Data Science and Engineering (DSE) Record, Volume 3, issue 1.

Prediction and Analysis of Knowledge-based Economy Indicators on GDP Growth Using Neuro-Fuzzy Technique

Itsarawadee Hema[1] and Narissara Eiamkanitchart [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

Itsarawadee_h@cmu.ac.th

Abstract. The objectives of this independent study consisted of two areas. Firstly, to select appropriate variables of the knowledge-based economy indicators as alternative indicators for predicting Gross Domestic Product (GDP) growth. Secondly, to develop models for forecasting the GDP growth rate using neuro-fuzzy technique and compare the model performance. The data used in this work were collected from the World Bank through an Application Programming Interface, consisting of 5 regions: East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, and South Asia. The study investigated and identified the independent variables of the knowledge-based economy that could be used in the GDP growth rate prediction model along with the development of the Adaptive Neuro-fuzzy Inference System (ANFIS) to predict the GDP growth rate. The performance assessment used the prediction results to compare with the Linear Regression (LR) and Artificial Neural Network (ANN) models, using the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results showed that ANFIS provided the highest accuracy in predicting GDP growth rate in 14 of 15 experiments from three types of data: training dataset, testing dataset, and unseen dataset), while the ANN and LR models are less accurate, respectively. The East Asia & Pacific region has the lowest error of all regions; with the average MAE and RMSE of the testing and unseen datasets at 0.265% and 0.345%, respectively.

Keywords: GDP growth, Knowledge-based economy, ANFIS, forecasting.

1 Introduction

Nowadays, Countries around the world are experiencing disruption in their economic growth due to the coronavirus (COVID-19) outbreak. Since the GDP growth rate is the significant number which reflects the economic condition during this pandemic, all countries have reported these numbers of quarter 1/2020 recently and all countries are affected as expected. In Fig. 1, it can be seen that the events that are important to the economic growth of all regions in the past 20 years can be divided into two major

events: the 2008-2009 Financial Crisis and the COVID-19 pandemic. However, it was found that all regions were greatly affected by the epidemic situation of the coronavirus, while going back to the financial crisis, some regions are less affected such as the East Asia & Pacific, Middle East & North Africa and South Asia regions.

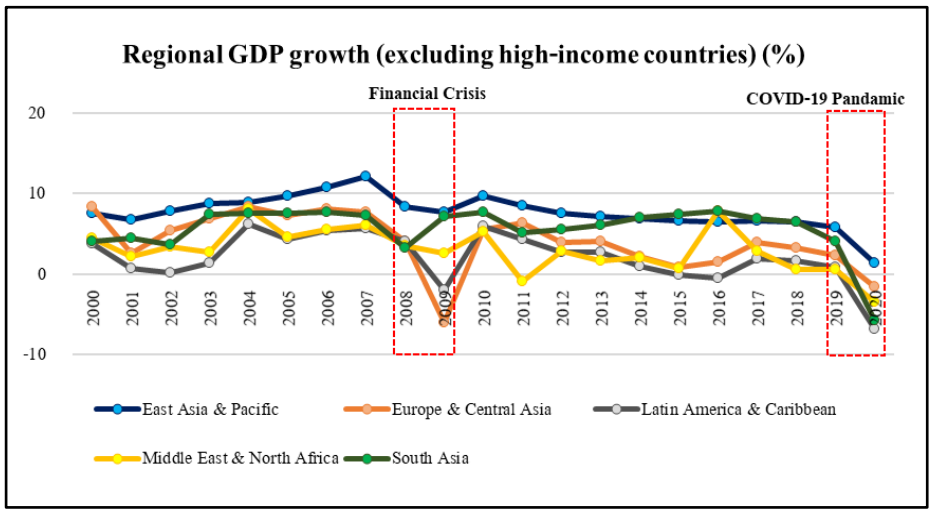

Fig. 1 Gross Domestic Product (GDP) growth by region since 2000-2020.

