EVALUATING DEEP LEARNING METHODS ON FORECASTING OF ECONOMIC CYCLE INDICATORS

JOOST JANSEN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2022873

COMMITTEE

dr. Gonzalo Nápoles
Kyana van Eijndhoven

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 20th, 2024

WORD COUNT

6182

EVALUATING DEEP LEARNING METHODS ON FORECASTING OF ECONOMIC CYCLE INDICATORS

JOOST JANSEN

Abstract

Accurate forecasting of economic cycle indicators is of crucial importance for policymakers and business executives who make decisions based on incomplete and delayed economic data. Traditional models like Autoregressive Integrated Moving Average (ARIMA) have been widely used for economic forecasting. This study investigates the relative predictive performance of the deep learning models Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) compared to the ARIMA model, using economic cycle indicators of the Dutch economy. This study employs a dataset comprising twelve macroeconomic indicators, including gross domestic product (GDP), unemployment rates, and house prices. For each indicator, an ARIMA, LSTM and GRU model is built and optimized for the prediction of various forecasting horizons. The models were evaluated based on their MSE scores. The results demonstrate that GRU models outperformed both LSTM and ARIMA models on all forecasting horizons, particularly in scenarios involving one step ahead predictions, where the average MSE scores were reduced by 41.73% relative to the ARIMA model. The ability of GRU models to maintain lower error rates across different horizons suggests their potential for both short-term and long-term policy making.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

Data Source: The data for this thesis was acquired from the Central Bureau of Statistics Netherlands (CBS). The obtained data is stored in the publicly accessible database, Statline, freely available for use. This thesis project did not involve collecting data from human participants or animals. The CBS retains ownership of the data during and after the completion of this thesis. Figures: All figures used in this thesis were created by the author using Python libraries. Code: The code used in this thesis is not written by someone else, and does not come from another study. However, documentation of the python packages as well as code generated by chatGPT version 3.5 (OpenAI, 2023) were used to support carrying out the research. The software frameworks utilized are Pandas version 2.2.0 (McKinney, 2010), NumPy version 1.26.0 (Harris et al., 2020), Matplotlib version 3.8 (Hunter, 2007), Seaborn version 0.13.2 (Waskom, 2021), Sci-kit Learn version 1.4.1 (Pedregosa et al., 2011), SKtime version 0.28.0 (Löning et al., 2019), Keras version 2.15.0 (Chollet et al., 2015), Statsmodels version 0.14.2 (Seabold & Perktold, 2010), and Plotly version 5.9.0 (Inc., 2015). Technology and Typesetting: The thesis was typeset using the standard LaTeX format provided by Tilburg University, ensuring compliance to Tilburg University standards. No other typesetting tools or services were used. Additionally, no reference management software was employed beyond what is integrated into the LaTeX template. The author did not use any tools or services for paraphrasing, spell checking, or grammar corrections.

2 INTRODUCTION

2.1 *Motivation*

Policymakers and executives in businesses take decisions in real time based on incomplete information about current economic conditions. They largely rely on official statistics released by governing authorities to evaluate the state of the economy (ECB, 2021). The release of official statistics are lagged varying from a few days or weeks to several months after the reference period. Therefore, stakeholders have interest in accurate forecasting of economic cycle indicators[1]. Forecasting these indicators has been done for decades using traditional economic models (Vafin, 2020). This study aims to investigate whether deep learning methods could outperform traditional economic models in predicting economic cycle indicators.

2.2 *Project Definition*

This study uses economic cycle indicators of the Dutch economy. These indicators were sourced from the database of the CBS, which systematically tracks thirteen macroeconomic concepts and publishes them online. Together, these thirteen indicators collectively provide insights into the Dutch economy. The ARIMA model serves as the baseline for forecasting these indicators. The deep learning methods LSTM and GRU are applied to compare to the baseline performance. These deep learning models are able to learn long-term dependencies and are very popular for working with sequential data such as time series data (Yamak et al., 2019).

2.3 *Societal and Scientific Relevance*

Economic models like ARIMA have been widely used for economic indicator forecasting (Peter & Silvia, 2012; Vafin, 2020). There is, however, a growing interest in exploring deep learning techniques, in particular due to their ability to handle complex temporal patterns (J. M.-T. Wu et al., 2022; Yurtsever, 2023b). Figure 1 visualizes the frequencies of published papers concerning the prediction of macroeconomic indicators ("Dimensions", 2024). The first graph shows published papers containing the words 'ARIMA' and 'Macroeconomic forecasting', while the second graph shows published papers containing the words 'LSTM' and 'GRU' and 'Macroeconomic forecasting'.

[1] Economic cycle indicators are macroeconomic metrics that provide insights into the current state and future direction of an economy.

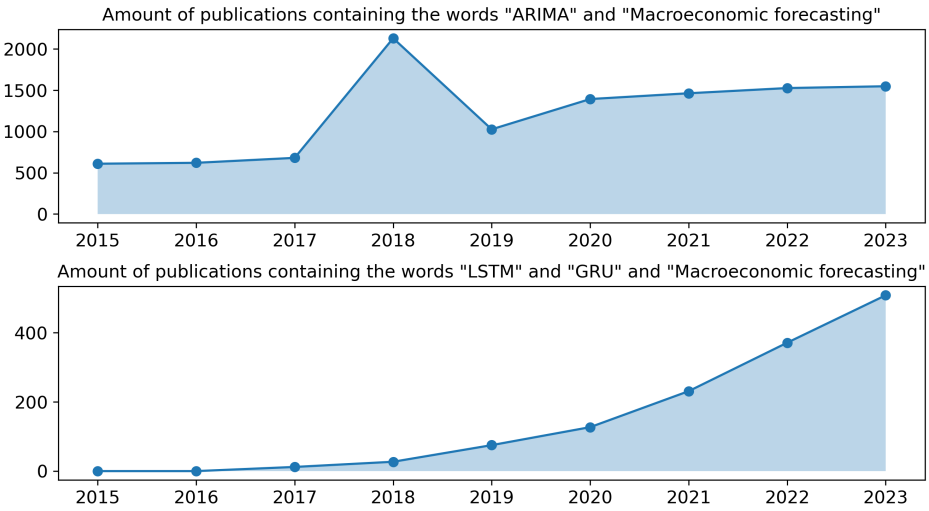

Figure 1: Frequency of publications ("Dimensions", 2024).

Figure 1 illustrates the widespread usage of ARIMA in forecasting macroeconomic indicators. It also illustrates the emerging trend of using deep learning methods for such predictions. However, a comparison of traditional economic methods and deep learning methods in the context of Dutch business cycle indicators is lacking. As stated before, accurate predictions of economic cycle indicators are beneficial for short-term decision making by policymakers and executives. The forecast is essential as it allows policymakers to anticipate to economic trends before making policy decisions (Pescatori & Zaman, 2011). These policies can include fiscal policies implemented by the government and monetary policies implemented by central banks, which operate independently of the government. Some examples of fiscal policies are government spending, taxation, or grants. Examples of monetary policies are interest rates and money supply. The further into the future predictions extend, the better policymakers and business executives can respond and adapt. Therefore, it is crucial to investigate the extent to which these indicators can be accurately forecasted over varying time horizons. This research might influence the prediction methods used by stakeholders.

2.4 *Research Questions*

Main RQ. To what extent can deep learning methods predict economic cycle indicators based on historical time series data when compared to traditional forecasting methods?

The primary aim of this research is to examine the predictive power of economic and deep learning methods on economic cycle indicators. The main research question is broken down into the following sub-questions:

RQ1. How do LSTM and GRU perform in predicting economic cycle indicators one-step ahead compared to ARIMA in terms of Mean Squared Error (MSE)?

RQ2. How do different window sizes influence the models' forecasting capabilities?

RQ3. What is the predictive performance of different indicators by always using the best-performing model?

The performance of the RNN models will be compared to the ARIMA model using MSE as the performance metric. Evaluating the results of the different models results in finding the best performing model. Varying in window sizes will determine how far ahead indicators can be predicted accurately. Accurate predictions further into the feature enhances policy-makers in their short-term policy making. It might be the case that some macroeconomic indicators may be easier to predict than others. Therefore, the best performing model will be used to predict the different indicators, resulting in an overview of the predictiveness of the indicators.

2.5 *Main Findings*

The main findings of this research provide insights into the predictiveness of LSTM and GRU compared to the traditional ARIMA model in forecasting Dutch macroeconomic cycle indicators. The results suggests that both LSTM and GRU generally outperform ARIMA in terms of forecasting accuracy, measured by MSE. Notably, GRU models excel, significantly reducing MSE by 41.73% compared to ARIMA when predicting one step ahead, suggesting its capability in managing sequential data and its potential in economic forecasting. Moreover, the GRU models consistently show the lowest error in terms of MSE across all forecasting horizons. This study also highlights the variability in predictability across different economic

indicators, indicating that GDP and house prices are relatively easy to predict, while bankruptcies and hours worked are more difficult to predict. This holds for both the traditional model as well as the deep learning models. These findings might provide guidance in enhancing macroeconomic forecasts conducted by stakeholders. This could improve decision-making processes for both short and long-term decision making in fiscal policies, monetary policies, and decisions made by business executives.

3 LITERATURE REVIEW

3.1 *Purpose of the Literature Review*

The purpose of this literature review is to provide an overview of existing research related to the forecasting of macroeconomic indicators using various time series models. Specifically, the models ARIMA, LSTM, and GRU. This review aims to summarize the current state of knowledge, highlight significant findings, and identify gaps in the literature that this thesis intends to address. The review will also justify the choice of methodologies and models employed in this research.

3.2 *Review of Existing Literature*

Forecasting macroeconomic indicators using time series models like ARIMA or deep learning methods like LSTM and GRU have been researched in the past. This section presents significant papers regarding time series forecasting with traditional and deep learning models, their application and their results.

3.2.1 *Traditional Time Series Forecasting Models*

Traditionally, time series forecasting has been done using the ARIMA model, developed by Box et al. (1970). ARIMA is still a widely used time series model for forecasting economic indicators due to the accuracy in forecasting, interpretability and cheap computational costs (Kontopoulou et al., 2023). Demonstrating the continuing relevance of ARIMA models in macroeconomic research, Wabomba et al. (2016) published an article on forecasting Kenya's GDP. This study used annual GDP data from 1960 to 2012 to model and forecast GDP one year ahead using ARIMA models. The models were evaluated using the metric MSE. The best model was then used to predict GDP for the next five years. The model that yielded the lowest MSE had an MSE score of 0.0099, demonstrating the highest accuracy in forecasting the Kenyan GDP one step ahead. Even more recent papers exist with similar approaches in predicting GDP of various countries using ARIMA. For instance, Ali and Haleeb (2020) modelled the GDP of Sudan using ARIMA. Another example is the study executed by Abonazel and Abd-Elftah (2019), which also used MSE as performance metric. The findings are comparable to those of Wabomba et al. (2016), with MSE scores between 0.0076 and 0.0104. Although ARIMA is widely used in forecasting macroeconomic time series, it has several drawbacks. ARIMA models assume linear relationships, whereas macroeconomic time

series data may be more complex in nature. another drawback is that constant standard deviation in errors is assumed, which may not be the case in reality (Siami-Namini et al., 2018).

