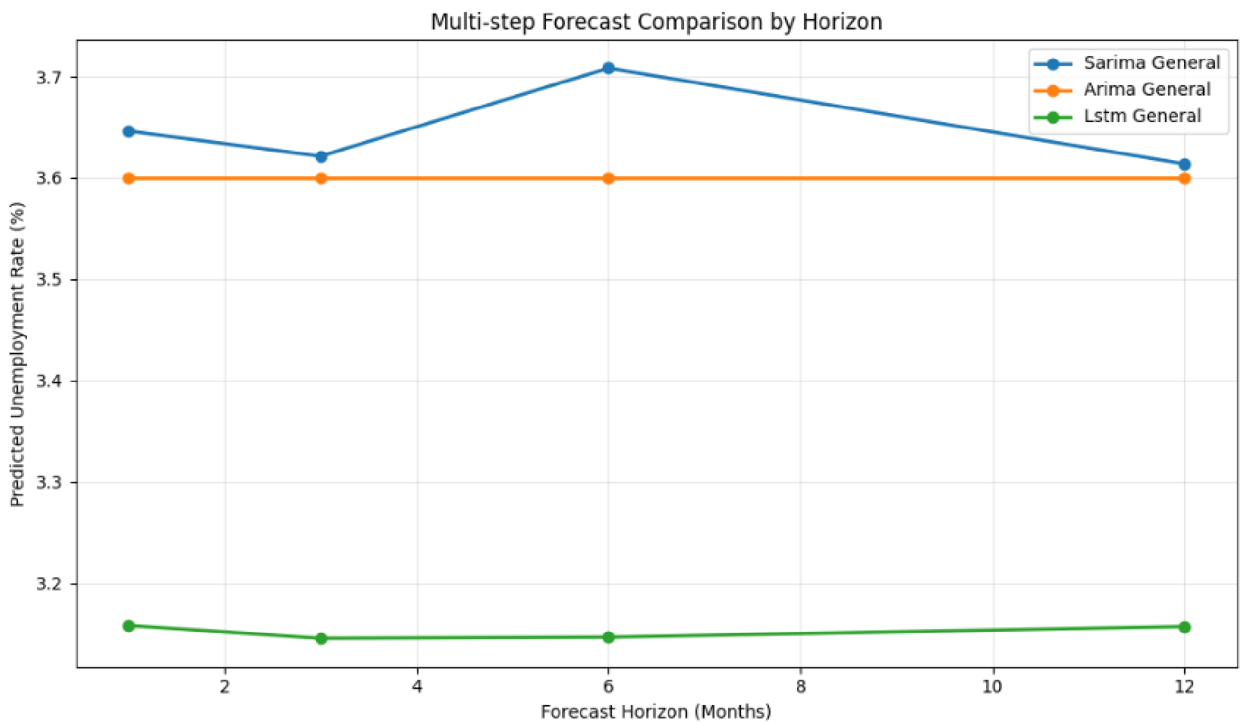

Figure 9. *Multi-Horizon Forecasting for General Unemployment Data*

4. AI Demonstrator

The AI demonstrator finalized the practical implementation of our research findings, translating sophisticated forecasting models into an accessible web-based interface for evaluating unemployment prediction methodologies. This demonstration platform enables the assessment of model performance and usability across various forecasting approaches, providing insights into practical analysis of Malaysia's Labour Market Dynamics.

4.1 System Architecture and Implementation

The demonstrator was developed using Python-based web technologies, specifically Dash and Plotly frameworks for interactive data visualisation. The architecture implements real-time API integration with Malaysian labour statistics services through a data management layer that handles automated retrieval, preprocessing, and validation of unemployment data across multiple datasets, including general unemployment, seasonally adjusted figures, and demographic breakdowns.

Model integration utilises joblib for SARIMA and ARIMA model persistence alongside TensorFlow/Keras for LSTM neural network deployment. The system provides a unified interface for comparing statistical and machine learning approaches, enabling the evaluation of different methodological frameworks without requiring specialised technical knowledge. The modular component architecture separates data processing, statistical analysis, transformation utilities and forecasting interfaces to support the systematic assessment of each analytical component.