Recently, all affected countries have initiated policy formulations to revitalize the domestic economy in various forms. Also, many countries around the world have recognized the importance of the development process in the use of information technology as the driving force of the new economy. Blinder [1] has suggested that productivity growth can be built through improving the quality of the workforce through education and training to enhance skills, building the readiness of the technological infrastructure as well as technological developments through research and innovation. Therefore, in most developing countries, which are countries where is an efficiency-driven Economy, it is necessary to recognize the importance of creating sustainable development. Currently, the world is changing from a resource-based economy to the knowledge-based economy which the important components are the development of the potential of manpower and empowering creativity that can help revive the shrinking economy around the world. In addition, the components of factors that can be used to measure the knowledge-based economy can be divided into 4 categories: computer infrastructure; Information and communication technology infrastructure Education and skill-building, and research for technology development and innovation [2].

As aforementioned, it can be seen that the recognition of the importance of knowledge-based economic factors can create sustainable economic growth for the countries and regions. Since the GDP growth forecast can show the level of development and the production efficiency of the country, an accuracy of forecasting is

therefore something that should be taken into account in order to form the framework of the country's key policy-making in all departments. Thus, this independent study aims to improve the efficiency of models used to forecast GDP growth and can analyze variables that are important in the context of the current world situation, especially the knowledge-based economy indicators, by using the Neuro-fuzzy technique, which will be useful in further development of models used in economic forecasting.

2 Literature Review

2.1 Previous Studies Using Neuro-Fuzzy Technique

Mirbaghani [3] studied the GDP growth rate of Irian by applying the fuzzy logic technique and fuzzy neural networks to forecast the GDP growth and measured the accuracy of the forecast with the root mean square error. This study used eight supply-side variables, including investment and physical costs, power of labor, human costs, businesses, credit, and financial variables. inflation, government, political situation. The results revealed that the fuzzy neural networks model was able to predict the GDP growth with the highest accuracy (RMSE of 4.5158e-005) compare to the fuzzy logic model (RMSE of 0.0081).

Mladenovic et al. [4] analyzed the management of health care expenditures and the rate of GDP growth. The adaptive neuro-fuzzy technique was used to select the source variable affecting the dependent variable forecast and measured the accuracy of the forecast with the root mean square error. The efficiency was then compared with Artificial Neural Networks and Genetic Programming methods. The study found that Adaptive Neuro-fuzzy Inference System gives the RMSE minimum 0.9531 and R-Square maximum is 0.9941 comparatively.

Mardani et al. [5] revealed the relationship between energy demand, economic growth, and the amount of carbon dioxide emissions of the G20 countries using the Adaptive Neuro-fuzzy technique in the period 1962–2016. By using the aforementioned techniques, the study created a model to try forecasting carbon dioxide emissions by using two independent variables, namely energy demand and economic growth, and measured the accuracy of the forecast with the square root mean square error. The results showed that the adaptive network fuzzy inference technique showed error values in the data studied in the range of 0.0119% - 0.0732%.

3 Data and Methodology

3.1 Data

The data was gathered from the World Bank Organization's Application Programming Interface (API) method based on the knowledge-based economy indicators [6].

The variables from the World Bank Development Indicator (WDI) were collected using secondary time-series data in the period 2000 - 2018. divided into five regions as follows: East Asia & Pacific (EAP), Europe & Central Asia (ECA), Latin America & Caribbean (LAC), Middle East & North Africa (MNA), and South Asia (SAS). The selected variables are related to the knowledge economy which consists of three pillars in forecasting GDP growth (annual, %) as follows:

1) Education and Human Resources: primary enrollment **(%** gross**)**, secondary enrollment **(%** gross**)**, and tertiary enrollment **(%** gross**)**

2) Innovation System: patent application and scientific & technical journal articles

3) Information Infrastructure: fixed telephone subscriptions **(**per 100 people**)**, mobile cellular subscribers **(**per 100 people**)** and individuals using the internet **(%** of the population**)**

There are also other variables to avoid the omitted variable problem including household consumption as % of GDP, exports of goods and services as % of GDP, population growth (%).