3.2.2 *Deep Learning Models for Time Series Forecasting*

The LSTM model, developed by Hochreiter and Schmidhuber (1997), and the GRU model, developed by Cho et al. (2014), will be applied to address the main drawbacks of ARIMA. These models do not assume linearity or constant variance in errors. LSTM and GRU networks can model complex, nonlinear relationships within the data, which may be present in macroeconomic indicators. Recent studies have demonstrated the effectiveness of these deep learning models in forecasting economic indicators. Hamiane et al. (2023) published an article on forecasting the US real GDP time series using LSTM networks. This study utilized quarterly GDP data from 1947 to 2022, collected from the Federal Reserve Bank of St. Louis. GDP data from the previous three quarters were used to predict one step ahead. The LSTM model employed a single hidden layer architecture. This structure achieved an MSE score of 0.010, showing prediction accuracy close to the findings of Wabomba et al. (2016). A study by Yurtsever (2023a) explored the predictability of unemployment rates in France, Italy, the US, and the UK using LSTM, GRU and a hybrid model combining both LSTM and GRU. The results show that GRU consistently outperforms LSTM on all three evaluation metrics across all four countries. The hybrid models showed comparable performance to the GRU models, suggesting that GRU might be better in capturing dependencies in macroeconomic time series.

3.2.3 *Comparative Studies*

Studies comparing ARIMA with LSTM and GRU on time series data have found mixed results, depending on the specific dataset and forecasting horizon. Yenilmez and Mugenzi (2023) compared LSTM and ARIMA in the prediction Rwanda's GDP per capita. The study utilized the annual GDP per capita of Rwanda from 1960 to 2021, collected from the World Bank's official website. The models were evaluated using multiple metrics, including MSE. The results show that ARIMA outperforms LSTM in predicting this indicator on almost every performance metric, emphasizing the continued relevance of ARIMA to this day. GDP is an indicator researched in this study as well. Therefore, we might observe similar results in predicting the Dutch GDP, where ARIMA could be very competitive as baseline model. Siami-Namini et al. (2018) published an extensive comparison of the performance of ARIMA, LSTM and GRU on financial time series data. Historical monthly financial time series were extracted from January 1985

until August 2018. The data included the Nikkei 225 index, the NASDAQ index, the S&P 500 price index and some more indexes. This study shows that LSTM outperforms ARIMA in predicting all the indexes researched by a reduction of 85% in terms of Root Mean Squared Error (RMSE). Therefore, it might be the case that the investment indicator is better predicted by LSTM than by ARIMA or GRU. Additionally, it is essential to acknowledge that comparative studies present varying findings, indicating that there may not be a single best model for all indicators. It is reasonable to assume that some indicators are best predicted by different models. Yamak et al. (2019) argues that LSTM and GRU models perform better on larger datasets, as indicated by previous research. However, due to the relatively small dataset used in their study, ARIMA outperformed these models. On the one hand, given the limited number of data points in this research, it is plausible that ARIMA might be a better model for forecasting most macroeconomic indicators. On the other hand, the number of data points used by Siami-Namini et al. (2018), falls within the same range, namely monthly data dating back for a couple of decades. This suggests LSTM could outperform ARIMA in forecasting most macroeconomic indicators. This study will demonstrate which model is likely to be the best forecasting model for each indicator.

3.2.4 *Model Comparison on Various Forecasting Horizons*

Goncalves et al. (2023) assesses the impact of forecasting horizon on the performance of traditional econometric models and machine learning models in forecasting stock market prices. This study used the daily closing prices data of the Brazil IBX50, between 2012 and 2022. They compared the performance of ARIMA and LSTM on a variety of forecasting horizons. Their results suggests that ARIMA predicts better for small forecasting horizons, and that ARIMA loses predictive power when the forecasting horizon increases. They also concluded that LSTM models are more consistent on increased forecasting horizons, outperforming ARIMA. The results of this study are consistent with the findings of Gavilanes (2022), who studied the predictive performance of LSTM and ARIMA on a univariate time series over various forecasting horizons. In this study, exchange rates were predicted by both models on a short-term and long-term period. The results show that ARIMA loses predictive power over time. This study also investigates the predictive performance of ARIMA and LSTM on varying forecasting horizons. We might observe the same results, where deep learning models outperform traditional economic models in predicting macroeconomic time series of the dutch economy over longer periods of time.

3.3 *State-of-the-Art and Knowledge Gaps*

3.3.1 *Current State-of-the-Art in Macroeconomic Forecasting*

A milestone paper in the field of time series forecasting is the paper about the M4 Competition, where 100,000 time series and 61 forecasting methods are used with the aim to enhance forecasting accuracy ("The M4 Competition: 100,000 time series and 61 forecasting methods", 2020). The competition utilized a total of 19,402 macroeconomic time series. The paper discusses the performance of traditional statistical methods like ARIMA and compares them with more recent machine learning and deep learning models. Both ARIMA and simple RNN models are utilized in this study. ARIMA performed surprisingly well against the RNN models according to this study, although combining traditional statistical methods with deep learning approaches resulted in the most accurate forecasts. The findings in this paper align with the findings of the most recent hybrid aproach papers. Many of the hybrid models that are currently used in the literature combine ARIMA and an artificial neural network (Kumar et al., 2023). A key example is a study done by Dave et al. (2021), who aims to provide an accurate prediction of Indonesia's future exports by integrating ARIMA and LSTM models. The idea of combining the traditional model with the deep learning model is to take advantage of the best features of both methods, namely ARIMA's strength in handling linear components and LSTM's ability to manage non-linear components of the data. The results of this study demonstrate that the forecasting performance of the traditional ARIMA model is comparable to that of the LSTM model across all four performance metrics, including Mean Absolute Percentage Error (MAPE). Specifically, the standalone ARIMA model achieved a MAPE of 9.38%, while the LSTM model achieved a MAPE of 8.56%. However, the hybrid ARIMA-LSTM model significantly outperforms both individual models, obtaining a MAPE of 7.38%. These scores imply that the hybrid approach is more accurate, demonstrating the advantages of combining linear and non-linear modeling techniques for time series forecasting. Despite the extensively studied and high-performing hybrid models, this research focuses on comparing the predictive power of ARIMA, LSTM, and GRU on Dutch macroeconomic indicators, as this specific study is still lacking. Therefore, hybrid models fall outside the scope of this study. However, they could potentially be applied in further research.

3.3.2 *Knowledge Gaps*

Despite extensive research dedicated to forecasting macroeconomic indicators, some knowledge gaps remain. Numerous studies have employed

traditional methods as well as RNN models to predict these indicators. For example, the earlier discussed M4 competition ("The M4 Competition: 100,000 time series and 61 forecasting methods", 2020). However, there is a noticeable lack of comparative analysis on the predictive performance of ARIMA, LSTM, and GRU on Dutch macroeconomic indicators. The literature does not address how ARIMA, LSTM, and GRU models compare in terms of forecasting accuracy. Additionally, there is limited research on the predictiveness of each indicator. Addressing these gaps would provide valuable insights into the predictive performance of these methods on forecasting Dutch macroeconomic indicators. Furthermore, this study provides a comparison between indicators in their predictiveness. Moreover, this study includes a variety of forecasting horizons, which enables comparison of the performance of each model and indicator across various forecasting lengths.

3.4 *Justification of Methodological Choices*

The time series used in this study are univariate, meaning only one variable is recorded over time. ARIMA is designed for univariate time series as it uses past values of a single variable in the prediction of future values. Therefore, ARIMA is well-suited as baseline model. The choice for the deep learning models LSTM and GRU for comparison is based on their structure. These models are designed to adress the limitations of traditional Recurrent Neural Networks (RNN) in learning long-term dependencies. This could be beneficial in macroeconomic forecasting, where trends and cyclic patterns are present. To ensure the time series data are not random and to identify seasonal or cyclic patterns, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for each time series will be plotted. These plots visualize the structure within the data (Elsaraiti & Merabet, 2021). For each of the time series, an ARIMA, LSTM and GRU models will be fitted for forecasting. A performance metric is needed to compare the predictive performance of these models. There are numerous metrics that are suitable for time series, such as MAPE, Mean Absolute Error (MAE), RMSE or MSE. All of these metrics are present in one or more reviewed papers. Each metric offers a slightly different perspective on model accuracy. MAPE measures errors as percentages, which could be useful in comparing performance across different scales. MAE calculates a simple average absolute error. RMSE and MSE square the errors, penalizing larger errors more. This is important when bigger errors have greater impacts in practical uses of the models.

4 METHOD

4.1 *Research Methodology*

The research methodology involves several key phases. First, indicator data is sourced from Statline. Numerous pre-processing steps needs to be taken to ensure its suitability for building the models. After pre-processing, the dataset is divided into training (80%), validation (10%), and test sets (10%). The training set is used for building the baseline ARIMA model and training the deep learning models. The validation set is used for optimizing the hyperparameters of both deep learning models. Once the models are built and optimized, their test predictions are compared against the true test set using MSE as the performance metric. The performance of the baseline ARIMA model is compared to that of the deep learning models to identify the best performing model. The best performing model is used for the prediction with different forecasting horizons to observe this affects prediction accuracy. Figure 2 shows a visual representation of the methodological approach.

