6 CAN THE CLASSICAL ECONOMIC MODEL IMPROVE THE PERFORMANCE OF DEEP LEARNING? A GDP FORECASTING EXAMPLE

Taoxiong Liu[1]
Huolan Cheng[1,*]

Abstract

Model ensemble is considered as a powerful tool to deal with the overfitting to train data when Deep Learning (DL) models is applied to small size sample. With the application to GDP forecasting, we find significant overfitting to the validation set which also limit the power of model ensemble. We propose the Filtering Ensemble Method (FEM) which use the Classical Economic Model (CEM) as a benchmark to filter overfitted DL models. Results show that the FEM successfully improves the performance of DL models, and the Two-step Prediction Method (TSPM) further enhances their accuracy. Besides, regression equations confirm the possibility of overfitting of DL models on validation sets and the effectiveness of CEMs in filtering overfitted DL models. The study highlights the importance of combining DL models with CEMs in macroeconomic forecasting and suggests that incorporating economic knowledge is critical for the successful application of DL models in economics.

Keywords: GDP forecasting, Deep Learning, Attention, LSTM, ARIMA, VAR

JEL: C45, C53, E27

1. Introduction

The advent of Machine Learning (ML) and Deep Learning (DL) has been widely recognized as a powerful technology that can be applied to economic forecasting (Varian, 2014; Bajari et al., 2015; Chalfin et al. 2016; Balla et al. 2022). In recent years, there has been a proliferation of literature on economic forecasting using ML or DL (Krauss, Do & Huck, 2017;Siami-Namini, Tavakoli & Siami-Namin, 2018; D. Vrontos, Galakis & D. Vrontos, 2021;Richardson, Mulder & Vehbi, 2021; Barkan *et al*., 2023). However, the effectiveness of these methods in delivering the promised improvements in forecast accuracy remains uncertain. A recent literature review by Petropoulos et al. (2022) raised questions regarding whether big data techniques would significantly enhance the accuracy of macroeconomic forecasts. In this paper, we focus on GDP forecasting, and investigate ways to effectively enhance the DL's forecasting ability with the help of Classical Economic Model（CEM）.

1 Institute of Economics, School of Social Sciences, Tsinghua University, Beijing, 100084, China. E-mail addresses: liutx@tsinghua.edu.cn (T. Liu), chl20@mails.tsinghua.edu.cn (H. Cheng).
** Corresponding author*

DL models possess the ability to represent complex functions, theoretically enabling them to express almost any function by adding more units and layers. However, due to their inherent complexity, training DL models is a formidable task that necessitates an enormous amount of data, which is unfeasible for macroeconomic forecasting. Traditional macroeconomic variables are usually quarterly or monthly, with sample sizes in the hundreds. Although high-frequency data can be employed in the age of big data, the nature of small samples cannot be altered when conventional macroeconomic variables such as GDP, Consumer Price Index (CPI), and Fixed Asset Investment are the predicted objects. When applying complicated DL models to macroeconomic prediction problems, overfitting is an inescapable problem, wherein DL models perform well in-sample but poorly out-of-sample. Additionally, DL models are susceptible to being trapped in a local optimum, because they are typically trained by stochastic gradient descent, where initial values for parameters are assigned randomly, and the optimization problem is usually not convex. To alleviate overfitting, regularization and early stopping have been proposed, and the ensemble of many models also mitigates the problem of local optimum. However, are these measures adequate to ensure optimal performance of DL models in macroeconomic forecasting? The answer, we believe, is negative. These methods primarily focus on alleviating the overfitting to the training data and utilize the model performance on the validation set to gauge the degree of overfitting. Nonetheless, to the best of our knowledge, no literature has explicitly examined the problem of overfitting the validation data, which is intuitively conceivable when the validation set is repeatedly used to evaluate the model. While the model ensemble, necessitating the training of hundreds of models is commonly considered as a powerful way to fight against overfitting and local optimum, it may simultaneously exacerbate the overfitting to the validation data. For training the DL model with a small sample, usually hyperparameter values are initiated stochastically and the training is halted once the performance in the validation set meets a predetermined criterion, only those initial hyperparameter values and corresponding models that fit the validation set moderately well remain. If the validation set is used multiple times, and its size is small, the remaining models will include those that occasionally fit the validation set well but might deviate significantly from the true model. Hence, there is a risk of overfitting in the validation sample when utilizing DL models in macroeconomic forecasting, which may explain the unstable and unsatisfactory improvements in forecast accuracy. Addressing overfitting in the validation set is, therefore, a crucial research problem.

In this paper, we find that the overfitting to the validation set does matter while we train and assemble lots of DLs for the GDP forecasting. we present a novel approach to mitigate the overfitting to the validation data when applying DL models to macroeconomic forecasting. We employ the CEM as a benchmark for model selection. Our rationale is that the good DL model should not deviate significantly from the CEM, which has been verified by extensive economic research. Thus, for the models performing fairly well on the validation set, the greater the deviation of the DL model from the CEM, the higher the probability of the overfitting to the validation data. Drawing on the literature review, we adopt widely used CEMs in GDP forecasting, such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregressive (VAR), as the benchmarks. We propose a Filtered Ensemble Method (FEM) that uses the benchmark model to filter the DL models for ensemble. As a comparison, we also assemble DL models by simple averaging. We find that the FEM method produces lower forecasting errors than the simple averaging method. Furthermore, given that a hybrid model has been shown to improve forecasting accuracy in previous studies (Zhang, 2003; Cadenas & Rivera, 2010; Babu & Reddy, 2014; Liu & Xu, 2015), we explore the applicability of FEM in a hybrid model. The hybrid model involves using a CEM to forecast GDP, followed by using a DL model to investigate the remaining residual, which is called as Two-step Prediction Method (TSPM). The results show that the hybrid model with FEM has the highest forecasting accuracy, compared to CEMs, DL models by simple averaging, and DL models with FEM. Lastly, we demonstrate the existence of overfitting on the validation set by regression, which is the main source of the improvement of the FEM method.

The subsequent section of this paper is structured as follows. The second part provides an introduction to the frequently used CEMs for GDP forecasting, as well as the commonly employed DL models in time series. The third part of this paper introduces the research idea and research scheme employed. The empirical results and analysis are then presented in the fourth part, which compares the forecasting effects of various models and highlights the importance of CEMs. Finally, the fifth part summarizes the research findings.

2. CEMs for the GDP and DL Models in Time Series

2.1. CEMs for the GDP

The GDP forecasting is a prominent area of research in economics, which has a long history dating back to Tinbergen (1939, 1974) and Klein (1970), who developed large macroeconometric models based on Keynesian theory. These models rely on national economic accounting procedures and estimate linear equations for each sector (e.g., consumption, investment, and exports), which are then combined with the constant national income equation. However, the forecasting ability of large macroeconometric models was doubted, debated, and criticized after the oil crisis in the 1970s, with Lucas' criticism being the most famous. In response, economists began to construct more rigorous econometric models to examine economic time series. Traditional GDP forecasting is divided into two approaches. The first approach involves building time series models, such as Autoregressive (AR), ARIMA and VAR, to forecast the GDP (Boero, 1990; Salazar & Weale, 1999; Fukuda, 2007; Artis & Okubo, 2010; Kim & Swanson, 2018; Kuck & Schweikert, 2021). The second approach involves constructing linkage equations for various economic sectors in an economy, such as Real Business Cycle (RBC) and Dynamic Stochastic General Equilibrium (DSGE) models (Kydland & Prescott, 1982; Červená & Schneider, 2014; Smets, Warne & Wouters, 2014; Fair, 2019; Yang, 2020; Chin, 2022). Despite their differences, both approaches seek to provide accurate GDP forecasts by utilizing economic theory and empirical data.

Comparisons of the out-of-sample prediction effects of the two types of models have been conducted by several researchers (Adolfson, Lindé & Villani, 2007; Adolfson et al., 2007; Christoffel,Coenen & Warne, 2008, 2011; Rubaszek & Skrzypczynski, 2008; Liu, Gupta & Schaling, 2009; Kolasa, Rubaszek & Skrzypczynski, 2012;Edge, Kiley & Laforte, 2010; Wolters, 2011; Edge & Gürkaynak, 2010; Del Negro & Schorfheide, 2013; Wieland & Wolters, 2011). The results show that the predictive power of DSGE models is comparable to or slightly better than that of VARs, but not significantly different from that of simple univariate time series models such as ARIMA models. Consequently, univariate models such as ARIMA and AR are often used as reference models for comparison with developing models, while the DSGE model is not employed. Some researchers, for instance, have compared ARIMA models to DL algorithms (Siami-Namini, Tavakoli & Siami-Namin, 2018; Song et al. 2020; Weytjens Lohmann &Kleinsteuber, 2021). Moreover, considering the complex relationship of economic variables, multivariate models remain a popular forecasting model, and VAR is frequently used as the benchmark model (Thomakos & Guerard, 2004; Kuzin, Marcellino & Schumacher, 2011;Tallman & Zaman , 2020; Barkan et al., 2023).

2.2. DL Models in Time Series

In recent years, there has been a growing trend in the use of DL models in finance, while their application in macroeconomics is still limited. Among the sequential models in DL, Recurrent Neural Network (RNN) was the first model to use hidden states to preserve the processing results of previous periods, allowing information to propagate through the loop. As a result, RNNs have

gained popularity in finance research (Roman & Jameel, 1996; Balkin, 1997; Binner, Kendall & Chen, 2004;; Irsoy & Cardie, 2014; Dixon, 2018). However, RNNs do not support long-term memory, and gradient explosion or disappearance may occur when the sequence is too long. To address this limitation, Hochreiter & Schmidhuber (1997) proposed the Long Short-term Memory (LSTM) model, which uses forget gates to selectively remember information, preventing gradient explosion or disappearance. Gate Recurrent Unit (GRU) is a variant of LSTM, which is more simplified and converges faster than LSTM, but at the expense of accuracy. Comparing core models, Shewalkar (2018) found that LSTM outperformed RNN and GRU. LSTM is now widely used in the economic and financial fields. The Attention mechanism, which allows a model to selectively focus on important features or parts of the input data, has exploded in recent years. Vaswani et al. (2017) said Attention is all you need! However, Attention has not been extensively applied in the economic and financial domains. LSTM combined with Attention holds high potential for applications in the economic and financial domains.