3.2 Methodology

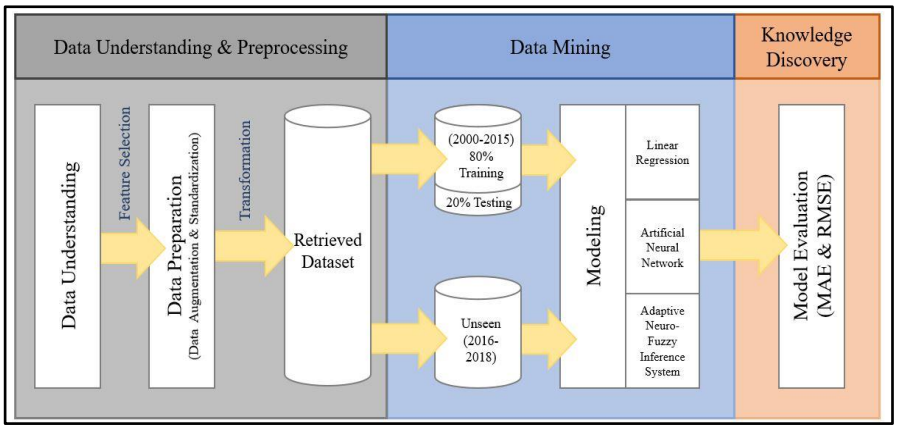

Fig. 2 Study Process

1) Data Understanding

Data understanding is one of the important processes before developing the model. It is a process for data analysis, leading to the appropriate use of variables. This study applied descriptive statistics and the correlation analysis between variables. The results reveal that the data in each of the selected variables have the highest and the minimum values which are quite different, especially in innovation system variables (patents and journals). It is therefore imperative to analyze the variables that are important to the GDP growth by region. The feature selection is applied to choose the best inputs combination from each group of variables using the linear regression analysis.

2) Data Preparation

This study uses the data preparation consisting of data augmentation technique to increase the data points in the dataset using the packages of Tsaug [7-9], Also, standardization technique is applied to scale those values, in order to reduce any bias from differentiation of magnitude, range, and units using the packages of Scikit-learn.

3) Modeling

Adaptive Neuro-fuzzy Inference System (ANFIS)

As has been introduced, the approach used for GDP growth prediction in this study is to explore the Neuro-fuzzy technique. The reason to choose an ANFIS is that it has been widely applied [10]. One of the advantages of ANFIS is, it is a combination of ANN and fuzzy systems using ANN learning capabilities to obtain fuzzy if-then rules with appropriate membership functions, which can learn something from the imprecise data that has been inputted and leads to the inference. Another advantage, it can make effective use of the self-learning and memory abilities of neural networks and brings a more stable training process. As seen in Fig. 2, ANFIS is constructed with five layers; the input layer is an antecedent parameter, three hidden layers are rule-based with three constant parameters and one consequent parameter, and one output layer. In the first layer, it transforms an input into a degree between 0 and 1 which is the so-called premise parameter. It is an activation function with membership function such as triangular, trapezoidal, Gaussian, or generalized-bell. The second layer will estimate each incoming signal to each neuron using the product operator; in the third layer, it will normalize all input signals while the fourth layer will do fuzzification. Finally, in the last layer, it will summarize all the weighted output values.