4.2 Application Structure and Functionality

The Market Overview module displays real-time unemployment indicators and interactive time series visualisation across configurable periods. This interface enables the assessment of data presentation methods and their effectiveness in communicating labour market trends to different stakeholder groups.

The Data Explorer provides interactive analysis capabilities for examining multiple unemployment metrics simultaneously. This component facilitates the evaluation of data exploration workflows and their utility for preliminary analysis and pattern identification in unemployment datasets with users having the ability to choose between line and area charts as their preferred visualization.

The Statistical Analysis module implements comprehensive testing procedures, including stationarity assessment through Augmented Dickey-Fuller tests, normality evaluation and autocorrelation analysis via ACF and PACF visualisation. This component enables systematic evaluation of data preprocessing requirements and statistical validation approaches essential for our time series modelling.

The Data Transformation module demonstrates preprocessing methodologies, including log transformation, differencing operations and Box-Cox optimisation. Real-time validation through ADF testing provides immediate feedback on the effectiveness of transformations, supporting the assessment of data preparation strategies we used for various modelling approaches.

The Forecasting Hub implements automated prediction generation across LSTM, SARIMA, and ARIMA models with configurable forecast horizons from one to twelve months. The system supports forecasting across multiple demographic datasets, enabling users to select between general employment data for overall unemployment rate predictions or youth-specific datasets for targeted analysis of unemployment rates within the 15-24 and 15-30 age groups. This demographic segmentation capability enables a comparative assessment of forecasting accuracy across different population groups and facilitates an evaluation of age-specific unemployment dynamics. The component facilitates direct comparison of model performance through standardized evaluation metrics and visualization of prediction accuracy and the confidence intervals of our models across both general and demographic-specific forecasting scenarios.