In summary, in the field of macroeconomic forecasting, simple univariate time series models, such as ARIMA, have demonstrated their durability and stability over time. And due to the influence of various factors such as consumption, investment, and exports on the economic system, multivariate VAR models have become a popular choice for macroeconomic forecasting, too. Nevertheless, with the increasing complexity of the current economic environment, traditional economic models have struggled to meet the high standards set by academics for GDP forecasting accuracy. Recently, DL models such as LSTM and Attention mechanisms have shown great promise in the field of financial forecasting. These models have the potential to revolutionize macroeconomic forecasting of GDP. However, there is a significant gap in research on the use of DL models for macroeconomic forecasting. Therefore, further exploration of the application of DL models in macroeconomic forecasting is necessary to identify their potential strengths and limitations.

3. Research Ideas and Schemes

DL models have demonstrated remarkable success in various domains owing to their flexibility and predictive power that enables them to learn from experience and data. However, their effectiveness in macroeconomic forecasting is limited due to the restricted availability of data samples in this field. In this paper, we aim to identify the potential reasons underlying the poor performance of DL models in macroeconomic forecasting and suggest strategies to enhance their forecasting abilities.

3.1. Research Ideas

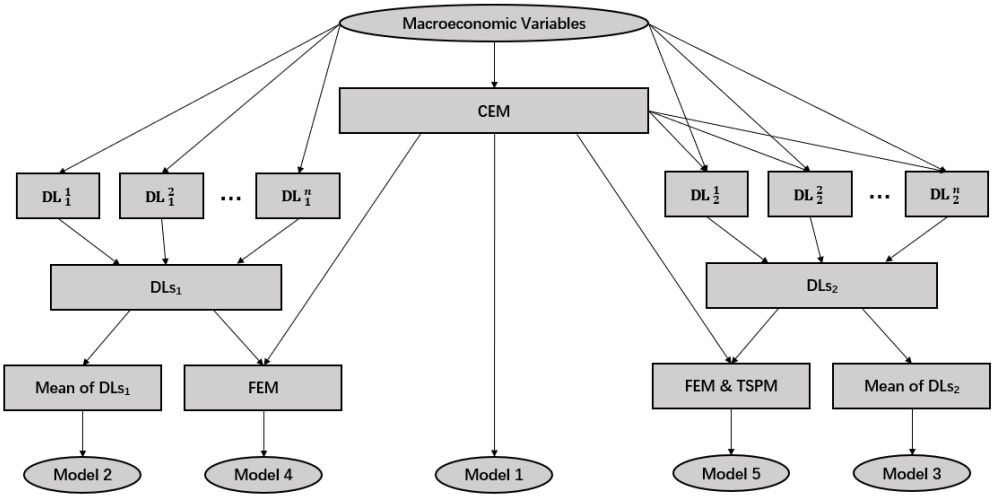
In macroeconomic forecasting, the limited sample size poses a challenge to the performance of DL models, as up to hundreds of thousands of parameters must be estimated with only a small sample of available data. This leads to the risk of overfitting the sample. Furthermore, the stochastic assignment of initial parameter values in DL models can lead to a local optimum problem in small data. While regularization, early stopping, dropout techniques, and assembling are commonly used to mitigate overfitting and local optimum problems, they may not be sufficient to guarantee good performance of DL models in macroeconomic forecasting. Existing overfitting mitigation strategies typically focus on the training set, without evaluating whether overfitting also occurs on the validation set. Since the overuse of information on the training set can result in overfitting on the training set, then overuse of information on the validation set can also result in overfitting on the validation set. In cases where parameter tuning and model evaluation rely solely on the validation set, such as when hyperparameters and the training times are determined based on the model's performance on the validation set, the retained model may only perform well on the validation set and not generalize to new data. Additionally, the use of ensemble approaches can amplify overfitting on the validation set, further limiting the generalization abilities of DL

models. Therefore, effectively using DL models in macroeconomic forecasting requires careful consideration of overfitting problems on both the training and validation sets.

To mitigate overfitting in the training set, techniques such as early stopping and regularization are commonly used. However, for overfitting in the validation set, we propose a novel approach called FEM, which utilizes CEM as a benchmark to filter DL models. The rationale behind this approach is that CEMs are grounded in macroeconomic theories and have been tested by real-world data over many years, thus providing a reflection of the actual operation of the economy. Therefore, a DL model with good fitting capacity should perform similarly to CEM on the validation set, and the significant deviation from CEM may indicate the risk of overfitting. To assess the efficacy of FEM, we design experiments to compare the performance of DL models with and without FEM.

Fig1 illustrates the experimental design, in which there are five models. (The acronyms for these models are listed in Table 1 which we will explain later.) The CEMs serve as the benchmark model, denoted as Model 1. For comparison purposes, we also utilize a simple average ensemble DL model, denoted as Model 2. In addition, we propose a DL model with FEM, denoted as Model 4, to mitigate the overfitting problem in the validation set. Prior research suggests that hybrid models can improve the forecasting ability of models. Therefore, we apply the FEM to the TSPM, which involves using a CEM to predict GDP, using a DL model to predict its residual, and finally assembling DL models using FEM. The DL model with FEM and TSPM is named as Model 5, while the model for comparison, the ensemble of DL models with TSPM, is denoted as Model 3.

Figure 1. Model flow



3.1.1. Filtering Rules for the FEM

The FEM approach for filtering DL models relies on determining the proximity of the DL model prediction results to the benchmark model prediction results and the true predicted target. To achieve this, the widely used Euclidean distance metric is utilized, with its formula presented below.

$$d(x,y) = \sqrt{\sum_{j=1}^{m}(x_j - y_j)^2} \quad (1)$$

where m denotes the sample size, and x and y denote the two groups of data to be assessed.

As the focus of our study is on the overfitting on the validation set, we restrict ourselves to utilizing only the validation set for our analysis. We begin by calculating the Euclidean distance between the prediction results of DL models and CEM, as well as those between DL models and the prediction target. The resulting distance sequences are then used to filter DL models, before they are integrated, based on the quantiles where lower quantiles indicate closer distances. For example, we could filter DL models based on their distance to the CEM falling within the [80% quantile, 100% quantile] interval and their distance to the prediction target falling within the [0% quantile, 20% quantile] interval. The final prediction is obtained by taking the ensemble of all retained DL models. In total, we choose four thresholds for each model utilizing FEM. They are:

Threshold 1, the distance to the CEM falling within [20% quantile, 100% quantile] and the distance to the target falling within [0% quantile, 20% quantile]

Threshold 2, the distance to the CEM falling within [20% quantile, 100% quantile] and the distance to the target falling within [0% quantile, 40% quantile].

Threshold 3, the distance to the CEM falling within [80% quantile, 100% quantile] and the distance to the target falling within [0% quantile, 20% quantile].

Threshold 4: the distance to the CEM falling within [80% quantile, 100% quantile] and the distance to the target falling within [0% quantile, 40% quantile].

3.1.2. Model 1 – The Benchmark Model

$$\hat{y}_t^1 = f(x_{t-1}, y_{t-1}) \tag{2}$$

Where y is the predicted target and x is the explanatory variable, and we just use $t-1$ to refer to lagging information. In this research, we utilize the ARIMA and VAR models as benchmark models. The forecasting formula for all models is expressed as $f(x_{t-1}, y_{t-1})$. And if the benchmark model is ARIMA, there is no explanatory variable. Then the forecasting formula is simplified to $f(y_{t-1})$. The benchmark model is used for comparison against the next four models.

3.1.3. Model 2 – The General DL Model

$$\hat{y}_{t,i}^2 = h(x_{t-1}, y_{t-1}), i = 1,2, \dots, n \tag{3}$$

$$\hat{y}_t^2 = \sum_{i=1}^{n} \hat{y}_{t,i}^2 / n \tag{4}$$

This model represents a general DL approach. The forecasting formula is denoted as $h(x_{t-1}, y_{t-1})$. The inputs of explanatory variables are consistent with Model 1, and the DL model is univariate if Model 1 is univariate, whose forecasting formula is simplified to $h(y_{t-1})$. And each $\hat{y}_{t,i}^2$ represent the prediction value obtained by the i'th DL model. To obtain n DL models for ensemble, we initialize the DL models with different starting parameter values. Model 2 does not employ the FEM, and the final prediction value is the average of n DL models.

3.1.4. Model 3 – The TSPM Model

$$\hat{\varepsilon}_{t,i}^2 = h(\hat{\varepsilon}_{t-1}^x, \hat{\varepsilon}_{t-1}^1), i = 1,2, \dots, n \tag{5}$$

$$\hat{y}_t^3 = \sum_{i=1}^{n} \hat{\varepsilon}_{t,i}^2 / n + \hat{y}_t^1 \tag{6}$$

TSPM involves using the CEM to predict the target variable firstly and using DL model to predict the CEM's residual. Thus Model 3 using the DL model to predict the residual of Model 1, represented by formula (5), where $\hat{\varepsilon}_{t-1}^x$ and $\hat{\varepsilon}_{t-1}^1$ are residuals of explanatory variables and target

variable in Model 1. And if the CEM is ARIMA, there is no residual of explanatory variable. Then the formula (5) is simplified to $\hat{\varepsilon}_{t,i}^{2} = h(\hat{\varepsilon}_{t-1}^{1})$. Like Model 2, n DL models are trained and utilized to generate the final forecast. And this model incorporates the TSPM but not the FEM. The ultimate prediction is the average of n DL models, plus the forecast produced by Model 1.

3.1.5. Model 4 – The DL Model Using the FEM

$$\hat{y}_t^4 = \sum_{i \in J} \hat{y}_{t,i}^2 * w_i \tag{7}$$

This model applies the FEM approach to filter out DL models that are likely to suffer from overfitting from all DL models in Model 2. It only retains DL models that are moderately fitted for integration, where J is the set of retained models, and w_i represents the weight assigned to each model. w_i is calculated according to residuals of i'th DL model. Specifically, w_i is the inverse of mean square error of i'th DL model.

3.1.6. Model 5 – The TSPM Model Using FEM

$$\hat{y}_t^5 = \sum_{i \in J} \hat{\varepsilon}_{t,i}^2 * w_i + \hat{y}_t^1 \tag{8}$$

This model applies the FEM to filter out DL models that are likely to suffer from overfitting from all DL models in Model 3. It only retains DL models that are moderately fitted for integration. The ultimate prediction is the weighted average of DL models, plus the forecast produced by Model 1.