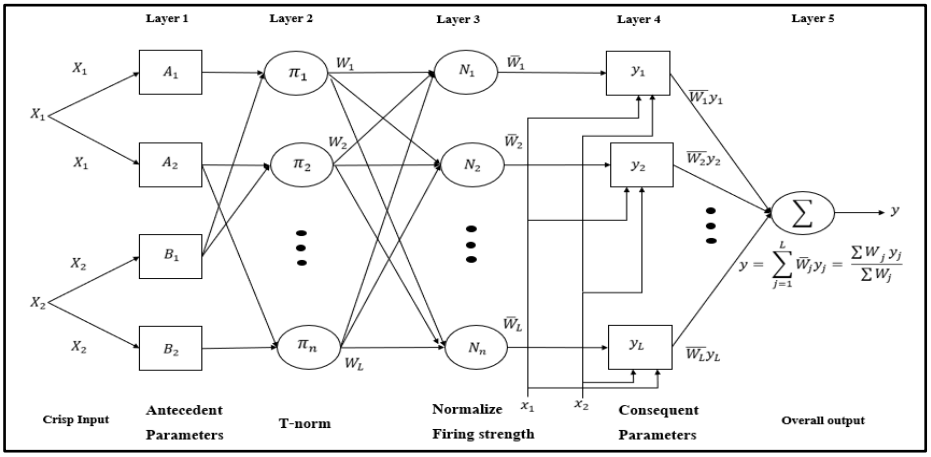

Fig. 3 ANFIS architecture with two inputs, one output

To develop a model for forecasting the GDP growth rate, an Adaptive Neuro-fuzzy Inference System (ANFIS) model will be constructed via MATLAB Program with Fuzzy Logic Toolbox [11] and compare the performance with the other two models to determine the most suitable model for use in the most efficient way, which consists of Linear Regression (LR) and Artificial Neural Network (ANN) models. In the study, the dataset is separated into two datasets (using period 2000–2015) after applying data augmentation technique including 80% of data for training the model and the rest of data for testing the model. Furthermore, an unseen dataset is for checking the model performance including original data in the period 2016-2018. In addition, three models will be established individually, and the parameters will be adjusted to train the models.

ANFIS is characterized by training and testing phases. However, determining the number of epochs, membership function type, and numbers is essential to construct the network and minimizes the modeling error. The shape of the membership function is modified during the training phase in a manner that causes the desired input/output relationship to be learned, and this step is repeated many times (epochs) until obtaining the required converge. This study will consider the appropriate type of membership function based on four types; triangular membership function, trapezoidal membership function, Generalized-bell membership function, and Gaussian membership function. Likewise, the number of input membership function are determined through the trial-and-error method. However, the model developer is free to choose and shape the membership function in accordance with the system demand, simplicity, speed, and convenience.

4) Evaluation

The model performance is evaluated using the statistical metrics, consisting of the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to measure the minimize error in the training dataset, testing dataset, and unseen dataset.

4 Results

The results of the experiment are divided into three parts. The first part is the feature analysis for selecting appropriate inputs in each region. The second part is an experiment to find the type of membership function and the number of input membership functions used in the model structure from the trial-and-error process. The last part is a comparison of model performance for each region's ANFIS model with the LR and ANN models.

4.1 Feature Analysis

The inputs selection for each group of experimental variables was performed using regression analysis. Only one input variable from each group was selected to avoid bias selection. Therefore, each region has a total of 54 sets to test (consisting of 3 variables in education and human resources, 2 variables in innovation system, 3 variables in information infrastructure and 3 other variables: 3x2x3x3 = 54 sets). The results showed that each region had different R-square values and the inputs combination. The selection criteria are; considering the set of variables from the p-value with probability less than 0.1 significance level [12] for at least three variables from each group of inputs along with the highest R^2 value. The best four inputs combination sets for each region are shown in Table 1.

Table 1. Finding the best inputs combination to describe the output

Region	The best 4 inputs combination	R-Square
EAP	Primary, Patents, Internet, Consumption	0.784
ECA	Tertiary, Journals, Mobile, Consumption	0.565
LAC	Primary, Journals, Telephone, Consumption	0.761
MNA	Tertiary, Patents, Internet, Population	0.443
SAS	Primary, Patents, Telephone, Export	0.580

From Table 1, the result showed that, in the East Asia & Pacific and Latin America & Caribbean regions, the best-selected set of variables can be used to describe the variation of GDP growth quite a lot with the R^2 78.4% and 76.1%, respectively, while in the South Asia, Europe & Central Asia and Middle East & North Africa regions, it was found that the best-selected set of variables could be used to describe the variation of GDP growth relatively little with the R^2 58.0%, 56.1% and 44.3%, respectively. The selected variables were then used to construct the model by using neuro-fuzzy technique.

4.2 Finding the Type and the Number of Input Membership Function

The experiments to determine the most appropriate type of the membership functions chosen for this study were divided into four types: the triangular membership function (trimf), the trapezoidal membership function (trapmf), the generalized-bell membership function. (gbellmf) and the Gaussian membership function (gaussmf). The number of membership functions that are selected to be tested using the trial-and-error method consisted of four forms: [2 2 2 2], [3 3 3 3], [2 3 2 3] and [3 2 3 2]. The selection of the most suitable membership function type and the number of input membership functions take measurements from the lowest RMSE.