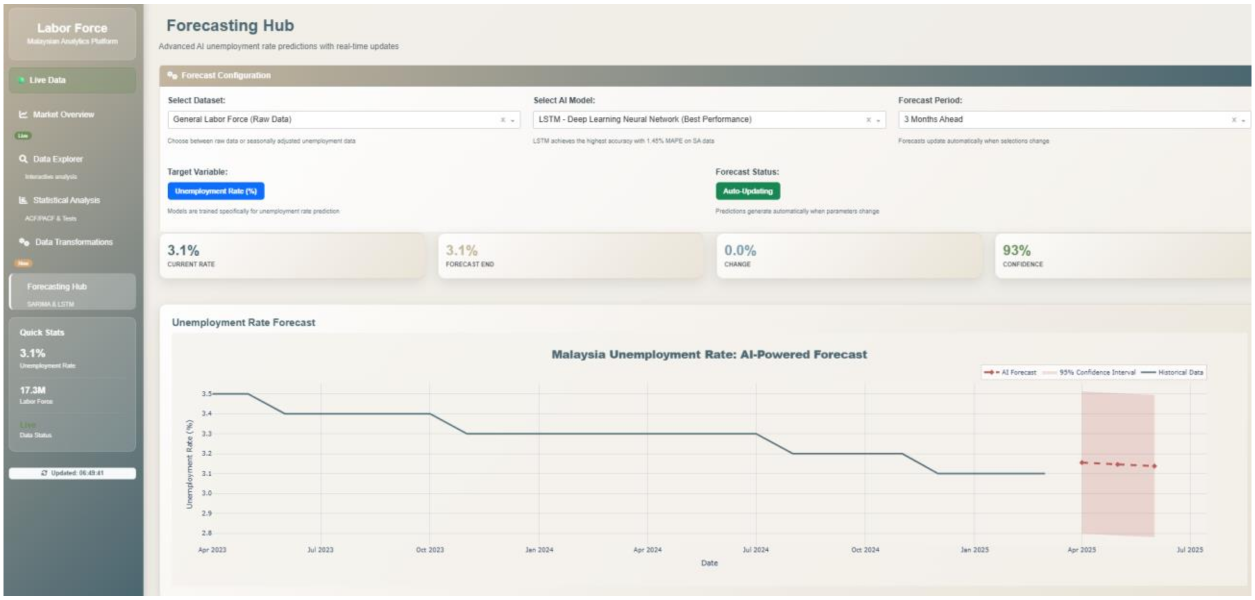

Figure 10. *AI Demonstrator's Forecasting Hub for general unemployment*

4.3 Technical Implementation and Assessment Capabilities

The system processes real-time data through API integration with Malaysian statistical services, implementing automated validation that includes temporal consistency checks and detection of missing values. This architecture enables the assessment of data quality management procedures and their impact on forecasting reliability.

Statistical analysis capabilities include multi-test validation frameworks and comprehensive correlation structure analysis supporting model identification procedures. The implementation enables the evaluation of statistical testing methodologies and their effectiveness for time series preprocessing and validation.

Model integration demonstrates deployment ready considerations through cross-validation testing and comprehensive accuracy measurement, including MAPE, RMSE, and directional accuracy metrics. Performance comparison across different modelling approaches provides empirical evidence for methodological assessment and selection frameworks.

4.4 User Interface and Accessibility Assessment

The interface design implements responsive layouts while maintaining professional presentation standards suitable for users. Navigation supports diverse user requirements by progressively disclosing advanced analytical features without compromising accessibility for basic operations.

Educational integration offers contextual explanations of statistical concepts and methodological approaches, facilitating the assessment of knowledge transfer effectiveness and user comprehension across diverse technical backgrounds. Lastly, Error handling and validation feedback systems demonstrate practical considerations for deployment in non-technical user environments.

4.5 Performance Evaluation and Validation

The demonstrator enables systematic assessment of forecasting methodologies through direct performance comparison across statistical and machine learning approaches. Historical validation capabilities demonstrate model accuracy through backtesting procedures, providing quantitative evidence for methodological evaluation.

The platform also facilitates the evaluation of uncertainty quantification approaches through confidence interval calculation and visualisation, enabling the assessment how reliable our forecasting model can be.

While we have not yet deployed the system, our team is considering future deployment of the analytical demonstrator using a free-tier platform such as Heroku. However, this deployment is contingent upon implementing an additional, more modern model based on recommendations gathered during the system development process.

5. Conclusions

This design project successfully addressed all four research questions through rigorous analysis of Malaysian unemployment forecasting methodologies. Our investigation revealed that LSTM neural networks significantly outperformed traditional statistical approaches, achieving exceptional 1.84% MAPE for general unemployment forecasting compared to SARIMA's 3.20% and ARIMA's poor 12.08% performance, establishing clear superiority of modern machine learning methods over conventional techniques.

The forecasting accuracy comparison demonstrated LSTM's consistent performance across multiple evaluation metrics, with youth-specific CEEMDAN-LSTM methodology achieving remarkable performance for the 15-30 age group. Multi-horizon forecasting validation confirmed reliable short-term prediction capabilities spanning 3-6 months, directly supporting workforce planning and potential policy formulation requirements.

Exploratory data analysis unveiled critical labor market insights, including pronounced COVID-19 impacts, strong seasonal patterns and demographic-specific unemployment vulnerabilities, with youth rates averaging 8.1 percentage points above general rates. The interactive AI demonstrator successfully translated forecasting models into accessible web-based tools, enabling real-time analysis and comparison across methodological frameworks.

This project provided invaluable learning experiences in advanced time series analysis, neural network implementation and end-to-end model development. We hope that the findings challenge conventional assumptions about statistical model adequacy for economic forecasting while demonstrating practical applications essential for graduating students entering Malaysia's evolving labor market landscape.

6. References

Aziz, A. A., & Foo, Y. C. (2024). Forecasting unemployment rate in Malaysia using ARIMA model. *International Journal of Research and Innovation in Social Science (IJRISS)*, 8(12), 464-476.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons..