3.2. Model Description

This section presents an overview of the models utilized for GDP forecasting. These models include both CEMs and DL models. To represent CEMs, we selected ARIMA and VAR. On the other hand, LSTM and Attention were chosen as the example DL models.

3.2.1. ARIMA and VAR

The forecasting formulas for ARIMA and VAR models are provided in Equations (9) and (10), respectively. The ARIMA and VAR models are constructed using the auto.arima and VAR functions in the R programming language, respectively. The lag order is selected automatically based on the information criterion.

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d y_t = c + \left(1 + \sum_{j=1}^{q} \theta_j L^j\right)\varepsilon_t \tag{9}$$

$$\begin{bmatrix} y_t \\ X_t \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \pi_{12,1} \\ \phi_{21,1} & \pi_{22,1} \end{bmatrix} \begin{bmatrix} \phi_{t-1} \\ \phi_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11,p} & \phi_{12,p} \\ \phi_{21,p} & \phi_{22,p} \end{bmatrix} \begin{bmatrix} y_{t-p} \\ X_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{2t} \\ \varepsilon_{2t} \end{bmatrix} \tag{10}$$

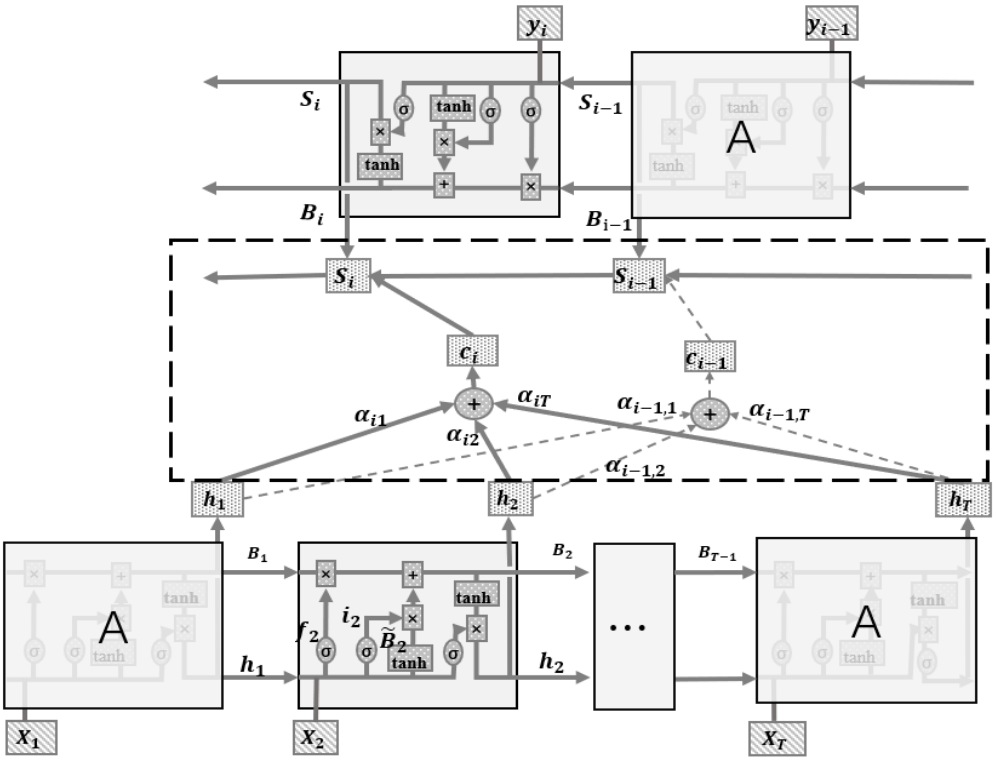
where p is the number of autoregressive terms, d is the number of differences, q is the number of moving average terms, y_t and X_t are time series and X_t may be matrix in VAR, ϕ_i is an autoregressive coefficient and may be matrix in VAR, L is a lag operator, c is an arbitrary constant, θ is a moving average coefficient, and ε_t is white noise.

3.2.2. LSTM and Attention

LSTM is a commonly used DL model that employs RNN for long-term memory and forgetting gates for short-term memory. However, in long time series, the influence of input at different time points on the prediction results may differ. In order to assess the impact of various inputs on output results, Treisman & Gelade (1980) introduced the Attention mechanism, which emulates human brain adaptation, assigns attention weights to different inputs, and highlights the

contribution of each input to the output. Attention did not gain much attention until 2014, when the Google Deepmind team applied it to image classification. Bahdanau, Cho & Bengio (2015) first applied it to machine translation tasks to achieve simultaneous translation and alignment. In 2017, the Google machine translation team extensively used Self-Attention to learn text representation. All these applications have achieved remarkable results. However, the application of the Attention mechanism in macroeconomic forecasting has been limited. In this paper, we incorporate the Attention mechanism into LSTM to forecast GDP. As shown in Figure 2, we use two inputs, X and y, where X represents the matrix of explanatory variables and y is the forecasting target. We construct an LSTM model for X and another for y, shown in the lower and upper parts of Figure 2, respectively. The middle part with the dotted wire frame represents the Attention mechanism.

Figure 2. LSTM and Attention



LSTM is a type of neural network that has three gates: the forget gate, the input gate, and the output gate. The forget gate determines which information to discard, as specified in Equation (11). The input gate updates the new memory based on the input data, as explained by Equations (12) and (13). The output gate combines both the short-term and long-term memory to produce the output h_t, represented by Equation (15). In the Attention mechanism, the main objective is to obtain the weight α_{it}, which represents the probability that the target variable y_i is aligned with the explanatory variables X_t, as demonstrated in Equations (17)-(18). Subsequently, we calculate the context vector c_i, which is presented in Equation (19), as an input to the hidden state of y_i. Finally, we feed the hidden state s_i into the subsequent network to predict y, as shown in Equation (21).

$$f_t = W_f \cdot [h_{t-1}, X_t] + b_f \tag{11}$$

$$i_t = W_i \cdot [h_{t-1}, X_t] + b_i \tag{12}$$

$$\tilde{B}_t = tanh(W_B \cdot [h_{t-1}, X_t] + b_B) \tag{13}$$

$$B_t = f_t * B_{t-1} + i_t * \tilde{B}_t \tag{14}$$

$$h_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) * \tanh(B_t) \tag{15}$$

$$tanh(x) = (e^x - e^{-x})/(e^x + e^{-x}) \tag{16}$$

$$e_{it} = a(s_{i-1}, h_t) \tag{17}$$

$$\alpha_{it} = exp(e_{it}) / \sum_{k=1}^{T} exp(e_{ik}) \tag{18}$$

$$c_i = \sum_{t=1}^{T} \alpha_{it} h_t \tag{19}$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{20}$$

$$\hat{y}_i = g(\hat{y}_{i-1}, s_i, c_i) \tag{21}$$

where h_t represents the hide state for X_t, B_t represents the cell state which constitutes the main line of events, running directly over the entire link with only a few linear interactions. $W.$ and $b.$ are the weight and bias of the current input, respectively. s_i is the hidden state for y_i. e_{it} is called as $\alpha_{it}{}'s$ associated energy. c_i is the context vector. $\sigma(\cdot)$ denotes the activation function, and the sigmoid function is general chosen.

3.3. Evaluation Index

In this study, three methods are employed for GDP prediction: moving window, cumulative window, and static window. In the moving window approach, the length of the training set observations remains constant, and the training sets shift one period after each prediction period. In the cumulative window approach, the training sets gradually increase, and all prior samples are included as the training set for the subsequent prediction period. In the static window approach, the model is trained only once, and all remaining periods are predicted. To evaluate the performance of the forecasting models, the mean squared error (MSE) is used as the measure of accuracy. The MSE is calculated as follows:

$$\mathrm{MSE} = \sum_{\mathrm{t}=1}^{\mathrm{N}} (\hat{\mathrm{y}}_{\mathrm{t}} - \mathrm{y}_{\mathrm{t}})^2 / \mathrm{N} \tag{22}$$

where N denotes the number of samples being predicted, and $\hat{\mathrm{y}}_{\mathrm{t}}$ and y_{t} denote the estimated and true values of the prediction, respectively.

To avoid potential selection bias, relying solely on MSE to evaluate models may not be sufficient. Therefore, this paper incorporates the Diebold-Mariano test (DM test), a hypothesis testing method, to supplement the model evaluation process. The DM test is utilized to compare two forecasting models of a time series and to determine which model is superior. The null hypothesis

of the DM test is that there is no difference between the two models, and the statistical formula is presented as follows:

$$DM = N^{\frac{1}{2}} \sum_{t=1}^{N} d_t / (N\sigma_d) \tag{23}$$

$$d_t = g(e_{1t}) - g(e_{2t}) \tag{24}$$

g is the loss function of interest, e.g., the quadratic loss $g(e) = e^2$ or the absolute loss $g(e) = |e|$, e_1 and e_2 are the errors from the two competing forecasts, and the σ_d^2 is the variance of d.

3.4. Short Names

Table 1 provides a comprehensive list of abbreviations used throughout this research paper, together with their corresponding full names. These abbreviations are frequently used in the analysis and discussion of the results, and providing their full names will help readers understand their meaning without confusion or ambiguity.