Table 2. Finding the type and the number of input membership function

Region	No. of MF	trimf	trapmf	gbellmf	gaussmf
EAP	[2 2 2 2]	0.455	0.381	0.430	0.434
	[3 3 3 3]	0.419	0.332	0.393	0.388
	[2 3 2 3]	0.449	**0.327**	0.401	0.427
	[3 2 3 2]	0.427	0.383	0.412	0.411
ECA	[2 2 2 2]	0.550	0.410	0.481	0.494
	[3 3 3 3]	0.370	0.296	0.328	0.361
	[2 3 2 3]	0.470	0.366	0.428	0.413
	[3 2 3 2]	0.409	**0.273**	0.329	0.321
LAC	[2 2 2 2]	0.565	0.447	0.509	0.522
	[3 3 3 3]	0.445	0.351	0.410	0.392
	[2 3 2 3]	0.503	**0.347**	0.439	0.468
	[3 2 3 2]	0.527	0.425	0.504	0.504
MNA	[2 2 2 2]	0.588	0.508	0.555	0.563
	[3 3 3 3]	0.556	0.503	0.507	**0.479**
	[2 3 2 3]	0.587	0.483	0.537	0.578
	[3 2 3 2]	0.561	0.497	0.512	0.532
SAS	[2 2 2 2]	0.439	0.356	0.363	0.347
	[3 3 3 3]	0.314	**0.263**	0.297	0.291
	[2 3 2 3]	0.334	0.265	0.312	0.304
	[3 2 3 2]	0.415	0.324	0.375	0.356

Note: The minimum error value in each data set is shown in bold.

From the Table 2, when comparing the error from the RMSE, it is found that each region had a type, and the number of appropriate membership functions varies. 1) In the East Asia & Pacific and Latin America & Caribbean regions, the results show that the trapezoidal membership function and the number of membership functions [2 3 2 3] are optimal, with the lowest RMSE at 0.327 and 0.347, consecutively. 2) In the Europe & Central Asia region, the result shows that the trapezoidal membership function and the number of membership functions [3 2 3 2] are optimal, with the lowest RMSE at 0.273. 3) In the Middle East & North Africa region, the result shows that the Gaussian membership function and the number of membership functions [3 3 3 3] are optimal, with the lowest RMSE at 0.479. 4) In the South Asia region, the result shows that the trapezoidal membership function and the number of membership functions [3 3 3 3] are optimal, with the lowest RMSE at 0.263. Then, the appropriate model structure is applied to develop the ANFIS model for predicting the GDP growth in each region.

4.3 Model Performance Comparison

After developing all 3 models, the evaluation of model performance for the knowledge-based economy variables and other variables affecting the GDP growth in each region is done by separating the data into 3 sets, consisting of training dataset,

testing dataset, and unseen dataset. The performance of the LR, ANN and ANFIS models are determined by the lowest MAE and RMSE. The results are shown in Table 3.

Table 3. Model Performance Results

		Training		Testing		Checking		Avg. of Testing and Unseen datasets	
Region	**Models**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
EAP	LR	0.579	0.761	0.673	0.883	1.856	1.917	1.265	1.400
	ANN	0.476	0.672	0.582	0.623	0.085	0.107	0.334	0.365
	ANFIS	**0.313**	**0.467**	**0.460**	**0.616**	**0.069**	**0.075**	**0.265**	**0.345**
ECA	LR	1.563	1.931	1.811	2.201	3.915	5.185	2.863	3.693
	ANN	1.407	1.765	1.636	2.029	1.501	1.583	1.569	1.806
	ANFIS	**0.561**	**0.838**	**0.716**	**1.025**	**0.493**	**0.660**	**0.605**	**0.842**
LAC	LR	1.127	1.389	1.023	1.354	**0.823**	**0.930**	**0.923**	**1.142**
	ANN	1.031	1.283	0.929	1.302	1.168	1.401	1.049	1.352
	ANFIS	**0.563**	**0.721**	**0.631**	**0.937**	1.518	1.716	1.075	1.327
MNA	LR	1.057	1.349	1.164	1.410	4.678	5.023	2.921	3.217
	ANN	0.875	1.201	0.908	1.211	4.791	5.991	2.850	3.601
	ANFIS	**0.626**	**0.896**	**0.679**	**0.939**	**1.556**	**1.952**	**1.118**	**1.445**
SAS	LR	0.775	0.955	0.673	0.887	2.628	3.014	1.650	1.951
	ANN	0.599	0.801	0.621	0.790	0.748	0.905	0.685	0.848
	ANFIS	**0.249**	**0.386**	**0.344**	**0.626**	**0.445**	**0.558**	**0.395**	**0.592**

Note: 1) The minimum error value in each data set is shown in bold.
2) The MAE and RMSE values are averaged of original data and augmentation data.