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Ismail, N., Hussin, F., Satar, N. S. M., & Awang, M. A. (2022). Forecasting the unemployment rate in Malaysia during COVID-19 pandemic using ARIMA and ARFIMA models. *ResearchGate*.

Li, T., Hua, M., & Wu, X. (2022). Stock index prediction based on time series decomposition and hybrid model. *Entropy*, 24(2), 146. https://www.mdpi.com/1099-4300/24/2/146

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3), 599-607.

Kamarudin, A.N & Vizie, T. (2024). Comparing forecasting accuracy of time series models on Malaysian unemployment data. *Proceedings of the Postgraduate Seminar on Science & Mathematics*, 174-184.

Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4144-4147.

7. Appendix

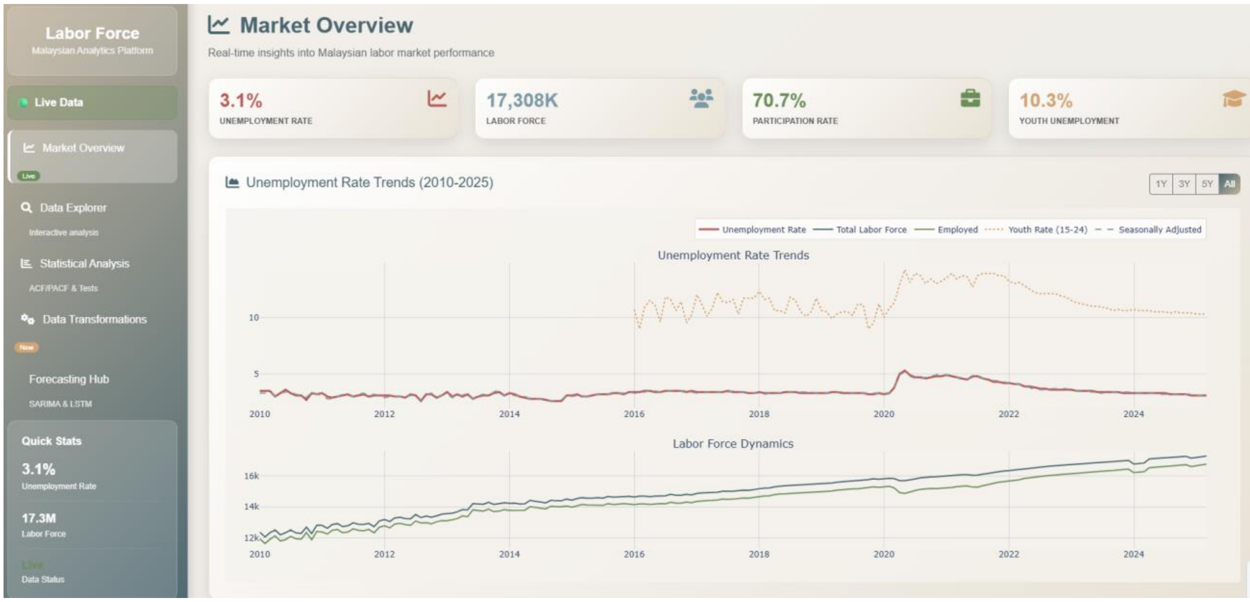

Figure 11. *AI Demonstrator's Market Overview*

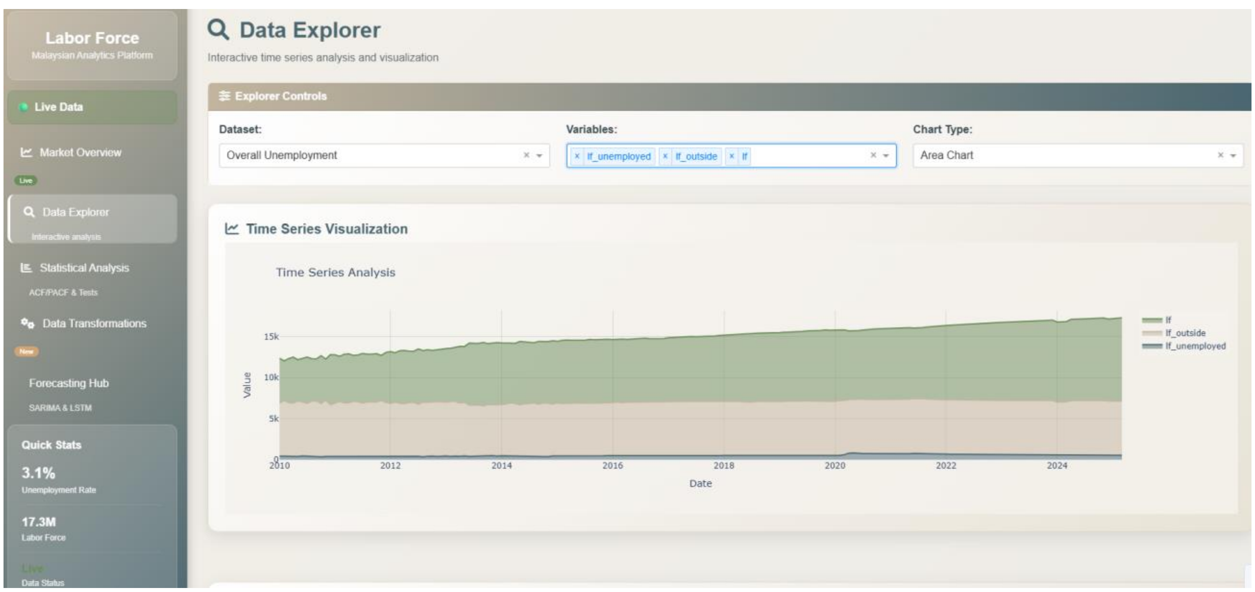

Figure 12. *AI Demonstrator's Data Explorer*

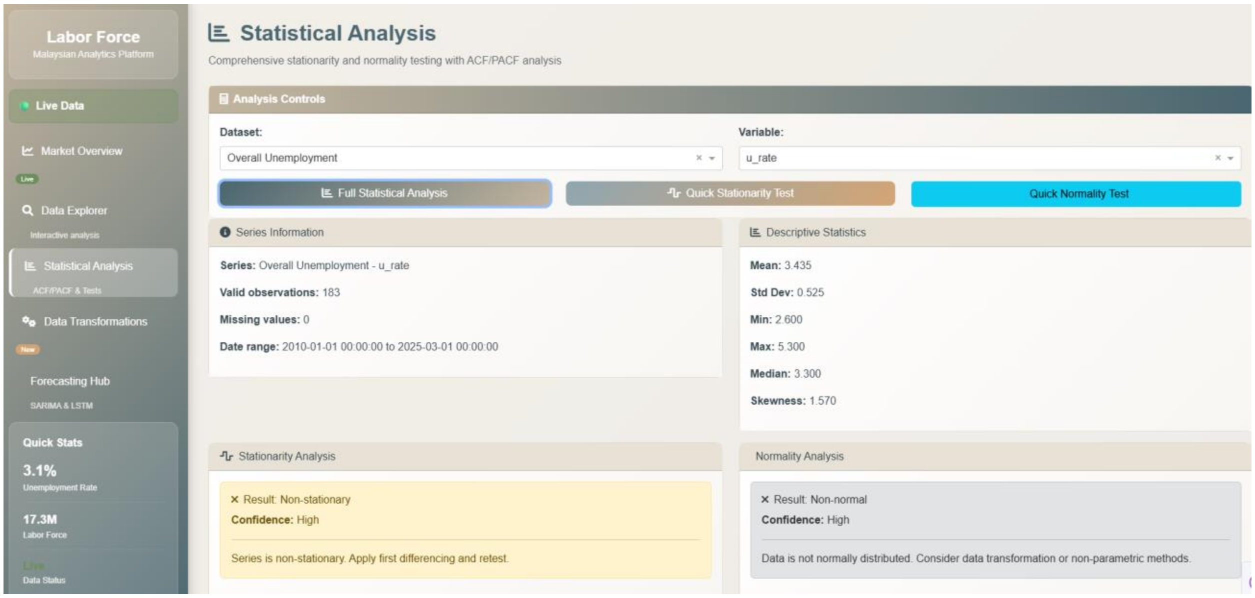

Figure 13. *AI Demonstrator's Statistical Analysis*