Table 1 Short Names in the Article

Abbreviation	Full Name	Abbreviation	Full Name
AR	Autoregressive	TSPM	Two-step Prediction Method
ARIMA	Autoregressive Integrated Moving Average	u	Unemployment Rate
CEIC	Comprehensive Economic Information Corporation	VAR	Vector Autoregressive
CEM	Classical Economic Model	w	Wage
CPI	Consumer Price Index	M1	Model 1
DL	Deep Learning	M2	Model 2
DM test	Diebold-Mariano test	M3	Model 3
DSGE	Dynamic Stochastic General Equilibrium	M4.1	Model 4 using Threshold 1
Exp	Export	M4.2	Model 4 using Threshold 2
FEM	Filtering Ensemble Method	M4.3	Model 4 using Threshold 3
GC	Government Consumption	M4.4	Model 4 using Threshold 4
GDP	Gross Domestic Product	M5.1	Model 5 using Threshold 1
GRU	Gate Recurrent Unit	M5.2	Model 5 using Threshold 2
Imp	Import	M5.3	Model 5 using Threshold 3

Inv	Investment	M5.4	Model 5 using Threshold 4
LASSO	Least Absolute Shrinkage and Selection Operator	A	M2 vs. M4.1
LSTM	Long Short-term Memory	B	M2 vs. M4.2
M	Money supply	C	M2 vs. M4.3
ML	Machine Learning	D	M2 vs. M4.4
MSE	Mean Squared Error	E	M3 vs. M5.1
n	Population Growth Rate	F	M3 vs. M5.2
PC	Private Consumption	G	M3 vs. M5.3
r	Interest Rate	H	M3 vs. M5.4
RBC	Real Business Cycle	I	M1 vs. M2
RNN	Recurrent Neural Network	J	M1 vs. M3
s	Saving rate	K	M2 vs. M3

4. Empirical Results & Analysis

4.1. Data

The aim of this study is to forecast the GDP growth rate for ten selected representative economies with bigger sample size, namely the United States, France, the United Kingdom, Japan, Canada, Australia, Germany, South Korea, South Africa, and Taiwan, China. The selection of explanatory variables is based on GDP accounting and macroeconomic theory. Specifically, we choose five indicators, namely private consumption (PC), government consumption (GC), investment (Inv), export (Exp), and import (Imp), according to the expenditure method of calculating GDP. Additionally, we supplement the selection of variables with those from the IS-LM model, Solow model, Harold-Domar model, Phillips curve, and other economic theories, such as money supply (M), CPI, interest rate (r), savings rate (s), population growth rate (n), unemployment rate (u), and wage (w). We collect all data from the Comprehensive Economic Information Corporation (CEIC) database2, and all variables are seasonal frequency and adjusted as growth rates, ending in the fourth quarter of 2019. The descriptive statistics of some of the variables are presented in Table 2. To ensure stationarity of the variables, we conduct a stationarity test on each of them, and if necessary, perform differencing until they become stationary. Subsequently, we employ the Least Absolute Shrinkage and Selection Operator (LASSO) to select the most important variables from the aforementioned 12 explanatory variables.

Table 2 Descriptive Statistics

Country	Variable	Obs.	Mean	SD	Min	Median	Max
USA	GDP	287	6.42	3.46	-3.47	6.13	19.65
	M	287	6.26	2.84	0.23	6.14	13.51
	CPI	287	3.46	2.88	-2.79	2.83	14.43
	PC	287	-0.02	1.33	-6.16	-0.03	8.08

[2] *https://insights.ceicdata.com/login*

Country	Variable	Obs.	Mean	SD	Min	Median	Max
	GC	287	6.60	5.65	-5.90	6.02	49.70
	Inv	287	6.77	6.09	-15.18	6.75	29.03
	Exp	287	7.73	11.49	-27.25	7.82	44.97
	Imp	287	9.14	11.07	-29.83	8.26	56.12
	s	287	20.52	2.49	13.30	20.6	24.90
	u	287	5.74	1.64	2.57	5.53	10.67
	r	287	3.43	20.2	-48.94	0.00	73.52
	w	287	-0.04	0.94	-7.01	0.03	7.11
	n	287	1.30	0.49	-0.67	1.21	2.86
UK	GDP	256	7.83	4.99	-3.52	6.20	26.89
	M	256	9.29	9.30	-3.24	7.98	74.25
	CPI	256	5.03	4.84	-0.45	3.25	26.57
	PC	256	7.68	4.87	-3.78	6.06	25.54
	GC	256	8.10	6.68	-1.65	7.45	47.03
	Inv	256	8.37	7.45	-14.04	7.49	30.03
	Exp	256	8.64	8.81	-10.00	7.57	39.61
	Imp	256	8.72	10.05	-14.46	7.54	55.46
	s	256	18.62	2.73	12.32	18.93	24.62
	u	256	3.34	13.01	-21.81	2.79	54.10
	r	256	0.21	21.24	-90.05	1.53	96.76
	w	256	45.24	32.54	5.00	41.15	101.10
	n	256	0.42	0.26	-0.12	0.42	0.92

Note: Because of space, descriptive statistics are shown for USA and UK variables only.

4.2. Results and Analysis

The dataset is divided into three parts, namely, the training set, the validation set, and the prediction set, in a 6:2:2 ratio. The training set and validation set are combined to construct the ARIMA and VAR models. In contrast, the DL models are trained on the training set alone and tuned on the validation set. The prediction set is utilized to compare the prediction performance of each model.

4.2.1. Out-of-Sample Prediction Effect of Each Model

The main objective of this study is to investigate whether the use of CEMs can effectively reduce overfitting of DL models on validation sets and thereby reduce the prediction error. If the FEM can reduce the out-of-sample prediction errors of the DL models, it indicates that the FEM has successfully alleviated the problem of overfitting on validation sets. In this research, ARIMA and VAR are used as CEMs, and five models are produced for each CEM. Strictly, there are eleven models for each CEM, since Model 4 and Model 5 have different thresholds leading to different selection rules and predictions. Table 3 presents the MSE of each model. For ease of comparison of prediction effects, the MSE of each model was compared in pairs, and the comparison results are shown in Table 4, which shows the number of economies for which the MSE of the model in the row is smaller than that of the corresponding model in the column, out of ten economies. For example, the number "10" in the first cell of the "Moving" column under "Model 1" column indicates that there are ten economies where the MSE of Model 2 is smaller than that of Model 1 under the

moving window, when ARIMA is the benchmark model. The "Total" column provides the summary of all three forecast windows.

Since Model 4 is obtained by applying the FEM to Model 2, according to Table 4, the performance of the FEM is reflected by the numbers of Model 4 that outperform Model 2, which are marked red. Specifically, when the CEM is ARIMA, these numbers are 27, 26, 25, and 25, accounting for 90.00%, 86.67%, 83.33%, and 83.33%, respectively. When the CEM is VAR, these numbers are 24, 23, 26, and 23, accounting for 80.00%, 76.67%, 86.67%, and 76.67%, respectively. These results demonstrate the strong efficacy of FEM as a filter for DL models, regardless of whether they are univariate or multivariate. Additionally, Table 3 reveals that, in many cases, Model 2 performs worse than Model 1, yet some Model 4 performs better than Model 1. For example, under the cumulative window, the Model 4 with Threshold 3 for Germany, and under the static window, the Model 4 with Threshold 2 for Australia outperform Model 1 when using ARIMA as the benchmark model. Furthermore, under the moving window, the Model 4 with Threshold 1 for Germany, and under the static window, the Model 4 with Thresholds 3 and 4 for the USA outperformed Model 1 when using VAR as the benchmark model. These cases demonstrate the good performance of FEM. To investigate whether FEM is effective when TSPM is used, we compared the performance of Model 5 to Model 3 in Table 4. When the CEM is ARIMA, the four ratios of Model 5 outperforming Model 3 are 83.33%, 83.33%, 86.67%, and 86.67%. When the CEM is VAR, these ratios are 80.00%, 80.00%, 76.67%, and 66.67%. Furthermore, nearly all Model 5s with different thresholds outperform Model 1. These results demonstrate that the FEM can be effective even when the TSPM is used, and combining the FEM with the TSPM allows the DL model to outperform the CEM.

We present the percentages of MSE reduction, which are the averages across ten economies, of Model 4 to Model 2, Model 5 to Model 3, and Model 4 & 5 to Model 1 in Figure 3. When comparing Models 4 to Model 2, we observe that, except for Models 4 using threshold 3 and 4 under moving window when using ARIMA as CEM, all other Model 4s perform better than Model 2, with the highest improvement percentage of 77.61%. Similarly, when comparing Models 5 to Model 3, we find that almost all Models 5s perform better than Model 3, with the highest improvement percentage of 23.62%. Besides, when Models 5 are compared to Model 1, all percentages are bigger than zero, with the highest being 66.5% and the lowest being 7.03%. These results demonstrate the effectiveness of FEM, and the weaker the DL model used in the ensemble are, the greater the improvement percentage of the FEM.

Table 3 Prediction effects (MSE)