From Table 3, the results demonstrate that the ANFIS model was found to perform the best across all regions since the model yielded the lowest MAE and RMSE values compared to the LR and the ANN models in 14 of 15 experiments from 3 datasets in all 5 regions, while one of experiment from the unseen dataset of the Latin America & Caribbean region, the LR model outperformed the ANN and ANFIS models, consecutively. In addition, the developed ANFIS model is the best predictor of the GDP growth in the East Asia & Pacific region, while it is a weak predictor of the GDP growth in the Middle East & North Africa region. This can be implied that the efficiency of the developed model depends on the relationship of the input and output variables since in the Middle East & North Africa region has the lowest R^2 among 5 regions. In summary, the ANFIS model gives the lowest average values of MAE and RMSE from testing and unseen datasets in 4 regions; in the East Asia & Pacific region: 0.265% and 0.345%, Europe & Central Asia region: 0.605% and 0.842%, Middle East & North Africa: 1.118% and 1.445%, the South Asia: 0.395% and 0.592%, respectively, while in Latin America & Caribbean, it shows the LR model outperforms the others with the average MAE and RMSE of 1.142% and 1.125%, respectively. Fig. 4 illustrates the comparison graph of forecast values obtained from the ANFIS model with the original values.

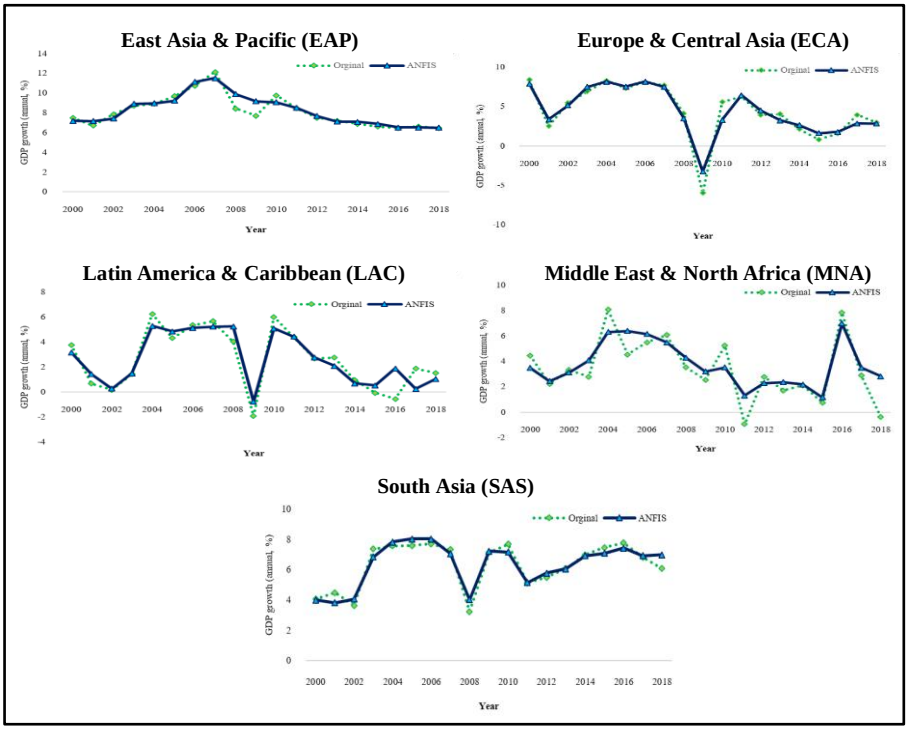

Fig. 4 Comparison of forecast values obtained from the ANFIS model to actual values

5 Discussion and Conclusion

According to its objectives, this study can be able to analyze the relationship between GDP growth rate and knowledge-based economic factors, including the ability to select suitable features to develop a forecast model using Neuro-fuzzy techniques with satisfactory accuracy. Thus, it can be concluded that the knowledge-based economy indicators can be applied as alternative indicators to predict GDP growth rate. Besides, ANFIS model has the best accuracy in forecasting the GDP growth compared to the LR and ANN models.