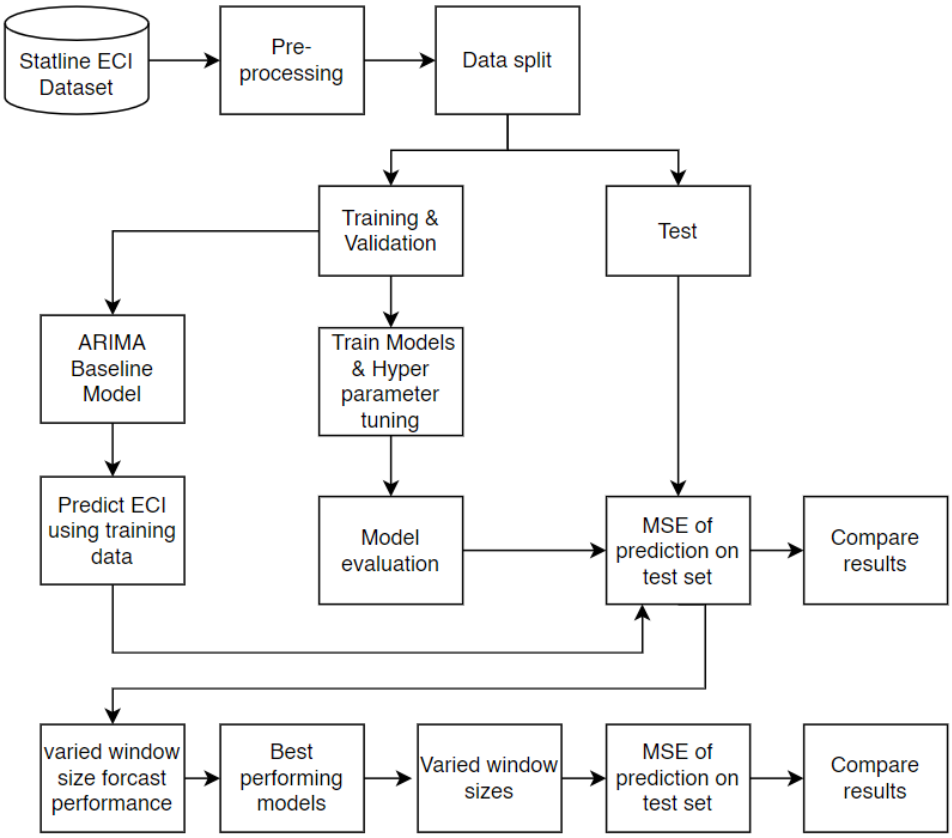

Figure 2: Visual representation of methodological approach.

4.2 *Models Overview*

4.2.1 *ARIMA*

The ARIMA model is widely used for time series forecasting, particularly effective for datasets that have temporal data. ARIMA is a powerful tool for forecasting future values using past values. Its capability to handle trends and seasonality makes it suited for macroeconomic indicators (Kontopoulou et al., 2023). The interpretability of ARIMA is a big advantage compared to LSTM and GRU, as ARIMA is relatively easy to understand, making it valuable for policymakers and stakeholders. Additionally, ARIMA models are efficient, with lower computational costs compared to more complex models.

4.2.2 *LSTM*

LSTM networks, are designed to capture long-term dependencies in sequential data. LSTM networks are known for overcoming the vanishing gradient problem, making them highly effective for learning from long sequences of data (Menculini et al., 2021). LSTM networks are well-suited for time series forecasting where long-term temporal dependencies play a crucial role.

4.2.3 *GRU*

GRU networks are designed to efficiently handle sequential data using gating mechanisms similar to those in LSTM networks but with a simplified structure. GRUs are known for their training efficiency and performance, being simpler and often faster to train compared to LSTM while achieving comparable results (Yamak et al., 2019). This makes GRUs particularly suitable for predicting macroeconomic indicators, balancing complexity and performance. They are effective for data with short to medium-term dependencies, which is common in certain macroeconomic indicators.

4.3 *Dataset Description*

The dataset includes thirteen economic cycle indicators compiled by CBS. These indicators are collected and published by CBS on a monthly or quarterly basis, with data going back to the 1980's, although the availability of data for each indicator may vary. The dataset thus consists of thirteen time series with several hundred timestamps. This data can be downloaded from Statline, CBS' official platform for statistical information, and can be exported to any format. Some examples of these indicators include GDP, unemployment rates, export levels, and bankruptcies data. The investment indicator does not have montly or quarterly data available in the statline database and will, therefore, not be used in this research. CBS has developed an online dashboard in which these thirteen indicators are graphically displayed, allowing policymakers to easily access and understand economic trends and developments. Figure 3 shows two economic cycle indicators.

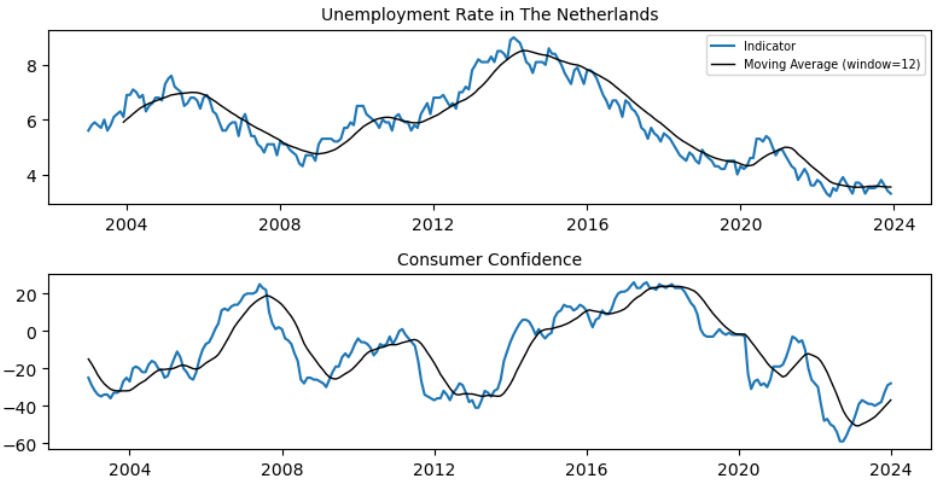

Figure 3: Unemployment Rate and Consumer Confidence in The Netherlands.

The indicators composed by the CBS are chosen because on individual level the indicators influence decision and policy making, and these indicators combined are used to quantify the economic conjuncture of The Netherlands. For each indicator, the measurement and their relevance will be discussed. The descriptions of all indicators are defined by the CBS (Centraal Bureau voor de Statistiek, 2016).

Exports measures the value of goods and services sold abroad, reflecting the competitiveness of the economy. They indicate the global demand for domestically produced goods and services. Policymakers use export data

to make up trade policies.

Producer confidence measures the sentiment among manufacturers about future production, orders, and stocks of finished products, providing insights into the manufacturing sector. This is measured through questionnaires. Business executives rely on this data to plan production schedules and invest in capacity expansions.

Unemployment rate measures the percentage of the labor force that is jobless and actively seeking employment. Among other things, unemployment rates influence government policy decisions relating to interest rates and job creation. Businesses use this data to make informed decisions about hiring, wages, and expansions.

Consumer confidence measures the optimism that consumers feel about the overall state of the economy and their personal financial situation. This is measured through questionnaires. This indicates their spending and saving behaviors, impacting business strategies in marketing and product development.

Hours worked measures the average number of hours worked by employees in a week. Business executives use this measurement to analyze productivity and labor cost efficiency, for example.

Gross domestic product measures the total market value of all goods and services produced within a country. It measures the of overall economic activity and health. The government keeps track of this indicator for making policies on taxation, spending, and borrowing.

House Prices measures the average cost of residential properties, providing insights into the real estate market's health. This indicator is used for monetary policy in adjusting interest rates. Real estate businesses and investors use this indicator to make investment decisions.

Manufacturing measures the output of the manufacturing sector. The relevance is equal to the relevance of the exports indicator, policymakers use manufacturing data to make up trade policies.

Consumption measures the total spending by households. It shows the consumer demand, guiding businesses in their marketing and production strategies to meet consumer needs.

Vacancies measures the number of job openings that are available but unfilled in the economy. It indicates labor market tightness. More vacancies suggest growing businesses and potentially robust economic conditions. Policymakers analyze vacancy data for labor market policies.

Bankruptcies measures the number of businesses that go bankrupt. This indicator provides insights into economic stress and the health of the business sector. It is used for evaluating the effectiveness of economic policies aimed at business sustainability.

Turnover of Temp Job Agencies measures the total revenues generated by temporary job agencies, reflecting the demand for temporary labor and the flexibility of the labor market. This measure is used to analyze changes in job trends, allowing policymakers and businesses to adapt to new economic situations more easily.

4.4 *Exploratory Data Analysis*

In order to gain insights into the data at hand, exploratory data analysis (EDA) was performed. This study investigated the size of the data, the distribution of the data, the autocorrelation and partial autocorrelation. The EDA in this subsection is visualized only for the indicators Unemployment Rate and Consumer Confidence due to clarity. Appendix A (page 37) provides time series visualisations for all indicators.

4.4.1 *Size of the data*

Table 1 provides insight into the indicators and the scopes of their respective datasets. This table describes the indicators investigated in this study, along with the tracked interval and the number of observations for each indicator.

Table 1: Dataset characteristics

Name Indicator	Interval	Period	Observations
Bankruptcies	Monthly	1981 - 2024	519
Consumer Confidence	Monthly	1986 - 2024	454
Consumption	Quarterly	1995 - 2023	116
Exports	Monthly	1995 - 2023	348
Gross Domestic Product	Quarterly	1995 - 2023	116
Hours Worked	Quarterly	1986 - 2024	77
House Prices	Monthly	2003 - 2022	348
Manufacturing	Quarterly	1995 - 2023	116
Producer Confidence	Monthly	1994 - 2024	363
Turnover Tempjob Agencies	Quarterly	2005 - 2023	76
Unemployment Rate	Monthly	2003 - 2023	252
Vacancies	Quarterly	1997 - 2023	108

4.4.2 *Distribution of the data*

Figure 4 visualizes the distribution of the indicators Unemployment Rate and Consumer Confidence. The overall shape provides a quick indication of how these indicators are distributed. For refference, a bell curve is displayed over both histograms. Consumer Confidence does not appear to be normally distributed. Indicators that follow a normal distribution often make the prediction results more accurate, whereas indicators that have a distribution including extreme outliers are likely to negatively influence the predictability (L. S.-Y. Wu et al., 2018).

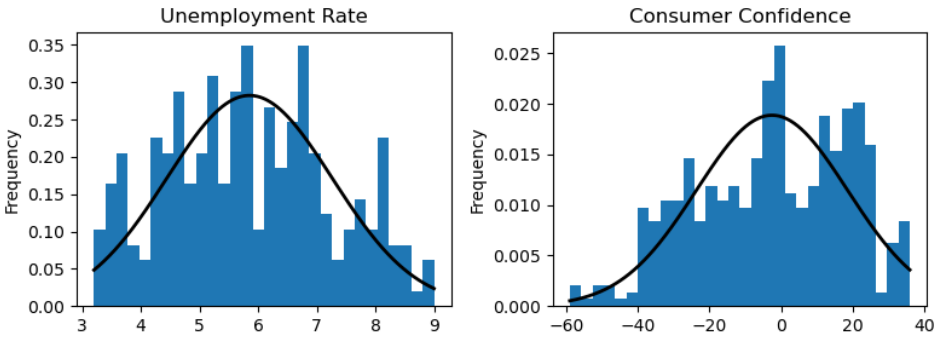

Figure 4: Unemployment rate and Consumer Confidence Histograms.