CEM		ARIMA										VAR									
Window	Model	USA	UK	France	Japan	Canada	Australia	Germany	SK	SA	TW	USA	UK	France	Japan	Canada	Australia	Germany	SK	SA	TW
Moving	M1	13.50	5.85	8.48	8.61	13.56	15.96	6.43	19.97	31.64	126.48	0.95	**0.88**	1.08	3.56	4.26	2.00	3.07	4.65	1.30	5.03
	M2	3.66	1.93	0.79	3.64	5.69	2.76	2.78	5.08	2.96	13.94	2.08	2.45	3.10	6.27	5.55	6.16	6.02	3.56	1.55	10.46
	M3	2.99	3.40	1.52	**3.04**	4.68	4.70	4.60	6.15	7.77	120.65	0.85	0.88	1.01	3.35	3.64	1.76	2.90	4.46	1.27	3.97
	M4.1	2.65	**1.87**	0.67	3.23	**4.42**	2.75	**2.40**	**4.76**	2.50	11.88	1.28	2.33	2.78	5.45	5.27	5.93	2.99	**2.80**	1.87	9.86
	M4.2	3.07	1.89	0.70	3.20	4.50	**2.71**	2.53	5.50	2.56	12.81	1.36	2.33	2.96	5.89	5.77	5.80	4.69	3.07	1.94	10.13
	M4.3	3.81	2.08	**0.64**	3.22	4.47	4.08	2.46	14.14	**2.15**	**8.94**	0.95	1.15	1.08	3.69	4.26	2.32	3.07	4.65	1.79	5.34
	M4.4	2.98	2.01	0.68	3.22	4.56	2.74	2.63	14.00	2.15	9.88	0.95	2.02	1.08	4.19	6.03	3.04	3.07	4.79	1.80	5.07
	M5.1	1.24	2.48	0.79	3.08	5.29	4.20	2.83	6.68	2.69	108.36	0.84	0.91	0.85	3.13	3.38	**1.72**	2.71	4.22	1.26	3.96
	M5.2	1.32	2.49	0.87	3.10	5.28	4.16	3.55	6.43	3.33	110.96	**0.83**	0.90	0.93	3.22	3.44	1.77	2.79	4.57	**1.26**	3.97
	M5.3	**1.18**	2.48	0.82	3.29	6.04	3.90	2.83	5.88	2.31	109.20	0.87	0.95	**0.78**	**3.06**	3.27	1.72	**2.63**	4.22	1.27	3.97
	M5.4	1.21	2.36	0.82	3.40	6.06	3.87	3.08	5.65	2.25	111.57	0.85	0.93	0.79	3.13	**3.24**	1.84	2.68	4.44	1.28	**3.94**
Cumulative	M1	0.66	**0.93**	0.42	2.96	2.75	2.39	1.29	3.20	1.64	4.80	0.88	0.97	1.08	4.00	4.23	2.08	3.08	4.78	1.29	4.86
	M2	3.84	1.90	0.82	3.37	6.01	2.39	3.03	5.23	2.69	14.21	1.66	1.90	3.33	6.20	6.11	5.41	3.23	3.57	1.51	10.07
	M3	0.57	0.95	0.33	2.56	2.57	2.10	1.13	2.46	1.24	4.46	0.85	0.93	0.96	3.80	3.74	1.82	2.97	4.61	**1.23**	3.72
	M4.1	2.85	1.88	0.70	3.01	4.40	2.29	2.84	5.53	2.32	12.52	1.04	1.96	2.96	5.41	5.43	5.28	2.69	**3.09**	2.02	9.38
	M4.2	3.31	1.92	0.70	3.01	5.21	2.34	2.79	6.06	2.36	13.53	1.14	1.84	3.15	5.67	6.32	5.39	**2.50**	3.19	1.90	9.74
	M4.3	0.66	1.27	0.47	3.14	2.75	2.22	1.24	3.20	1.69	4.80	0.88	0.98	1.08	4.39	4.23	2.77	3.08	4.74	1.49	5.01
	M4.4	0.66	1.33	0.62	3.02	2.76	2.24	1.55	3.20	1.92	11.24	0.88	2.02	1.08	4.43	6.72	3.36	3.03	4.76	1.49	5.02
	M5.1	0.52	0.99	0.32	2.46	**2.41**	1.96	1.10	2.24	1.13	4.01	**0.84**	0.93	**0.85**	3.70	3.45	**1.79**	2.92	4.58	1.29	3.62
	M5.2	0.53	0.97	0.32	2.50	2.49	2.01	**1.09**	2.26	1.13	4.16	0.84	0.94	0.88	3.73	3.52	1.79	2.95	4.52	1.23	3.61
	M5.3	0.52	1.03	**0.32**	**2.39**	2.43	**1.82**	1.10	**2.21**	1.04	**4.01**	0.84	**0.92**	0.85	3.74	**3.35**	1.86	2.88	4.54	1.36	3.53
	M5.4	**0.52**	1.01	0.32	2.43	2.48	1.88	1.10	2.22	**1.03**	4.01	0.85	0.93	0.86	**3.64**	3.46	1.80	2.92	4.61	1.30	**3.44**
Static	M1	0.49	**0.94**	0.38	2.70	2.67	2.40	1.51	4.21	2.51	5.05	1.09	1.09	1.39	6.13	4.38	2.33	3.68	7.05	1.29	4.76
	M2	4.84	2.04	1.39	3.35	6.38	2.42	2.84	7.45	6.22	14.97	2.49	1.89	4.30	5.92	5.88	6.20	5.85	9.18	3.64	12.09
	M3	0.44	0.95	0.32	2.36	2.49	2.09	1.20	2.80	1.86	4.83	1.09	1.04	1.24	6.03	3.79	2.14	3.68	7.06	**1.21**	3.43
	M4.1	2.57	2.13	0.94	2.97	5.86	2.42	2.51	14.65	4.65	13.17	1.60	2.05	3.94	**5.38**	6.02	5.92	**3.19**	7.91	5.54	10.54
	M4.2	4.13	2.01	1.12	3.02	5.68	2.39	2.61	10.76	5.21	13.89	2.07	2.07	4.15	5.57	6.09	6.10	4.01	8.36	4.89	11.01
	M4.3	0.49	0.94	0.38	2.70	2.67	2.45	1.51	4.21	2.51	5.05	1.09	1.09	1.39	6.13	4.38	2.33	3.68	7.05	1.29	4.76
	M4.4	0.49	3.57	0.38	3.36	2.67	2.46	2.81	4.21	2.51	5.05	1.09	2.69	1.39	5.46	4.38	2.33	3.68	7.05	1.29	4.76
	M5.1	0.42	0.96	0.32	2.15	2.32	1.97	0.92	2.29	1.81	4.24	1.10	1.02	1.22	6.03	3.82	**2.02**	3.57	7.04	1.25	3.41
	M5.2	0.42	0.95	0.32	2.18	2.39	1.99	0.96	2.33	1.76	4.45	1.09	1.02	**1.22**	5.97	**3.66**	2.07	3.60	7.05	1.25	**3.36**
	M5.3	**0.41**	0.96	0.32	**2.13**	**2.19**	**1.92**	0.92	2.24	1.60	**4.18**	**1.05**	1.00	1.23	5.90	4.09	2.04	3.57	**6.88**	1.32	3.75
	M5.4	0.41	0.96	**0.32**	2.15	2.25	1.92	**0.91**	**2.22**	**1.56**	4.19	1.08	**0.99**	1.28	5.79	3.71	2.06	3.51	6.96	1.33	3.78

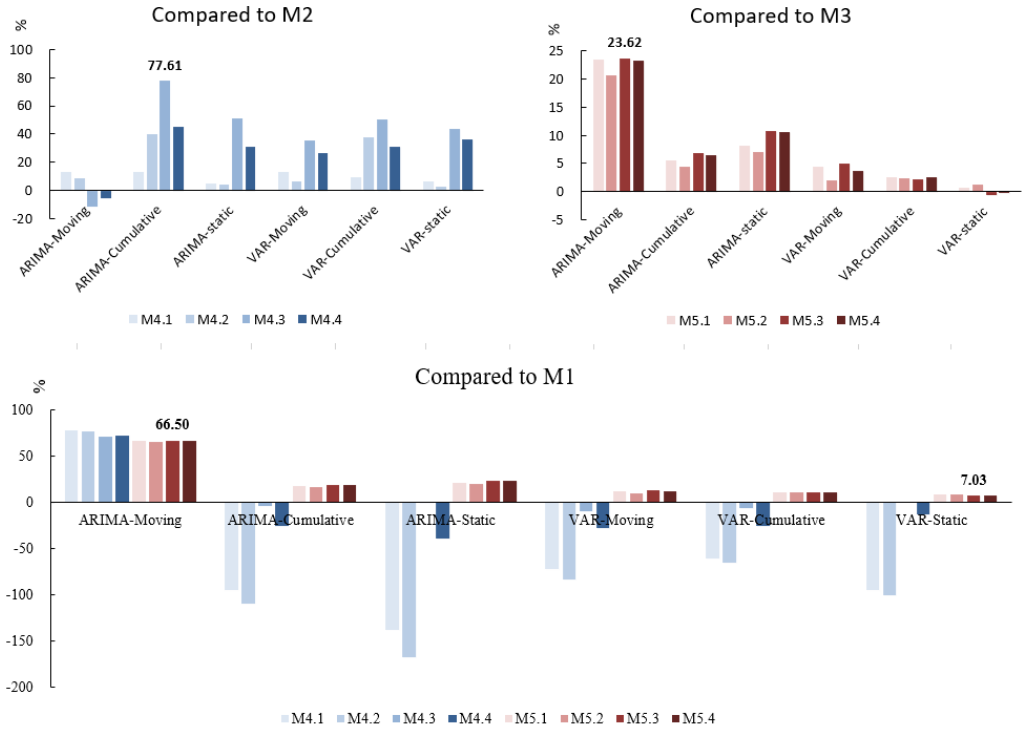
Note: M1 to M3 represent Model 1 to Model 3, and M4.1 to M4.5 represent Model 4 with different thresholds. For instance, M4.1 represents Model 4 with Threshold 1. Similarly, M5.1 to M5.4 represent Model 5 with different thresholds. The red and bold font in the results indicates the minimum MSE of the eleven models compared.

Table 4 Comparison of the models

CEM	Model	Model 1				Model 2				Model 3				Model 4			
		Moving	Cumulative	Static	Total	Moving	Cumulative	Static	Total	Moving	Cumulative	Static	Total	Moving	Cumulative	Static	Total
ARIMA	M2	10	1	0	11(36.67%)	-	-	-	-	-	-	-	-	-	-	-	-
	M3	10	9	9	28(93.33%)	3	10	10	23(76.67%)	-	-	-	-	-	-	-	-
	M4.1	10	1	0	11(36.67%)	10	9	8	27(90.00%)	9	0	0	9(30.00%)	-	-	-	-
	M4.2	10	1	1	12(40.00%)	9	8	9	26(86.67%)	8	0	0	8(26.67%)	-	-	-	-
	M4.3	10	2	0	12(40.00%)	6	10	9	25(83.33%)	7	0	1	8(26.67%)	-	-	-	-
	M4.4	10	1	0	11(36.67%)	8	10	7	25(83.33%)	8	0	0	8(26.67%)	-	-	-	-
	M5.1	10	9	9	28(93.33%)	5	10	10	25(83.33%)	7	9	9	25(83.33%)	2	10	10	22(73.33%)
	M5.2	10	9	9	28(93.33%)	3	10	10	23(76.67%)	7	9	9	25(83.33%)	2	10	10	22(73.33%)
	M5.3	10	9	9	28(93.33%)	3	10	10	23(76.67%)	8	9	9	26(86.67%)	3	10	9	22(73.33%)
	M5.4	10	9	9	28(93.33%)	3	10	10	23(76.67%)	8	9	9	26(86.67%)	2	10	10	22(73.33%)
VAR	M2	1	1	1	3(10.00%)	-	-	-	-	-	-	-	-	-	-	-	-
	M3	9	10	8	27(90.00%)	9	9	9	27(90.00%)	-	-	-	-	-	-	-	-
	M4.1	2	2	2	6(20.00%)	9	8	7	24(80.00%)	1	2	2	5(16.67%)	-	-	-	-
	M4.2	1	2	1	4(13.33%)	8	8	7	23(76.67%)	1	2	1	4(13.33%)	-	-	-	-
	M4.3	0	1	6	7(23.33%)	8	9	9	26(86.67%)	0	0	2	2(6.67%)	-	-	-	-
	M4.4	0	2	6	8(26.67%)	7	7	9	23(76.67%)	0	0	3	3(10.00%)	-	-	-	-
	M5.1	9	10	9	28(93.33%)	9	9	9	27(90.00%)	9	8	7	24(80.00%)	9	8	8	25(83.33%)
	M5.2	9	10	9	28(93.33%)	9	9	9	27(90.00%)	7	8	9	24(80.00%)	9	8	9	26(86.67%)
	M5.3	9	9	9	27(90.00%)	9	9	10	28(93.33%)	8	8	7	23(76.67%)	10	10	9	29(96.67%)
	M5.4	9	9	9	27(90.00%)	9	9	10	28(93.33%)	6	7	7	20(66.67%)	10	10	8	28(93.33%)

Note: The "Total" column is obtained by summarizing the "Moving", "Cumulative", and "Static" column. The percentage in parentheses indicates the overall percentage by which the model in the row outperforms the model in the column across all three forecast windows. For instance, the value in the first cell of the first 'Total' column, 11(36.67%), implies that, when using ARIMA as the reference model, there are 11 instances in which Model 2 performs better than Model 1 across all ten economies and three forecasting windows, accounting for 36.67% of the total.