In terms of improving the efficiency of forecasting in economic work, this study had established the ANFIS model for forecasting GDP growth by applying the knowledge-based economy factors in each category and other factors that would best describe the variation of a GDP growth rate. The results showed that the proposed ANFIS model was found to be the most efficient and accurate for forecasting across all regions with the lowest MAE and RMSE in 14 of 15 experiments, which the performance was consistent with the previous studies [3-5], while the ANN model and the LR model showed less, consecutively. In conclusion, the ANFIS model in this study seemed to predict

well in the East Asia & Pacific, Europe & Central Asia, and South Asia regions. Nevertheless, the error from the ANFIS model of the Middle East & North Africa region was the highest one since the relationship between the selected input variables and output variable was very low. Thus, they might not be good predictors for GDP growth in this case. For the implementation of future work, there are two possible options to improve the performance. The first option is collecting more data in order to have more information to train the model, especially in terms of input variables. Frankly, if the input variables are more closely related, it may affect the model's accuracy performance to improve. As a final point, other techniques for data augmentation should be explored more so that they can be more useful in machine learning.

References

1. A. S. Blinder, "The Internet and the New Economy," Brookings Institution., Washington D.C., Policy Brief Vol. 30, pp. 1-8, June 2000.
2. M. Z. M. Junoh, "Predicting GDP Growth in Malaysia Using Knowledge-based Economy Indicators: A Comparison between Neural Network and Econometric Approaches," Sunway College Journal, Vol.1, pp. 39-50, 2004.
3. M. Mirbagheri, "Fuzzy-Logic and Neural Network Fuzzy Forecasting of Iran GDP Growth," African Journal of Business Management, Vol. 4, Issue. 6, pp. 925-929, 2010.
4. Mladenovic, I., Milovancevic, M., Mladenovic, S. S., Marjanovic, V. and Petkovic, B. (2016). Analyzing and Management of Healthcare Expenditure and Gross Domestic Product (GDP) Growth Rate by Adaptive Neuro-fuzzy Technique. *Computer in Human Behavior*, 64(1), 524-530.
5. I. Mladenovic, M. Milovancevic, S. S. Mladenovic, V. Marjanovic and B. Petkovic, "Analyzing and Management of Healthcare Expenditure and Gross Domestic Product (GDP) Growth Rate by Adaptive Neuro-fuzzy Technique," Computer in Human Behavior, Vol. 64, No.1, pp. 524-530, 2016.
6. Ö. Karahan, "Input-Output Indicators of Knowledge-based Economy and Turkey," Journal of Business, Economics & Finance, Vol. 1, Issue 2, pp. 21-36, June 1, 2012.
7. W. T. Tsaug. Tsaug Package, July 2019 [Online]. Available: https://tsaug.readthedocs.io/en/stable/ [Access July 20, 2021].
8. P. Matias, D. Folgado, H. Gamboa, and A. Carreiro, "Time Series Segmentation Using Neural Networks with Cross-Domain Transfer Learning," *Electronics*, Vol. 10, Issue. 15, pp. 1805, 2021.
9. I. Y. Javeri, M. Toutiaee, I. B. Arpinar, T. W. Miller, J. A. Miller, Improving Neural Networks for Time Series Forecasting using Data Augmentation and AutoML, March 2021 available: https://arxiv.org/abs/2103.01992 [Access July 20, 2021].
10. พยุง มีสัจ, ระบบฟัซซีและโครงข่ายประสาทเทียม. กรุงเทพมหานคร: สำนักพิมพ์มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2553.
11. D. Bystrov and J. Westin, Practice. Neuro-Fuzzy Logic Systems: MATLAB Toolbox GUI. July 2020 [Online]. Available: https://www.researchgate.net/profile/ Mohamed_Mourad _Lafifi/post/how_to_write_Neural_Network_and_ANFIS_MATLAB_code_for_multiple_output [Access July 20, 2021].
12. J. Kim, How to Choose the Level of Significance: A Pedagogical Note, 2015 [Online]. Available: https://mpra.ub.uni-muenchen.de/66373/1/MPRA_paper_66373.pdf [Access July 20, 2021].