4.4.3 *Autocorrelation and partial autocorrelation*

ACF plots visualize the degree of similarity between a given time series and a lagged version of itself. Figure 5 visualizes two ACF plots of the indicators Unemployment Rate and Consumer Confidence. ACF plots for all indicators can be found in Appendix B (page 38).

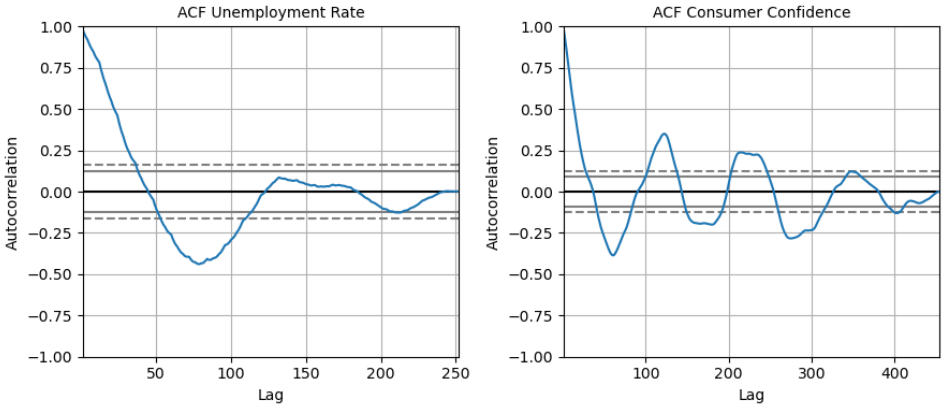

Figure 5: Autocorrelation plots.

Both plots show correlations outside of the significance bounds, suggesting the time series are not random. The autocorrelation decreases over time. This indicates that consecutive values tend to be similar, while the correlation weakens, meaning that values from, for instance, five years ago are less associated with recent values than values from one year ago. The fluctuations of the ACF above and below the significance bounds indicate periodic fluctuations. These fluctuations suggest presence of seasonality or cyclical behaviour. Figure 6 shows the PACF plot for the same indicators.

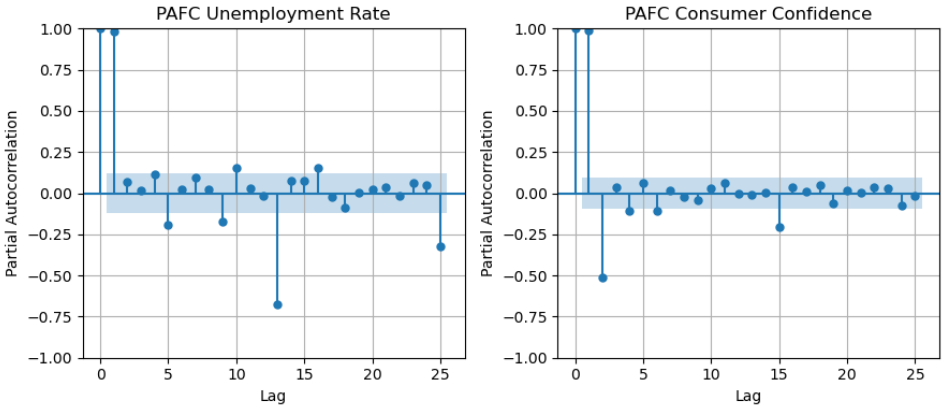

Figure 6: Partial Autocorrelation plots.

PACF measures the correlation between the series and a lagged version of itself, controlled for the effect of other lags in between. PACF is particularly useful in determining the order of an autoregressive (AR) process in time series modeling. Significant peaks in PACF suggest the number of lag terms needed in an AR model. PACF plots for all indicators can be found in Appendix C (page 39).

4.5 *Data scaling*

One of the goals of this research was to determine the predictiveness of indicators. In order to compare results between different indicators, the data was scaled using a standard scaler. The choice for the scaler was based on the fact that the standard scaler is robust to outliers and provides a standardized scale centered around zero, making it well suited for comparability.

4.6 *data split*

The time series data was transformed into sequences of observations of a period of 1 year. All sequences for the time series data were formed using the sliding window approach, visualised in Figure 7. Each line corresponds to a sequence of observations and the forecasting horizon, visualized as the orange dots, corresponds to the true future values we aim to predict. This approach results in X and y values, where X denotes the input data and y denotes the true label data. Figure 7 shows how the sequences were made for the monthly data. Quarterly interval indicators have a sequence

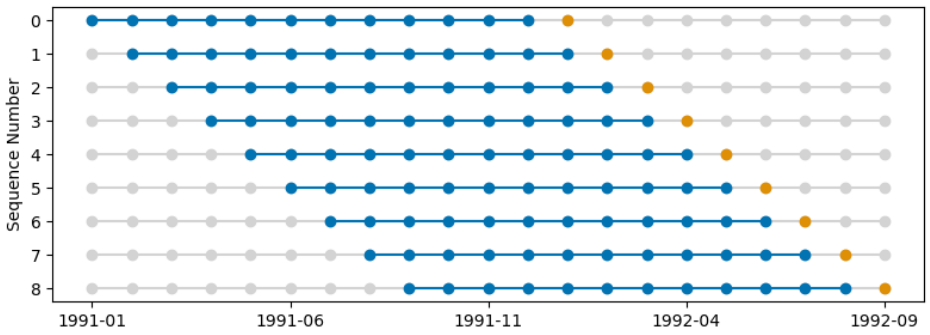

Figure 7: Sliding window approach for sequencing.

length of four. The second step in the data splitting process is splitting these sequences with their corresponding labels in training, validation and test data. This is done though train-test-split where 80% is placed in the

training set, 10% is placed in the validation set and 10% is placed in the test set.

4.7 *Evaluation Metrics*

The predictiveness of the indicators was measured by the commonly used evaluation metric MSE. Since the predicted values are predicted on a continuous scale, a regression metric is prefered. MSE penalizes large errors more heavily than small errors due to squaring the differences, which makes it suitable as an evaluation metric for accurately predicting these macroeconomic indicators.

5 RESULTS

5.1 *Hyperparameter Tuning*

The hyperparameter tuning for the deep learning models is done through the KerasTuner framework, which is specifically designed for optimizing hyperparameters in Keras models (O'Malley et al., 2019). For each of the twelve indicators, an attempt was made to find the optimal hyperparameters for the LSTM and GRU models through Random Search. Firstly, the search space is determined by commonly researched ranges of the hyperparameters. The search space is shown in Table 2.

Table 2: Search space for hyperparameter tuning.

Number of units	Dropout	Learning Rate	Activation Function
32 - 128	0 - 0.26	0.001, 0.005, 0.01	ReLu, Sigmoid, Tanh

The number of units is searched for in steps of four. The dropout rate is searched for in steps of 0.02. In order to not further increase the search space, some parameters are fixed. The sequence length is set to a year, the number of epochs is 15 and the batch size is set to 1. Along with the possible values for the other hyperparameters, a total of 2808 combinations can be made. Therefore, grid search would be way too computationally expensive, and random search is chosen. A total of 100 random combinations are searched for the LSTM and GRU models for each indicator. Figure 8 shows one of the 26 parallel coordinate plots. Each line in the plot represents a combination of hyperparameters that is tried.

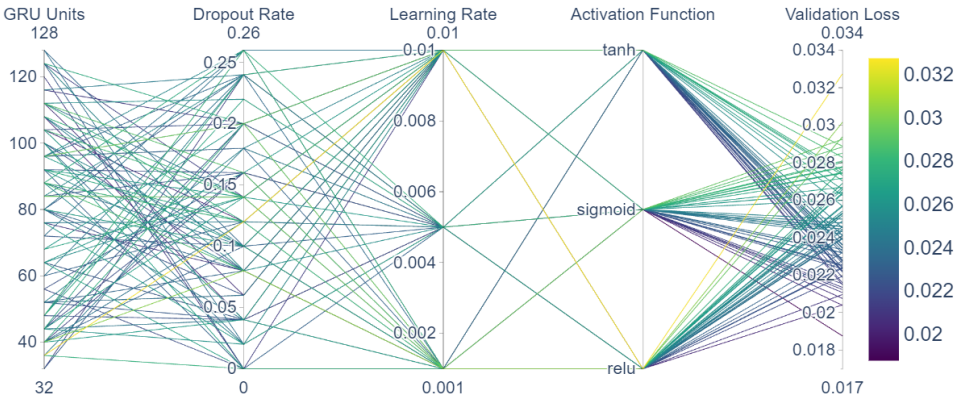

Figure 8: Parallel coordinates plot for the GRU model on exports data.

5.2 *Performance on training and validation sets*

Appendix D (page 39) shows a table with the optimal hyperparameter settings for the deep learning models for each indicator. For each model, the training and validation loss are stored. The loss function in this case is the MSE, where the objective is to minimize MSE. Figure 9 shows the training and validation loss of the best LSTM models for the unemployment rate and consumer confidence plotted over the epochs. The figure illustrates that for both unemployment rate and consumer confidence, the training and validation loss drops over the epochs. For consumer confidence, the validation loss lies above the training loss, which might indicate overfitting. However, there is no point in which the validation loss keeps increasing while training loss decreases. The gap between the lines in the unemployment rate plot are fairly close to each other, suggesting good generalization.

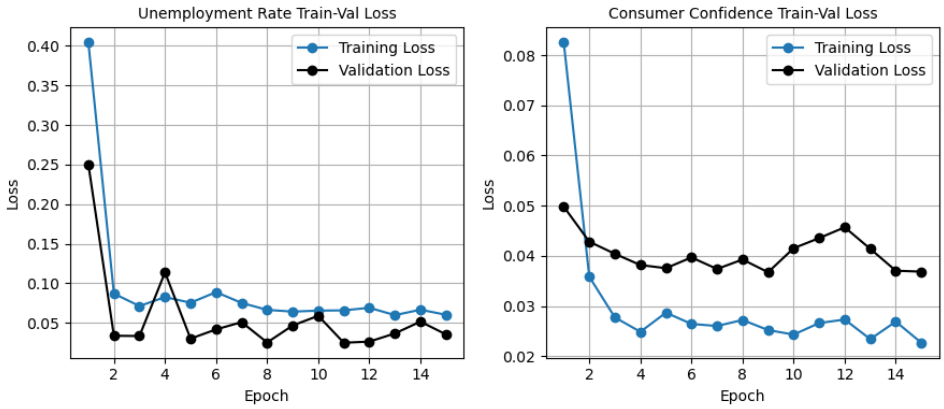

Figure 9: LSTM training and validation loss over epochs.