Figure 3. Percentage of Improvement



Furthermore, upon analysis of Table 3, it is observed that when the CEM is ARIMA, for the nine economies, with the exception of the UK, the MSE minimums were obtained by Model 5, which incorporates both FEM and TSPM. Specifically, Model 5 with threshold 3 and 4 account for six and three instances, respectively. Similarly, for the five models utilizing VAR as the CEM, the MSE minimums were relatively dispersed, but Model 5 still accounted for six instances. This highlights that Model 5, which combines the strengths of both linear and nonlinear models, is the best forecasting model. Hence, it can be concluded that the FEM and TSPM complement each other and yield better predictive performance.

4.2.2. Compare Models Using Diebold-Mariano Test

This section presents the results of the DM test, which complements the MSE comparison section. In particular, we focus on the difference between Model 4 and Model 2, as well as the difference between Model 5 and Model 3. Additionally, we examine the differences between Model 1 and Model 2, Model 1 and Model 3, and Model 2 and Model 3. Table 5 displays the results.

Table 5 P-values of the DM tests

CEM		ARIMA										VAR									
Window	VS.	USA	UK	France	Japan	Canada	Australia	Germany	SK	SA	TW	USA	UK	France	Japan	Canada	Australia	Germany	SK	SA	TW
Moving	A	0.01	0.91	0.31	0.21	0.33	0.64	0.15	0.68	0.00	0.05	0.08	0.18	0.70	0.10	0.59	0.64	0.00	0.02	0.02	0.24
	B	0.02	0.98	0.26	0.07	0.28	0.88	0.14	0.19	0.00	0.08	0.02	0.04	0.29	0.08	0.50	0.03	0.00	0.02	0.00	0.48
	C	0.81	0.53	0.30	0.37	0.46	0.18	0.35	0.00	0.00	0.20	0.06	0.09	0.04	0.09	0.40	0.01	0.01	0.32	0.38	0.19
	D	0.38	0.64	0.37	0.36	0.45	0.85	0.69	0.00	0.00	0.30	0.06	0.25	0.04	0.14	0.65	0.04	0.01	0.26	0.36	0.17
	E	0.00	0.04	0.01	0.00	0.00	0.00	0.02	0.61	0.00	0.01	0.01	0.00	0.00	0.02	0.01	0.03	0.00	0.01	0.00	0.09
	F	0.00	0.04	0.01	0.00	0.00	0.00	0.03	0.09	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.02	0.00	0.05	0.00	0.04
	G	0.00	0.03	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.07
	H	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.02	0.00	0.05
	I	0.00	0.00	0.00	0.01	0.02	0.00	0.04	0.00	0.00	0.00	0.06	0.05	0.04	0.07	0.40	0.00	0.01	0.32	0.54	0.16
	J	0.00	0.00	0.01	0.73	0.16	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	K	0.15	0.14	0.00	0.04	0.06	0.00	0.08	0.08	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Cumulative	A	0.33	0.95	0.42	0.10	0.14	0.38	0.63	0.47	0.00	0.05	0.09	0.56	0.47	0.16	0.08	0.80	0.32	0.12	0.00	0.09
	B	0.00	0.70	0.26	0.09	0.34	0.81	0.30	0.01	0.00	0.25	0.02	0.33	0.11	0.06	0.52	0.54	0.05	0.06	0.00	0.35
	C	0.01	0.14	0.11	0.72	0.13	0.31	0.05	0.05	0.05	0.07	0.08	0.06	0.05	0.18	0.27	0.01	0.84	0.29	0.77	0.19
	D	0.01	0.17	0.06	0.54	0.13	0.30	0.11	0.05	0.06	0.34	0.08	0.64	0.05	0.19	0.69	0.07	0.79	0.28	0.75	0.19
	E	0.00	0.04	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.03
	F	0.00	0.02	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.02	0.00	0.02	0.00	0.04
	G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.02	0.00	0.01	0.00	0.01
	H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.01	0.00	0.01	0.00	0.02
	I	0.01	0.03	0.07	0.48	0.13	0.87	0.06	0.05	0.04	0.07	0.08	0.06	0.05	0.11	0.27	0.01	0.84	0.27	0.60	0.17
	J	0.03	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	K	0.00	0.00	0.01	0.01	0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Static	A	0.00	0.40	0.01	0.01	0.12	0.38	0.05	0.00	0.00	0.00	0.01	0.06	0.10	0.19	0.81	0.62	0.00	0.00	0.00	0.00
	B	0.00	0.85	0.01	0.01	0.12	0.49	0.13	0.00	0.00	0.00	0.04	0.04	0.28	0.16	0.60	0.97	0.00	0.00	0.00	0.01
	C	0.00	0.02	0.02	0.22	0.09	0.48	0.10	0.02	0.00	0.05	0.03	0.09	0.02	0.90	0.21	0.00	0.05	0.35	0.00	0.06
	D	0.00	0.03	0.02	0.69	0.09	0.31	0.69	0.02	0.00	0.05	0.03	0.00	0.02	0.27	0.21	0.00	0.05	0.35	0.00	0.06
	E	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.78	0.02	0.19	0.00	0.00	0.00	0.00
	F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.00	0.01	0.00	0.00
	G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.02	0.00	0.00	0.00	0.00
	H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
	I	0.00	0.02	0.02	0.22	0.09	0.93	0.10	0.02	0.00	0.05	0.03	0.09	0.02	0.90	0.21	0.00	0.05	0.35	0.00	0.06
	J	0.01	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	K	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Note: The 'VS.' column indicates the models that the DM test compares. A-D represent Model 2 vs. the four Models 4s, respectively. E-H represent Model 3 vs. the four Model 5s, respectively. I, J and K means Model 1 vs. Model 2, Model 1 vs. Model3, Model 2 vs. Model3. The red numbers are those smaller than 0.05.

Table 6 Summary of the DM tests (P-value < 5%)

CEM	ARIMA				VAR			
VS.	Moving	Cumulative	Static	Total	Moving	Cumulative	Static	Total
A	3	2	6	11(36.67%)	3	1	5	9(30.00%)
B	2	3	6	11(36.67%)	6	2	6	14(46.67%)
C	2	2	5	9(30.00%)	3	2	4	9(30.00%)
D	2	1	5	8(26.67%)	3	1	5	9(30.00%)
E	9	10	10	29(96.67%)	9	10	8	27(90.00%)
F	9	10	10	29(96.67%)	10	10	9	29(96.67%)
G	10	10	10	30(100.00%)	9	10	9	28(93.67%)
H	10	10	10	30(100.00%)	9	10	10	29(96.67%)
I	10	3	5	18(60.00%)	4	2	4	10(33.33%)
J	8	10	10	28(93.33%)	10	10	10	30(100.00%)
K	5	10	10	25(83.33%)	10	10	10	30(100.00%)

To facilitate the analysis of the DM test results, we summarize cases with the P-value less than 5% in Table 6. The summary results indicate that when using ARIMA as the benchmark model, 32.5% (which is the average of 36.67%, 36.67%, 30%, and 26.67%) of Model 4s are significantly different from Model 2, and 98.33% (which is the average of 96.67%, 96.67%, 100.00%, and 100.00%) of Model 5s are significantly different from Model 3. All Model 5s using threshold 3 and threshold 4 are significantly different from Model 3. When the benchmark model is VAR, 34.17% (which is the average of 30.00%, 46.67%, 30.00%, and 30.00%) of Model 4s are significantly different from Models 2, and 94.17% (which is the average of 90.00%, 96.67%, 93.67%, and 93.67%) of Model 5s are significantly different from Model 3. When using ARIMA as the benchmark model, 60.00% of Model 1s are significantly different from Model 2, 93.33% are significantly different from Model 3, and 83.33% of Model 2s are significantly different from Model 3. Under the moving window, all Model 1s are significantly different from Model 2. This is because the ARIMA model performs poorly under the moving window, so the DL model performs better than it. When the moving window is excluded, only 40.00% of Model 1s are significantly different from Model 2. When using VAR as the benchmark model, 33.33% of Model 1s are significantly different from Model 2, all are significantly different from Model 3, and all Model 2s are significantly different from Model 3. The results suggest that the prediction performance of the general DL model is less different from that of the CEM. However, after using FEM, the DL model's accuracy improves. And the DL model using FEM and TSPM shows significantly better performance than the general DL model and CEM.

Based on the results of both the DM tests and MSE analysis, it can be concluded that the complex DL model alone performs poorly in GDP forecasting, even worse than the CEM. However, when the CEM is utilized to filter the DL models in ensemble, the predictive ability of the DL model improves significantly. Moreover, the FEM still improves the DL model's predictive ability when applied to the TSPM. Among all the models, the DL model using both the FEM and the TSPM has the best predictive ability.

4.3. Test of the Overfitting on Validation Sets

We argue that because of the small sample size, overfitting occurs not only on the training set but also on the validation set when DL models are used in macroeconomic forecasting. Therefore, we propose using the CEM as a benchmark and the FEM to reselect DL models to mitigate the

overfitting on the validation set. Our results indicate that this approach yields significant improvement in the accuracy of macroeconomic forecasts. To further investigate the possibility of overfitting on the validation set and the importance of the benchmark, we develop regression equations (25) and (26) to explore the relationship between the MSE on the prediction set and the MSE on the validation set. Specifically, we use Model 3 under the static window as an example, and construct regressions. If the overfitting on the validation set is present, a U-shaped or negative relationship between the MSE on the prediction set and the MSE on the validation set should be observed. Conversely, the absence of such a relationship suggests that overfitting on the validation set is may not be a concern.

$$mse_{fore} = \alpha + \beta_1 * mse_{valid} + \varepsilon \tag{25}$$

$$mse_{fore} = \alpha + \beta_1 * mse_{valid} + \beta_2 * mse_{valid}^2 + \varepsilon \tag{26}$$

where mse_{fore} represents the MSE on the prediction set, mse_{valid} represents the MSE on the validation set, and mse_{valid}^2 valid represents the squared term of mse_{valid}.