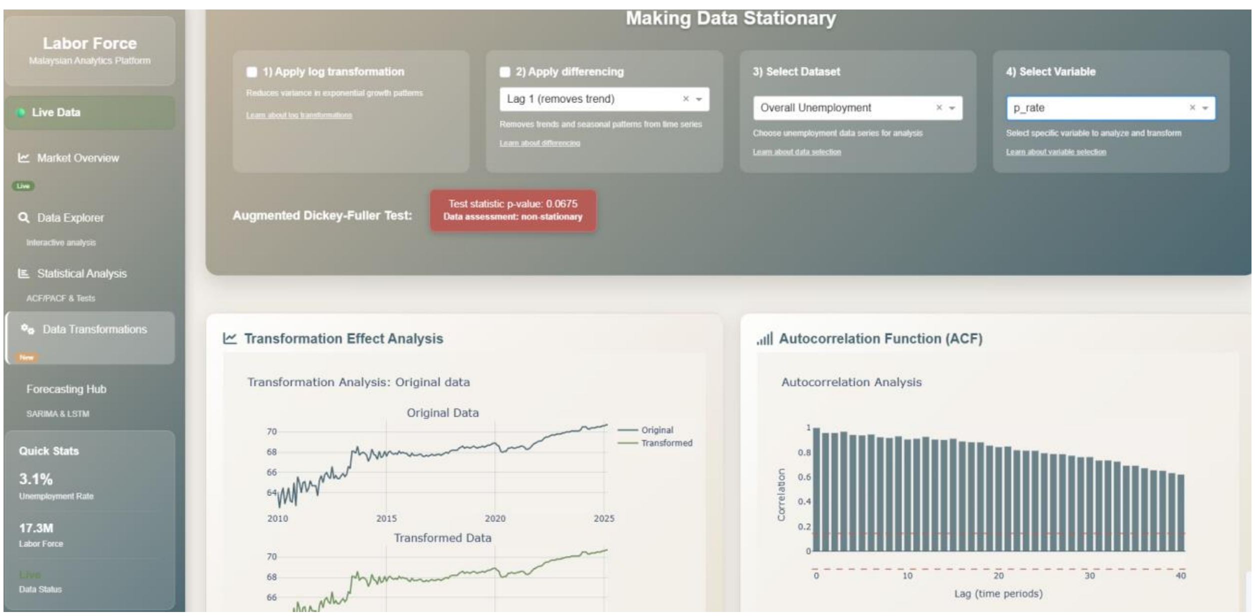

Figure 14. *AI Demonstrator's Data Transformation*

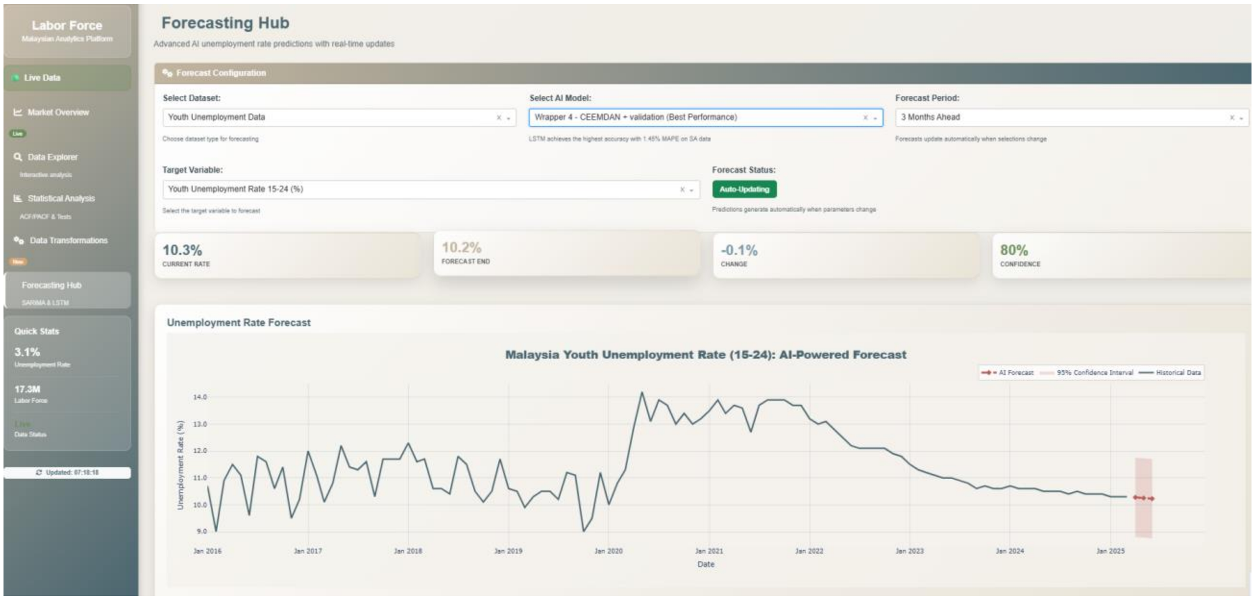

Figure 15. *AI Demonstrator's Forecasting Hub for youth employment*

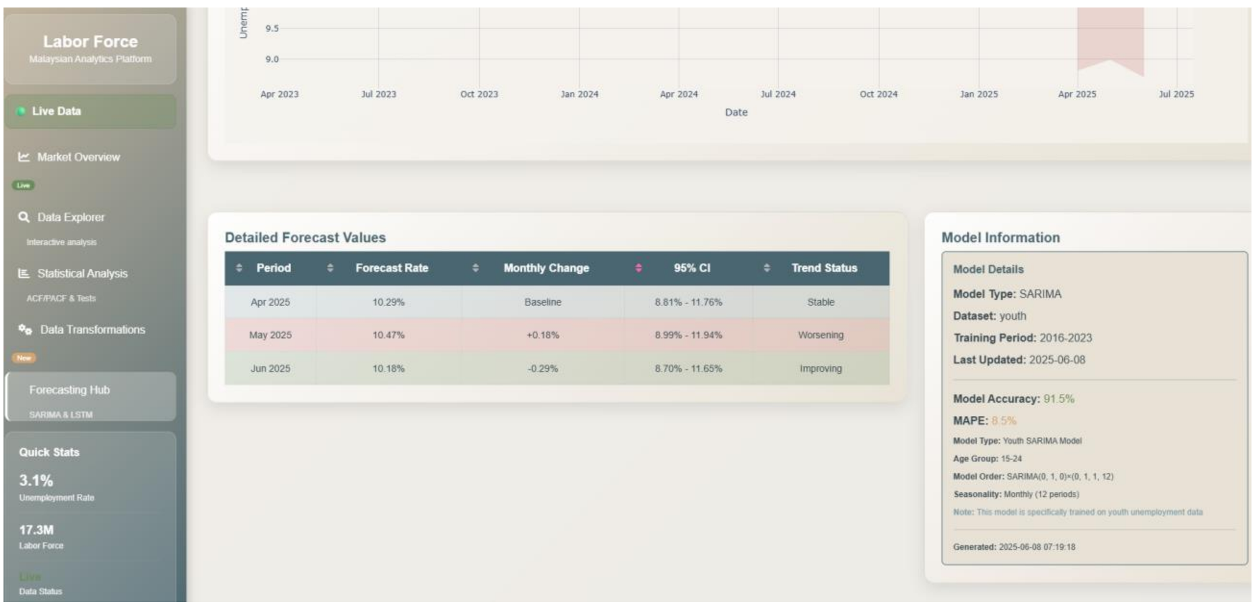

Figure 16. *AI Demonstrator's Forecasting Hub for youth employment with Detailed Forecast Values and Model Information.*