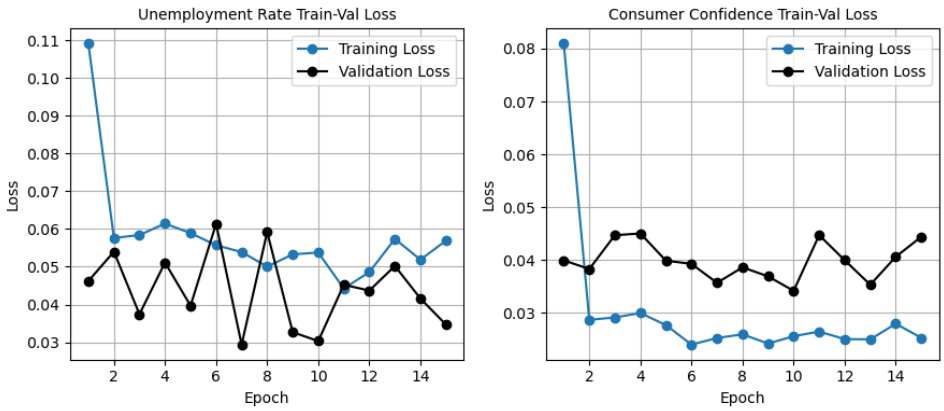

Figure 10: GRU training and validation loss over epochs.

Figure 10 shows the training and validation loss of the best GRU models for unemployment rate and consumer confidence plotted over the epochs. Like the LSTM models, the training and validation losses tend to decrease over the epochs, except for the validation loss of consumer confidence, which fluctuates around 0.04. The statements made about the gap between the training and validation loss lines in the LSTM plots also apply here. One of the goals of this research is to compare the deep learning models in predictiveness as well as comparing the predictiveness of indicators itself. Therefore, the box plot in Figure 11 visualizes the validation loss of all the trained models. Each box plot provide the median and the spread of the validation loss per indicator for both deep learning models. At a brief glance, one can compare the performance of the indicators and deep learning models. It seems that hyperparametertuning for GRU in general does have less influence of the validation loss, since for almost all indicators the spread is smaller than the spread of LSTM. A keen eye may have noticed that not all indicators are visualized in the box plots. The indicators producer confidence, hours worked and bankruptcies have validation losses so high that it is not possible to plot without stretching the image. See Appendix D (page 39) for a table with the best validation losses for each indicator.

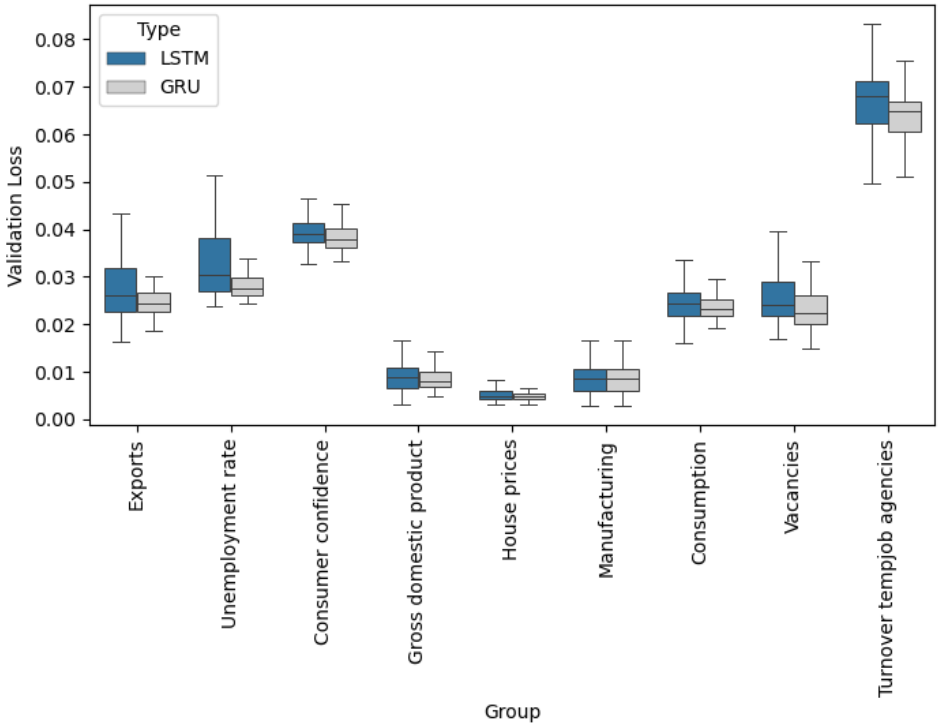

Figure 11: Boxplots of validation losses per indicator.

5.3 *Performance on test sets*

Table 3 presents the MSE results of time series forecasting using three different methods: ARIMA, LSTM, and GRU, where ARIMA is used as baseline model. Each deep learning method was evaluated both in a tuned and not tuned state. The MSE scores are presented across the 12 different indicators. Generally, tuning reduced MSE for the deep learning methods across most indicators. And for those three indicators where the Tuned model performed slightly worse, the error difference is relatively small. In most cases, deep learning models outperformed the baseline ARIMA model. For two indicators, the ARIMA model performed best. For indicators 1, 5, 10, and 12 the ARIMA model is greatly outperformed by the deep learning models, with an average MSE that is half as low. For indicators 2, 3, 8, and 9, the difference is close to zero. When the MSE values for each of the tuned models are summed up, a reduction of 16.12% in terms of MSE is observed for LSTM compared to ARIMA. For GRU, this percentage is 41.73%. In conclusion, The LSTM and GRU models, when tuned, often provide better or comparable performance to the AutoArima model. Based on the results presented in Table 3, GRU generally outperforms LSTM and ARIMA when forecasting dutch macroeconomic indicators one step ahead, based on previous values of the indicator.

Table 3: Performance of models on test set

	AutoArima	LSTM		GRU	
Indicator	Tuned	Not tuned	Tuned	Not tuned	Tuned
1	0.0612	0.0534	0.0328	0.0528	0.0439
2	0.0764	0.0751	0.0707	0.0707	0.0651
3	0.0375	0.0518	0.0332	0.0293	0.0297
4	0.0609	0.0501	0.0469	0.0475	0.0419
5	0.3334	0.3249	0.3233	0.3691	0.1492
6	0.0063	0.0082	0.0077	0.0091	0.0121
7	0.0044	0.0059	0.0071	0.0057	0.0064
8	0.0061	0.0056	0.0099	0.0099	0.0108
9	0.0106	0.0131	0.0102	0.0149	0.0118
10	0.0234	0.0316	0.0242	0.0209	0.0174
11	0.1566	0.1388	0.1329	0.1282	0.1381
12	0.2787	0.2229	0.1865	0.1277	0.0886

To clarify how the MSE values in Table 3 are calculated, Figure 12 displays the plots of four randomly selected predictions. These plots help visualize part of the consumer confidence predictions. The gray shaded area in

each plot shows the forecasting horizon, which compares the predicted values to the actual data. For each sequence in the test dataset, the squared difference between the forecasted and actual values was calculated. These errors were then averaged across all sequences to produce the MSE values listed in Table 3. The plots in Figure 12 demonstrate the forecasts from the three tuned models: LSTM, GRU, and ARIMA. In this example, the models were used to predict consumer confidence.

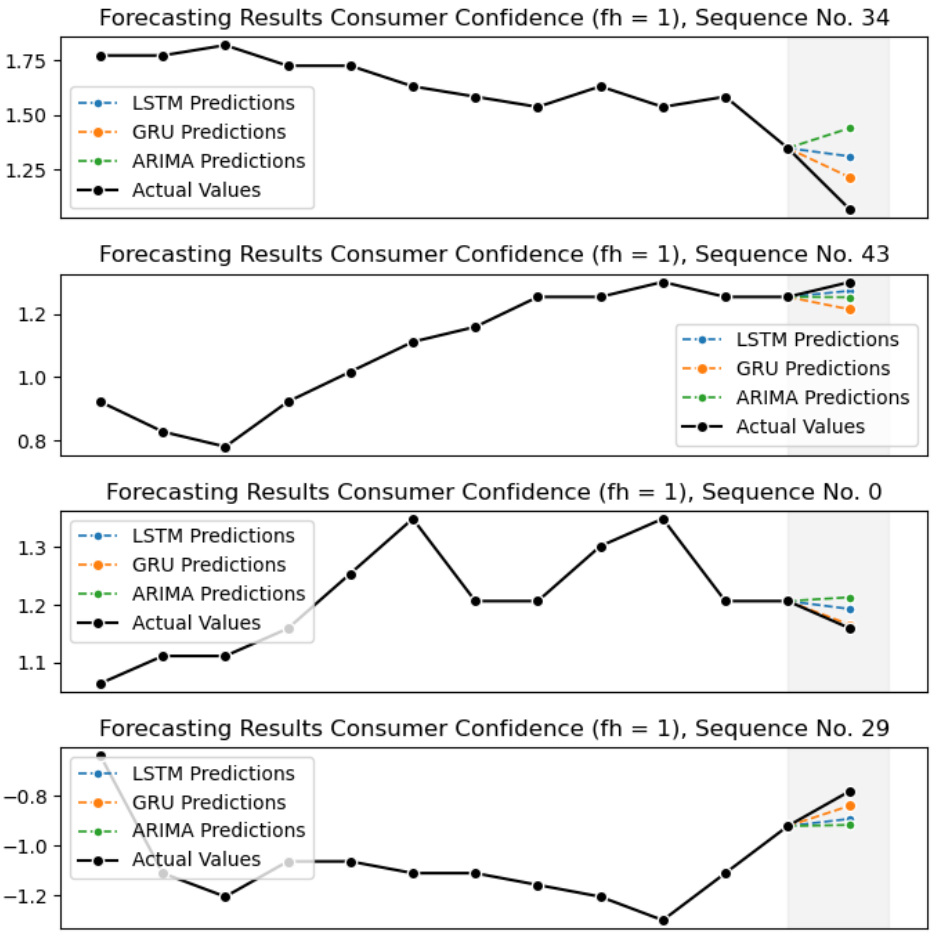

Figure 12: Forecasting of Consumer Confidence (fh = 1).