Furthermore, Equation (27) and (28) were constructed to explore how the introduction of the baseline model affects the relationship between the MSE on the prediction set and the MSE on the validation set. If the overfitting on the validation set is mitigated after controlling the Euclidean distance between the DL model's predictions and benchmark model's predictions on the validation set, the U-shaped or negative relationship between the MSE on the prediction set and the MSE on the validation set will be no longer significant or even reverse to an inverted U-shaped or positive relationship. These regression analyses allowed us to explore the impact of the benchmark role on reducing overfitting and improving the prediction accuracy of DL models.

$$mse_{fore} = \alpha + \beta_1 * mse_{valid} + \beta_3 * dis_{valid} + \varepsilon \tag{27}$$

$$mse_{fore} = \alpha + \beta_1 * mse_{valid} + \beta_2 * mse_{valid}^2 + \beta_3 * dis_{valid} + \beta_4 * dis_{valid}^2 + \varepsilon \tag{28}$$

where dis_{valid} represents the Euclidean distance between the DL model predictions and the benchmark model predictions on the validation set, and dis_{valid}^2 represents the squared term of dis_{valid}.

Table 7 presents the regression results for the United States, France, Canada, and Australia, and the results for other economies are similar. We investigate the possibility of overfitting on validation sets from equations e1 and e2, and we analyze whether overfitting on the validation set is still strong after controlling the Euclidean distance between the DL model's predictions and benchmark model's predictions on the validation set from equations e3 and e4.

For the United States, Both the two equation e2s indicate a positive U-shaped relationship between the MSE on the prediction set and the MSE on the validation set, suggesting the high possibility of overfitting on the validation set. However, after controlling the Euclidean distance between the DL model's predictions and benchmark model's predictions on the validation set, as shown in equations e4s, one significant positive U-shaped relationship switches to a significant negative U-shaped relationship, and another becomes no longer significant, indicating a reduction in overfitting on validation sets. Similarly, the first equation e2 for France shows a positive U-shaped relationship between the MSE on the prediction set and the MSE on the validation set, and shifts to negative U-shaped relationship after controlling the Euclidean distance between the DL model's predictions and ARIMA's predictions on the validation set. However, the second equation e1 and e2 for France all show no significant relationship between the MSE on the prediction set and the MSE on the validation set, and the insignificant relationship in equation e1 switches to significant positive relationship in the second equation e3, while the insignificant relationship in equation e2 remains insignificant in the second equation e4. Both changes reflect

that controlling the Euclidean distance between the DL model's predictions and VAR's predictions on the validation set does no bad to the DL models which show no significant overfitting. For Canada, the first equation e1 shows a negative relationship and the negative relationship is no longer significant in the first equation e3, which reflects the filtering effect of ARIMA. While the second equation e1 shows a positive relationship and the positive relationship is still significant in the second equation e3. The results for Australia are similar to those for the United States.

Table 7 Regression results

CEM	ARIMA				VAR			
Equation	e1	e2	e3	e4	e1	e2	e3	e4
Variable	mse_fore				mse_fore			
USA								
mse_valid	-0.23**	-4.87**	0.082**	5.52***	0.01	-14.20**	-0.06	8.71
	(0.09)	(2.38)	(0.04)	(1.35)	(0.22)	(6.19)	(0.20)	(6.99)
mse_valid_2		8.36*		-9.64***		20.46**		-12.76
		(4.26)		(2.37)		(8.91)		(10.09)
dis_valid			1.00***	1.26***			0.81***	0.29
			(0.04)	(0.13)			(0.14)	(0.36)
dis_valid_2				-3.01				6.40*
				(1.90)				(3.46)
Constant	0.07***	0.72**	-0.02**	-0.79***	0.02	2.48**	0.02	-1.47
	(0.02)	(0.33)	(0.01)	(0.19)	(0.08)	(1.074)	(0.07)	(1.21)
Observations	100	100	100	100	100	100	100	100
R-squared	0.06	0.10	0.87	0.89	0.00	0.05	0.24	0.27
France								
mse_valid	-3.70***	-42.59***	-1.43***	24.27***	0.50	-2.19	0.72*	-3.71
	(0.41)	(10.02)	(0.20)	(5.86)	(0.30)	(4.27)	(0.43)	(4.45)
mse_valid_2		81.78***		-50.50***		4.46		7.96
		(21.06)		(11.85)		(7.04)		(7.58)
dis_valid			2.19***	5.86***			0.23	1.00
			(0.10)	(0.61)			(0.32)	(0.84)
dis_valid_2				-41.62***				-2.49
				(6.88)				(2.70)
Constant	1.02***	5.60***	0.45***	-2.82***	0.28***	0.68	0.20	0.77
	(0.09)	(1.18)	(0.05)	(0.72)	(0.09)	(0.64)	(0.15)	(0.64)
Observations	100	100	100	100	100	100	100	100
R-squared	0.45	0.53	0.91	0.93	0.03	0.031	0.03	0.05
Canada								
mse_valid	-1.78***	32.04***	-0.01	43.30***	0.64***	-1.80	0.83**	1.94
	(0.41)	(7.72)	(0.38)	(4.72)	(0.22)	(3.20)	(0.33)	(3.38)
mse_valid_2		-14.76***		-18.64***		0.78		-0.53
		(3.37)		(2.05)		(1.02)		(1.12)
dis_valid			3.43***	5.93***			0.21	-1.65**
			(0.41)	(0.79)			(0.27)	(0.71)
dis_valid_2				-7.55***				2.25***
				(2.66)				(0.78)

CEM	ARIMA				VAR			
Equation	e1	e2	e3	e4	e1	e2	e3	e4
Variable	mse_fore				mse_fore			
Constant	2.93***	-16.37***	0.74	-24.42***	-0.54	1.34	-0.88	-1.06
	(0.48)	(4.43)	(0.45)	(2.73)	(0.36)	(2.48)	(0.58)	(2.53)
Observations	100	100	100	100	100	100	100	100
R-squared	0.16	0.30	0.51	0.75	0.08	0.08	0.08	0.16
Australia								
mse_valid	-1.12	-22.52***	-3.16***	8.83***	-2.87**	-64.12**	-1.26	30.00
	(0.71)	(7.00)	(0.24)	(2.25)	(1.14)	(26.68)	(0.88)	(23.87)
mse_valid_2		9.12***		-5.10***		29.70**		-14.98
		(2.97)		(0.97)		(12.93)		(11.52)
dis_valid			5.01***	5.80***			2.88***	1.54*
			(0.17)	(0.32)			(0.33)	(0.78)
dis_valid_2				-0.52				3.07**
				(0.37)				(1.21)
Constant	3.97***	16.24***	5.53***	-1.42	3.47***	35.01**	1.58*	-14.63
	(0.73)	(4.05)	(0.24)	(1.30)	(1.17)	(13.77)	(0.91)	(12.38)
Observations	100	100	100	100	100	100	100	100
R-squared	0.02	0.11	0.90	0.93	0.06	0.11	0.47	0.53

*Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.*

In summary, our findings indicate that when using DL models for macroeconomic prediction, the overfitting on the validation set does exist, and the filtering effect of CEMs can assist in alleviating overfitting and improving the model's prediction capabilities when the overfitting on the validation set exists. Additionally, the use of the CEM does not significantly diminish the DL model's prediction capabilities when there is no significant overfitting on the validation set. Hence, the addition of the benchmark CEM can aid in improving the DL model's performance.

5. Conclusion & Outlook

The primary objective of this study is to improve the performance DL models in macroeconomic forecasting with the assistance of CEMs. Due to the limited availability of macroeconomic data, training DL models can be challenging, and overfitting is a significant issue that affect the accuracy of predictions. In this research, we have identified the problem of overfitting in DL models, specifically on the validation set, using GDP forecasting as an example. To address this issue, we propose using CEMs as benchmarks to filter DL models and using them for the ensemble. This approach has yielded promising results in GDP forecasting and can serve as a reference solution for forecasting other macroeconomic variables.

Our study also emphasizes the importance of incorporating economic knowledge into DL models to improve their performance in macroeconomic forecasting. Rather than applying DL models alone, we advocate for an effective combination of CEMs, which is structural and has definitely incorporated economic knowledge, and DL models to capitalize on the strengths of both. Additionally, we highlight the limitations of the DL in macroeconomic forecasting and the forecasting with small sample in other field, which can provide valuable insights for future advancements in both fields.

Our study has some limitations, including that we only use the traditional macroeconomic variables, and the CEMs are only used to filter the DL models. Future research can explore the

potential of incorporating emerging big data into macroeconomic forecasting or use CEMs for dataset augmentation, i.e., generating new samples for DL model training. Another possibility is to determine the appropriate complexity of DL models with the help of CEMs.

Acknowledgement

This work was granted by the Tsinghua Guoqiang Institute, the Tsinghua-Yinchuan Institute for the Internet of Water and the National Natural Science Foundation of China.

References

Adolfson, M., Lindé, J. and Villani, M., 2007. Forecasting Performance of an Open Economy DSGE Model. *Econometric Reviews*, 262–4, pp.289–328. https://doi.org/10.1080/07474930701220543.

Adolfson, M., Andersson, M., Linde, J., Villani, M. and Vredin, A., 2007. Modern forecasting models in action: improving macroeconomic analyses at central banks. *International Journal of Central Banking*, 3, pp.111–144. https://www.ijcb.org/journal/ijcb07q4a4.htm.

Artis, M. and Okubo, T., 2010. The UK intranational business cycle. *Journal of Forecasting,* 291-2, pp.71-93. https://doi.org/10.1002/for.1141.

Babu, C.N. and Reddy, B.E., 2014. A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. *Applied Soft Computing*, 23, pp.27-38. https://doi.org/10.1016/j.asoc.2014.05.028.

Bahdanau, D., Cho, K. and Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations ICLR15, Yoshua Bengio and Yann LeCun Eds.. http://arxiv.org/abs/1409.0473.

Bajari, P., Nekipelov, D., Ryan, S.P. and and Yang M., 2015. Machine Learning Methods for Demand Estimation. *American Economic Review*, 105 5, pp.481-85. https://doi.org/10.1257/aer.p20151021.

Balla, J., Huang, S., Dugan, O., Dangovski, R. and Soljacic, M., 2022. Ai-assisted discovery of quantitative and formal models in social science. arXiv preprint arXiv:2210.00563. https://doi.org/10.48550/arXiv.2210.00563.

Barkan, O., Benchimol, J., Caspi, I., Cohen, E., Hammer, A. and Koenigstein, N., 2023. Forecasting CPI inflation components with hierarchical recurrent neural networks. *International Journal of Forecasting*, 393, pp.1145-1162. https://doi.org/10.1016/j.ijforecast.2022.04.009.

Binner, J.M., Kendall, G. and Chen, S.-H. Ed., 2004. Applications of Artificial Intelligence in Finance and Economics. *Advances in Econometrics*. 19, pp.71-91. https://doi.org/10.1016/S0731-90530419013-0.

Boero, G., 1990. Comparing ex-ante forecasts from a sem and var model: An application to the italian economy. *Journal of Forecasting*, 9, pp.13-24. https://doi.org/10.1002/for.3980090103.

Cadenas, E. and Rivera, W., 2010. Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model. *Renewable Energy*, 3512, pp. 2732-2738. https://doi.org/10.1016/j.renene.2010.04.022.

Červená, M. and Schneider, M., 2014. Short-term forecasting of GDP with a DSGE model augmented by monthly indicators. *International Journal of Forecasting*, 303, pp. 498-516. https://doi.org/10.1016/j.ijforecast.2014.01.005.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S., 2016. Productivity and Selection of Human Capital with Machine Learning. *American Economic Revie*w, 106 5, pp. 124-27. https://doi.org/10.1257/aer.p20161029.

Chin, K.-H., 2022. Forecast evaluation of DSGE models: Linear and nonlinear likelihood. *Journal of Forecasting*, 416, pp.1099–1130. https://doi.org/10.1002/for.2850

Christoffel, K., Coenen, G. and Warne, A., 2008. The New Area-Wide Model of the Euro Area: Specification, Estimation Results and Properties. European Central BankWorking Paper, October 944. Available at: <https://www.proquest.com/publiccontent/working-papers/new-area-wide-model-euro-micro-founded-open/docview/1698252832/sem-2?accountid=14426>.

Christoffel, K., Coenen, G. and Warne, A., 2011. *Forecasting with DSGE models*. In: Clements. M.,Hendry, D.Eds., Oxford Handbook on Economic Forecasting. Oxford University Press, pp.89–128.

Del Negro, M. and Schorfheide, F., 2013. Bayesian macroeconometrics. *The Oxford Handbook of Bayesian Econometrics*, pp. 293–389.

Dixon, M., 2018. Sequence classification of the limit order book using recurrent neural networks. *Journal of Computational Science*, 24, pp.277-286. https://doi.org/10.1016/j.jocs.2017.08.018.

D. Vrontos, S., Galakis, J. and D. Vrontos, I., 2021. Modeling and predicting U.S. recessions using machine learning techniques. *International Journal of Forecasting*, 372, pp.647-671. https://doi.org/10.1016/j.ijforecast.2020.08.005.

Edge, R.M., Kiley, M.T. and Laforte, J.P., 2010. A comparison of forecast performance between federal reserve staff forecasts, simple forecasts, simple reduced-form models, and a DSGE model. *Journal of Applied Econometrics*, 25, pp.720–754. https://doi.org/10.1002/jae.1175.

Edge, R.M. and Gürkaynak, R., 2010. How useful are estimated DSGE model forecasts for central bankers?. *Brookings Papers on Economic Activity*, pp.209–259. http://www.jstor.org/stable/41012847.

Fair RC., 2019. Information content of DSGE forecasts. *Journal of Forecasting*, 386, pp. 519-524. https://doi.org/10.1002/for.2581.

Fukuda K., 2007. Forecasting real-time data allowing for data revisions. *Journal of Forecasting*, 266, pp.429-444. https://doi.org/10.1002/for.1032.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Comput*, 9(8), pp.1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, J.K., 2021. LSTM-based Sales Forecasting Model. *Transactions on Internet & Information Systems*, 15, pp1232–1245. https://doi.org/10.3837/tiis.2021.04.003.

Irsoy, O. and Cardie, C., 2014. Opinion Mining with Deep Recurrent Neural Networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 720-728. https://doi.org/10.3115/v1/D14-1080.

Kim, H.H. and Swanson, N.R., 2018. Methods for backcasting, nowcasting and forecasting using factor‐MIDAS: With an application to Korean GDP. *Journal of Forecasting*, 373, pp.281-302. https://doi.org/10.1002/for.2499.

Klein, L.R., 1970. An Essay on Theory of Economic Prediction. Markham Publishing Company Markham Economics Series, Chicago.

Kolasa, M., Rubaszek, M. and Skrzypczynski, P., 2012. Putting the New Keynesian DSGE Model to the Real-Time Forecasting Test. *Journal of Money, Credit and Banking*, 44, pp.1301-1324. https://doi.org/10.1111/j.1538-4616.2012.00533.x.

Krauss C., Do, X.A. and Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 2592, pp. 689-705. https://doi.org/10.1016/j.ejor.2016.10.031.

Kuck, K. and Schweikert, K., 2021. Forecasting Baden‐Württembergs GDP growth: MIDAS regressions versus dynamic mixed‐frequency factor models. *Journal of Forecasting*, 405, pp. 861-882. https://doi.org/10.1002/for.2743.

Kuzin, V., Marcellino, M., and Schumacher, C., 2011. MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 272, pp.529-542. https://doi.org/10.1016/j.ijforecast.2010.02.006.

Kydland, F.E. and Prescott, E.C., 1982. Time to build and aggregate fluctuations. *Econometrica*, 506, pp.1345–1370. https://www.proquest.com/scholarly-journals/time-build-aggregate-fluctuations/docview/214651219/se-2.

Liu, G., Gupta, R., and Schaling, E., 2009. A new-Keynesian DSGE model for forecasting the South African economy. *Journal of Forecasting*, 28, pp.387–404. https://doi.org/10.1002/for.1103.

Liu, T.X. and Xu X.F., 2015. Can Internet Search Behavior Help to Forecast the Macro Economy?. Economic Research Journal, 5012, pp.68-83.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E.,, Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A., Ziel., F., 2022. Forecasting: theory and practice. *International Journal of Forecasting*, 383, pp.705-871. https://doi.org/10.1016/j.ijforecast.2021.11.001.

Richardson, A., Mulder, T.V.F., and Vehbi, T., 2021. Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 372, pp.941-948. https://doi.org/10.1016/j.ijforecast.2020.10.005.

Roman, J. and Jameel, A., 1996. Backpropagation and Recurrent Neural Networks in Financial Analysis of Multiple Stock Market Returns. 29th Annual Hawaaii International Conference on System Sciences, 2, pp. 454-460. https://doi,org/10.1109/HICSS.1996.495431.

Rubaszek, M. and Skrzypczy´ nski, P., 2008. On the forecasting performance of a small-scale DSGE model. *International Journal of Forecasting*, 243, pp.498–512. https://doi.org/10.1016/j.ijforecast.2008.05.002.

Salazar, E. and Weale, M., 1999. Monthly data and short-term forecasting: an assessment of monthly data in a VAR model. *Journal of Forecasting*, 18, pp.447-462. https://doi.org/10.1002/SICI1099-131X19991218:7<447::AID-FOR726>3.0.CO;2-T.

Shewalkar, A.N., 2018. Comparison of RNN, LSTM and GRU on Speech Recognition Data. North Dakota State University. https://hdl.handle.net/10365/29111.

Siami-Namini, S., Tavakoli, N., and Siami-Namin, A., 2018. A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications ICMLA, pp.1394-1401. https://doi.org/ 10.1109/ICMLA.2018.00227.

Smets,F., Warne, A. and Wouters, R., 2014. Professional forecasters and real-time forecasting with a DSGE model. *International Journal of Forecasting*, 304, pp.981-995. https://doi.org/10.1016/j.ijforecast.2014.03.018.

Song X.Y., Liu Y.T., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L. and Cheng, Z., 2020. Time-series well performance prediction based on Long Short-Term Memory LSTM neural network model. *Journal of Petroleum Science and Engineering*, 186, 106682. https://doi.org/10.1016/j.petrol.2019.106682.

Tallman, E.W. and Zaman, S., 2020. Combining survey long-run forecasts and nowcasts with BVAR forecasts using relative entropy. *International Journal of Forecasting*, 362, pp.373-398. https://doi.org/10.1016/j.ijforecast.2019.04.024.

Thomakos, D.D. and Guerard, J.B., 2004. Naïve, ARIMA, nonparametric, transfer function and VAR models: A comparison of forecasting performance. *International Journal of Forecasting*, 201, pp.53-67. https://doi.org/10.1016/S0169-20700300010-4.

Tinbergen, J., 1939. Business Cycles in the United States, 1919–1932 – Statistical Testing of Business-Cycle Theories. League of Nations, Economic Intelligence Service, Geneva. http://hdl.handle.net/1765/14937.

Tinbergen, J., 1974. *The Dynamics of Business Cycles: A Study in Economic Fluctuations*. University of Chicago Press.

Treisman, A.M. and Gelade, G., 1980. A Feature-integration Theory of Attention. *Cognitive Psychology*, 121, pp.97-136. https://doi.org/10.1016/0010-02858090005-5.

Varian, H. R., 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), pp. 3. https://doi.org/10.1257/jep.28.2.3.

Vaswani, A., Shazeer, N., Parmar,N., Uszkoreit, J., Jones, L., Gomez, A.L., Kaiser, L. and Polosukhin I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

Weytjens, H., Lohmann, E., and Kleinsteuber, M., 2021. Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet. *Electron Commer Res*, 21, pp.371–391. https://doi.org/10.1007/s10660-019-09362-7.

Wolters, M., 2011. Forecasting Under Model Uncertainty. Goethe University Frankfurt Working Paper. https://www.econstor.eu/bitstream/10419/48723/1/VfS_2011_pid_231.pdf.

Wieland, V. and Wolters, M.H., 2011. The diversity of forecasts from macroeconomic models of the US economy. *Economic Theory*, 47, pp.247–292. https://doi.org/ 10.1007/s00199-010-0549-7.

Yang PR., 2020. Using the yield curve to forecast economic growth. *Journal of Forecasting*, 397, pp. 1057-1080. https://doi.org/10.1002/for.2676.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, pp.159-175. https://doi.org/10.1016/S0925-23120100702-0.