5.4 *Performance on test sets with increased forecasting horizons*

Table 4 presents MSE results of time series forecasting using three different methods: ARIMA, LSTM and GRU, with varying forecasting horizons of 3 and 6 time steps. For a better interpretation of the table, the best performing models with a forecasting horizon of three are highlighted

in light gray, and the best performing models with a forecasting horizon of six are highlighted in a darker gray. As expected, in most cases the MSE increases when the forecasting horizon increases. The average MSE increases from a forecasting horizon of three to six for AutoArima, LSTM, and GRU are 47.24%, 49.42%, and 51.09%, respectively. Noticeable is that most of the best performing models for both a forecasting horizon of three and six are performed using GRU. Again, on average, GRU is the best performing model. Additionally, it is notable that the increase in MSE when extending the forecast horizon is highly dependent on the specific indicator. For example, consider Indicator 10. The MSE for this indicator increases by up to six times when doubling the forecasting horizon for all three models. Conversely, for Indicator 11, the number of bankruptcies, the MSE only slightly increases when the forecasting horizon is doubled. This suggests that the sensitivity to forecast horizon length varies significantly between different indicators. Moreover, there is also an inconsistency in identifying the best model for each indicator across different forecasting horizons. For example, while the ARIMA model performs best for Indicator 6 and 7, at a forecasting horizon of 1-step ahead, the LSTM model outperforms it at a 3-step horizon, and GRU takes the lead at a 6-step horizon. Similarly, for Indicators 1 and 9, the LSTM model performs best at a 1-step horizon, but loses to GRU when increasing the forecasting horizon. The only model that consistently outperformed the other two models for several indicators across all forecasting horizons is the GRU. Specifically, for Indicators 4 and 12, the GRU model is the best-performing model at forecasting horizons of one, three, and six steps ahead.

Table 4: Performance of models on test set

	AutoArima		LSTM		GRU	
	fh = 3	fh = 6	fh = 3	fh = 6	fh = 3	fh = 6
1	0.0871	0.1682	0.0716	0.0847	0.0604	0.0792
2	0.6452	0.9969	0.5886	0.6601	0.6217	0.6992
3	0.0731	0.1237	0.0659	0.1244	0.0564	0.1081
4	0.1688	0.2354	0.1277	0.1931	0.1193	0.1648
5	0.5983	0.2911	0.3196	0.3202	0.2626	0.3078
6	0.0315	0.0232	0.0252	0.0153	0.0185	0.0296
7	0.0311	0.0298	0.0183	0.0172	0.0171	0.0156
8	0.033	0.0233	0.0344	0.0259	0.0208	0.0248
9	0.0432	0.042	0.0495	0.0454	0.0448	0.0511
10	0.1124	0.7115	0.1132	0.6629	0.1312	0.6071
11	0.1503	0.1767	0.1163	0.1512	0.1249	0.1579
12	0.2019	0.382	0.0748	0.098	0.0683	0.0907

Figure 13 provides visualizations of the unemployment rate predictions using four randomly selected sequences. These plots offer insights into part of the unemployment rate forecasts. The gray shaded area in each plot indicates the forecasting horizon, which is now three steps ahead instead of one. The plots compare the predicted values to the actual data. Figure 13 demonstrates the forecasts from the three tuned models: LSTM, GRU, and ARIMA, specifically for predicting the unemployment rate.

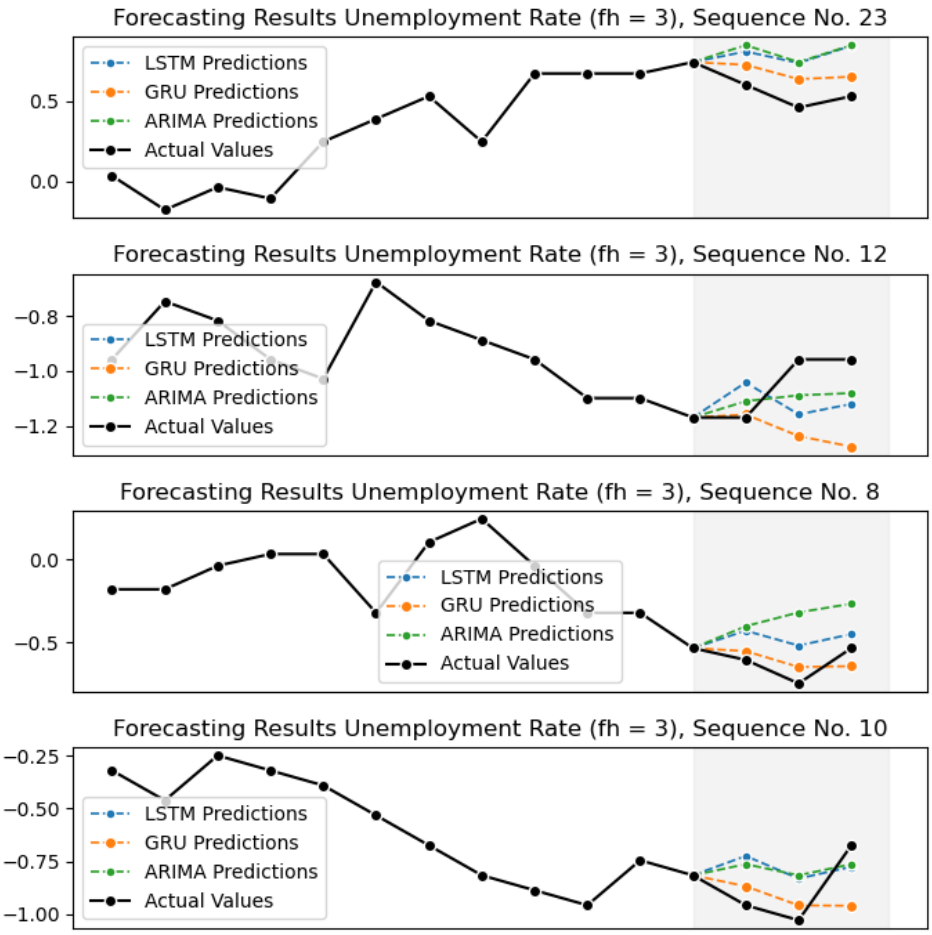

Figure 13: Forecasting of Unemployment Rate (fh = 3).

Figure 14 provides visualizations of the house price predictions using four randomly selected sequences. These plots offer insights into part of the house price forecasts. The gray shaded area in each plot indicates the forecasting horizon, which is now six steps ahead. The plots compare the predicted values to the actual data. Figure 14 demonstrates the forecasts from the three tuned models: LSTM, GRU, and ARIMA, specifically for predicting house prices.

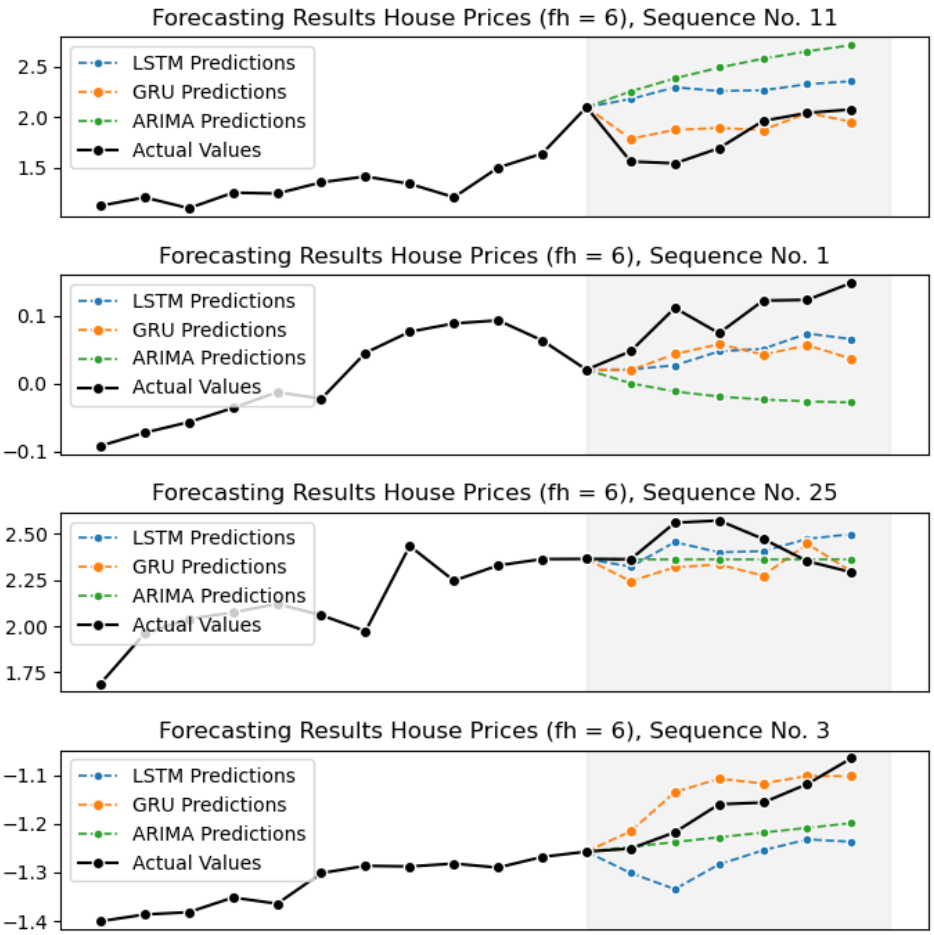

Figure 14: Forecasting of House Prices (fh = 6).

6 DISCUSSION

The main goal of this research was to examine the predictive power of economic and deep learning methods on economic cycle indicators. The evaluation of deep learning methods against ARIMA for forecasting economic cycle indicators has yielded several insightful observations. This discussion delves into the findings, the implications of the results, the limitations of the study, and potential future research directions.

6.1 *Performance Comparison*

Research question one investigated how LSTM and GRU perform in predicting economic cycle indicators one step ahead compared to ARIMA, in terms of MSE. The results indicate that deep learning models LSTM and GRU, generally outperform the traditional ARIMA model in predicting the Dutch Macroeconomic cycle indicators. GRU models, in particular, demonstrated superior performance, often achieving lower MSE values compared to both LSTM and ARIMA. To be precise, the average MSE dropped 16.12% for LSTM and 41.73% for GRU when compared to ARIMA. GRU significantly outperforms ARIMA and LSTM in forecasting one step ahead. This suggests that GRU's simpler architecture and efficient handling of sequential data could be better suited for this type of forecasting, especially in scenarios involving one-step ahead forecasting. Research question two examined how different window sizes influence the models' forecasting capabilities. The performance of the models were evaluated with increased forecasting horizons. As expected, the MSE increased with the length of the forecasting horizon for all models. However, GRU consistently showed better adaptability, maintaining relatively lower error rates even with extended horizons. This robustness indicates that GRU best captures temporal patterns in the Dutch Macroeconomic cycle indicators over varying time frames, compared to ARIMA and LSTM.

6.2 *Indicator Specific Observations*

Research question three examined the predictive performance of different indicators when always using the best-performing model. After training each model on their best hyperparameters, the economic indicators varied in predictability. Indicators like GDP, house prices, and consumption showed close to true value predictions. In contrast, certain indicators, such as bankruptcies and hours worked, had higher MSE values across all models, indicating difficulty in predicting these metrics accurately, independent from the model used. The results for the indicator GDP align

with the findings in the study of Yenilmez and Mugenzi (2023), where ARIMA outperformed the deep learning methods in the prediction of GDP. This was also the case for the results for unemployment rate. The GRU models outperform LSTM on every forecasting horizon, complementing research done by Yurtsever (2023a). Their results also align in the context of predicting unemployment rate, and the GRU models outperform LSTM models. The inconsistency in model performance across different indicators and forecasting horizons highlights the complexity of economic time series data. Economic indicators can have unique characteristics, seasonal patterns, and external influences that affect their predictability.

6.3 *Implications for Policy and Decision Making*

The predictive performance of deep learning models may have practical implications for policymakers and business executives. Moreover, more accurate forecasts of economic indicators can lead to better decisions regarding economic and monetary policies. For instance, precise predictions of unemployment rates can help with the timely implementation of job creation programs, while accurate GDP forecasts can inform tax policy adjustments. The ability of GRU models to maintain lower error rates across different horizons suggests their potential in both short-term and long-term policy making. Policymakers can rely on these models to anticipate economic trends and prepare strategies.

6.4 *Limitations*

6.4.1 *Scope and Boundaries*

This study focuses on predicting economic cycle indicators using ARIMA, LSTM, and GRU models, using historical data. The analysis is limited to a set of twelve macroeconomic indicators from the Dutch economy. These indicators are sourced from the CBS, and the study only uses data from this source. The research is constrained to the performance of these models within the specific context of the Dutch economy and the selected indicators.

6.4.2 *Methodological Constraints*

Due to the relatively small size of the test set and some small indicator datasets, for some indicators the test set became extremely small. This resulted in single predictions having a significant influence on the performance metrics. It could be the case that due to chance, relatively easy-to-predict values were placed in the test set, distorting the view of the

predictive performance. The same goes for the small size of the validation set. Although 100 models were trained to find the best hyperparameters, the limited size of the validation set could lead to inaccuracies in the true MSE. This issue is particularly relevant for datasets describing quarterly data, where the validation set might contain only a few data points. With such a small number of predictions forming the MSE, there is a higher chance that the selected best hyperparameters are not truly optimal, but rather the result of chance variations in the data. Consequently, the hyperparameters identified as the best may not consistently provide the best performance on larger datasets. Another limitation is the non-stationarity of the time series. The time series data were not transformed to be stationary before training the models. This is a disadvantage for the deep learning models, as patterns are easier to detect in stationary data. The ARIMA model, however, can inherently make the time series stationary through its differencing component. Lastly, while some hyperparameters were tuned using random search, many others were not optimized due to computational constraints. Parameters such as the optimizer, number of epochs, and batch size could be further optimized to improve performance.

6.4.3 *Generalizability*

The findings of this study are specifically applicable to the selected macroeconomic indicators of the Dutch economy and may not be directly generalizable to other contexts or economies. The small sample size and the specific characteristics of the Dutch economic indicators limit the ability to generalize the results to different datasets or indicators from different countries. Further research with larger and more diverse datasets is needed to validate the findings and enhance their generalizability.

6.5 *Future Research*

To enhance the reliability of the results, future research could optimize these models by addressing some of the methodological constraints of the current study. Increasing the validation and test size would reduce the high impact of individual predictions on the MSE, providing a more robust measure of model performance. It is also important to ensure data stationarity. Transforming the time series data to be stationary before training the deep learning models could improve their performance. Last but not least, upscale hyperparameter Tuning by Expanding the number of hyperparameters and their search space. Even though hybrid models were out of the scope of this research, future research could also entail combining statistical and machine learning models, since their proven

promising performance in recent studies. Applying hybrid approaches to these Dutch macroeconomic indicators might further improve predictive performance.

7 CONCLUSION

In summary, this study fills a gap in economic forecasting research by comparing the predictiveness of LSTM and GRU to ARIMA using macroeconomic indicator data from the Netherlands. The GRU model stands out, outperforming both the LSTM and the traditional ARIMA model in predicting economic indicators. The GRU models consistently achieve lower error rates across the forecast horizons, emphasizing their value in both short and long-term policy decisions. Policymakers might leverage these models to forecast economic trends and develop strategies.

REFERENCES

Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting egyptian gdp using arima models. *Reports on Economics and Finance*, *5*(1), 35–47. https://doi.org/10.12988/ref.2019.81023

Ali, H. M. H., & Haleeb, A. M. A. (2020). Modelling gdp for sudan using arima. *Munich Personal RePEc Archive*, (101207). https://mpra.ub.uni-muenchen.de/101207/1/MPRA_paper_101207.pdf

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Centraal Bureau voor de Statistiek. (2016, April). Beschrijving afzonderlijke indicatoren conjunctuurklok. https://www.cbs.nl/nl-nl/achtergrond/2016/17/beschrijving-afzonderlijke-indicatoren-conjunctuurklok

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. https://arxiv.org/abs/1406.1078

Chollet, F., et al. (2015). Keras.

Dave, E., Leonardo, A., Jeanice, M., & Hanafiah, N. (2021). Forecasting indonesia exports using a hybrid model arima-lstm. *Procedia Computer Science*, *179*, 480–487. https://doi.org/10.1016/j.procs.2021.01.031

Dimensions [Accessed: 2024]. (2024). https://app.dimensions.ai

ECB. (2021). Update on economic, financial and monetary developments. *ECB Economic Bulletin*, (5).

Elsaraiti, M., & Merabet, A. (2021). A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed. *Energies*, *14*(20), 6782. https://doi.org/10.3390/en14206782

Gavilanes, R. S. (2022, March). *Univariate time series forecasting: Comparing arima & lstm neural network to the random walk benchmark for exchange rates* [Master's thesis, Master of Science in Data Analytics for Business].

Goncalves, J. V. M., Alexandre, M., & Lima, G. T. (2023, November). *ARIMA and LSTM: A Comparative Analysis of Financial Time Series Forecasting* (Working Papers, Department of Economics No. 202313). University of São Paulo (FEA-USP). https://ideas.repec.org/p/spa/wpaper/2023wpecon13.html

Hamiane, S., Khalifi, H., Ghanou, Y., & Casalino, G. (2023). Forecasting the gross domestic product using lstm and arima. *IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 1–6. https://doi.org/10.1109/ICTMOD59086.2023.10438159

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

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

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Inc., P. T. (2015). *Collaborative data science*. https://plot.ly

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of arima vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, *15*(8). https://doi.org/10.3390/fi15080255

Kumar, B., Sunil, & Yadav, N. (2023). A novel hybrid model combining sarma and lstm for time series forecasting. *Applied Soft Computing*, *134*, 110019. https://doi.org/https://doi.org/10.1016/j.asoc.2023.110019

Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., & Király, F. J. (2019). Sktime: A unified interface for machine learning with time series.

The m4 competition: 100,000 time series and 61 forecasting methods [M4 Competition]. (2020). *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.04.014

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing prophet and deep learning to arima in forecasting wholesale food prices. *Forecasting*, *3*(3), 644–662. https://doi.org/10.3390/forecast3030040

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). Kerastuner.

OpenAI. (2023). Gpt-3.5: Generative pre-trained transformer 3.5 [Available at https://www.openai.com].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E.

(2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pescatori, A., & Zaman, S. (2011). Macroeconomic models, forecasting, and policymaking. *Economic Commentary*, (2011-19).

Peter, Ď., & Silvia, P. (2012). Arima vs. arimax–which approach is better to analyze and forecast macroeconomic time series. *Proceedings of 30th international conference mathematical methods in economics*, *2*, 136–140.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of arima and lstm in forecasting time series. *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, 1394–1401.

Vafin, A. (2020). Forecasting macroeconomic indicators for seven major economies using the arima model. *Sage Science Economic Reviews*, *3*(1), 1–16.

Wabomba, M. S., Mutwiri, M. P., & Fredrick, M. (2016). Modeling and forecasting kenyan gdp using autoregressive integrated moving average (arima) models. *Science Journal of Applied Mathematics and Statistics*, *4*(2), 64–73. https://doi.org/10.11648/j.sjams.20160402.18

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wu, J. M.-T., Tsai, M.-H., Cheng, C.-C., & Wu, M.-E. (2022). Predict the trend of economic indicators in time series based on recurrent neural network combined with leading indicators. *Journal of Intelligent and Fuzzy Systems*, *43*(2), 2179–2189.

Wu, L. S.-Y., Hosking, J. R. M., & Ravishanker, N. (2018). Reallocation Outliers in Time Series. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *42*(2), 301–313. https://doi.org/10.2307/2986234

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019). A comparison between arima, lstm, and gru for time series forecasting. *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*, 49–55.

Yenilmez, İ., & Mugenzi, F. (2023). Estimation of conventional and innovative models for rwanda's gdp per capita: A comparative analysis of artificial neural networks and box–jenkins methodologies. *Scientific African*, *22*, e01902.

Yurtsever, M. (2023a). Unemployment rate forecasting: Lstm-gru hybrid approach. *Journal for Labour Market Research*, *57*, 18. https://doi.org/10.1186/s12651-023-00345-8

Yurtsever, M. (2023b). Unemployment rate forecasting: Lstm-gru hybrid approach. *Journal for Labour Market Research*, *57*(1), 1–9.

APPENDIX A

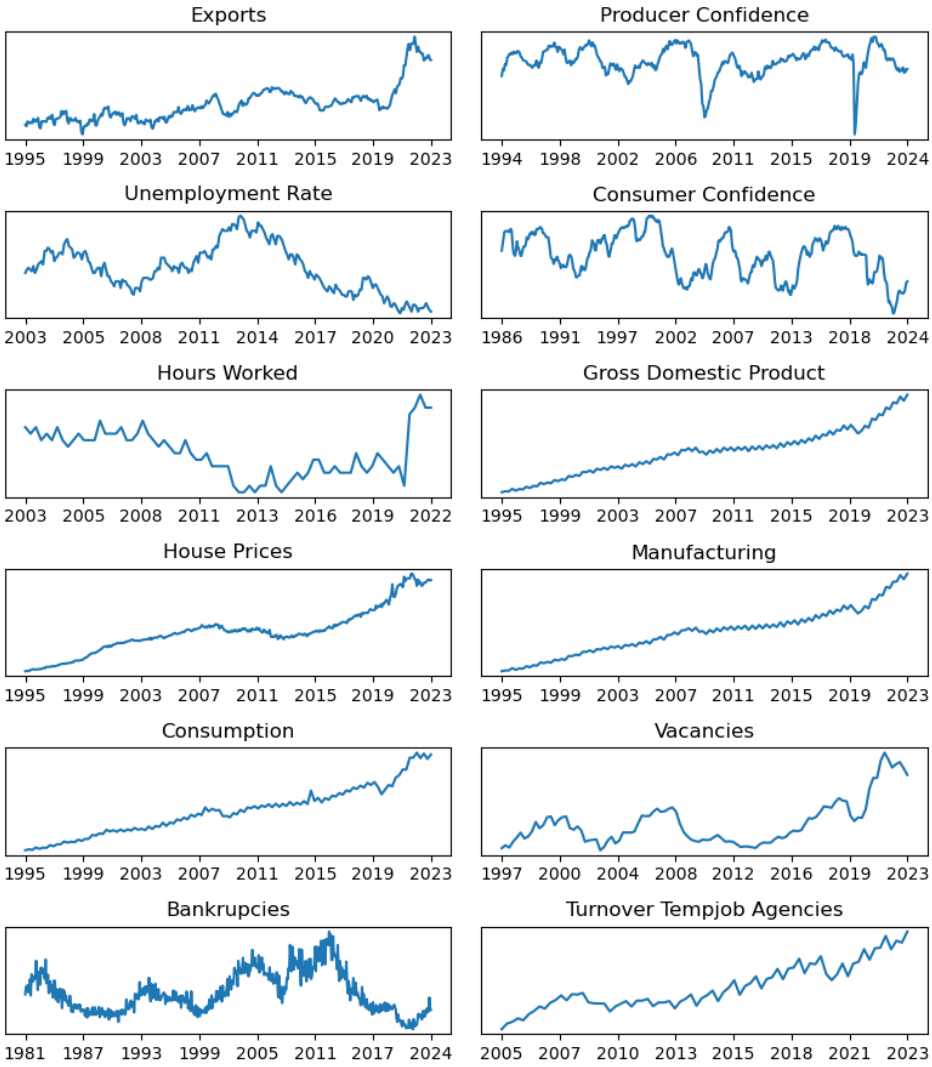

Figure 15: Indicator plots.

APPENDIX B

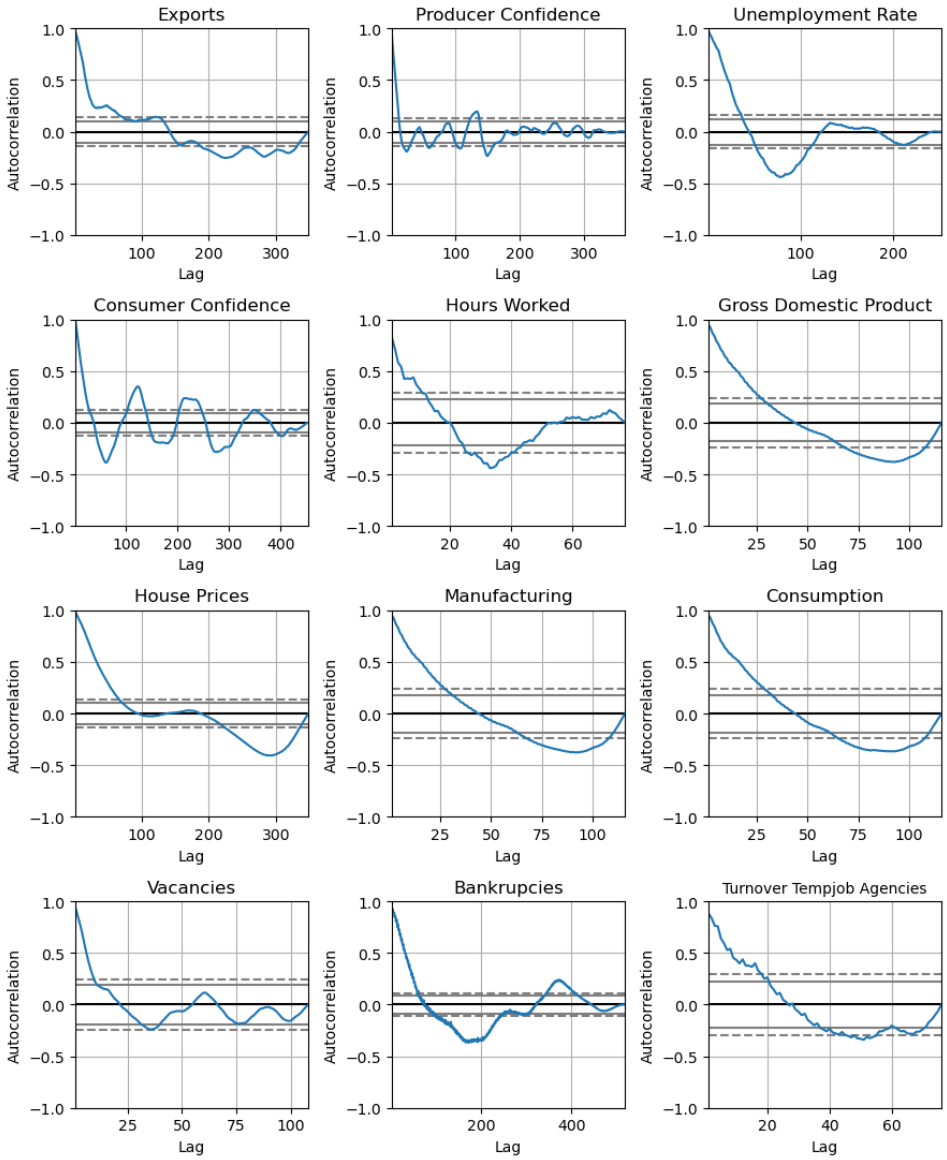

Figure 16: Autocorrelation Function plots.

APPENDIX C

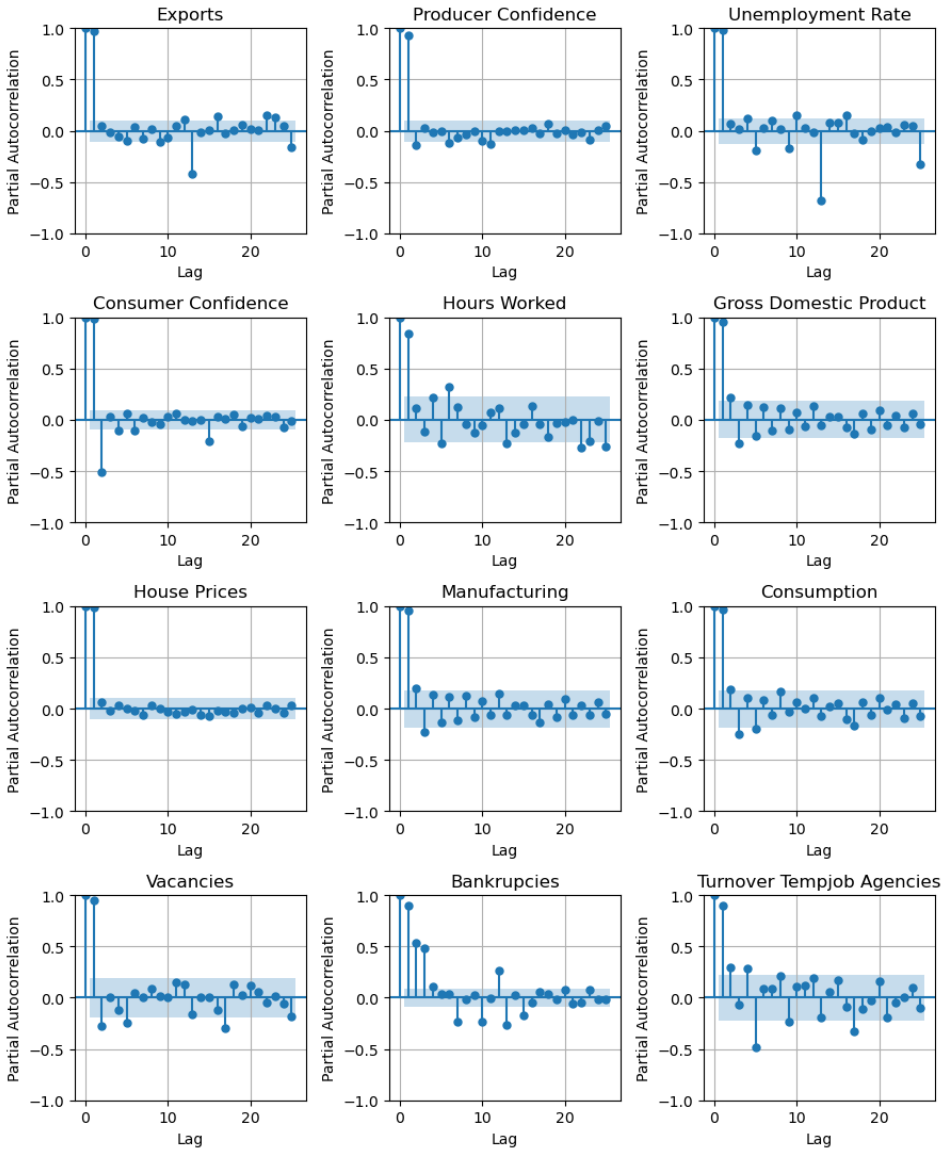

Figure 17: Partial Autocorrelation Function plots.

APPENDIX D

Table 5: Best Hyperparameters for the Deep Learning Models

	Model	Units	Dropout	Learning Rate	Activation	Val Loss
1	LSTM	32	0.2	0.01	ReLu	0.0164
	GRU	104	0.04	0.005	Sigmoid	0.0188
2	LSTM	68	0.04	0.01	Tanh	0.0368
	GRU	60	0.14	0.01	Sigmoid	0.0414
3	LSTM	112	0.16	0.01	Sigmoid	0.0239
	GRU	64	0.12	0.005	Sigmoid	0.0245
4	LSTM	72	0.04	0.005	Sigmoid	0.0328
	GRU	68	0	0.005	ReLu	0.0334
5	LSTM	76	0.02	0.01	ReLu	0.2067
	GRU	108	0.1	0.01	Sigmoid	0.213
6	LSTM	72	0.12	0.001	ReLu	0.003
	GRU	104	0.02	0.01	Tanh	0.0048
7	LSTM	116	0.2	0.005	Sigmoid	0.003
	GRU	124	0.08	0.005	ReLu	0.0031
8	LSTM	32	0.14	0.005	ReLu	0.0029
	GRU	36	0.08	0.001	ReLu	0.0028
9	LSTM	72	0.24	0.001	ReLu	0.016
	GRU	80	0	0.005	ReLu	0.0193
10	LSTM	104	0.16	0.005	Sigmoid	0.017
	GRU	32	0.18	0.01	Tanh	0.015
11	LSTM	96	0.08	0.001	ReLu	0.1192
	GRU	68	0.02	0.01	Sigmoid	0.1117
12	LSTM	108	0.06	0.01	Sigmoid	0.0348
	GRU	56	0.02	0.005	ReLu	0.0457