# HYBRID MODEL, NEURAL NETWORKS, SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR, AND ARIMA MODELS FOR FORECASTING TOURIST ARRIVALS

**[1]PURWANTO, [2]SUNARDI), [3]FENTY TRISTANTI JULFIA, [4]ADITYA PARAMANANDA**

[1, 2, 3, 4] Universitas Dian Nuswantoro, Semarang, Central Java, Indonesia

E-mail: [1]purwanto@dsn.dinus.ac.id, [2]sunardi@dsn.dinus.ac.id
[3]fentytristanti@gmail.com, [4]myaditya.nanda@gmail.com

## ABSTRACT

An autoregressive integrated moving average (ARIMA) model has been succeed for forecasting in various field. This model have disadvantages in handling the non-linear pattern. Artificial Neural Networks (ANN), Support Vector Machine (SVM) and K-Nearest Neighbor (k-NN) models can be considered to handle non-linear pattern. Neural network, SVM and k-NN models have also succeed for forecasting in various fields and these models yield mixed results of performance. In this paper, we propose a hybrid model combining ARIMA and Artificial Neural Networks model with optimum number of neuron in input layer, optimum number of neuron in hidden layer, optimum of activation function for forecasting tourist arrivals. The forecasting accuracies of the models are compared based on tourist arrivals time series data. The proposed hybrid model yield better forecasting accuracies results compared to ARIMA, K-Nearest Neighbor, neural network and Support Vector Machine with various kernel.

**Keywords:** *Hybrid Model, ARIMA, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine, Forecasting Tourist Arrivals*

## 1. INTRODUCTION

Many forecasting models with different technique have been developed by many researchers to solve their problems. Researchers have made every effort to improve predictive performance. Accuracy of forecasting models depends on the model and depends on the characteristic of the data. But it is very important to determine the best forecasting model based on the characteristic of the data.

Data mining models have been widely used in solving various problems for classification, regression, clustering and forecasting. For forecasting problems, many forecasting models have been proposed by researchers to solve their problems. The ARIMA model has been used to solve tourism problems, namely to estimate the number of tourists. The ARIMA and ARMA model for predicting the number of tourists has been performed in [1] - [3]. However, ARIMA model does not yield satisfactory results for non-linear data. K-Nearest Neighbor technique has been used for forecasting daily air arrivals in Mallorca Island

[4]. Olmedo [5] has compared near neighbor technique with neural network for travel forecasting problem. Ali and Shabri [6] have used SVM and Artificial Neural Network (ANN) for modelling Singapore Tourist Arrivals to Malaysia. Sitohang, Andriyana and Chadidjah [7] has applied SVM for forecasting Tourist Arrivals to Bali. Huang and Hou [8] have implemented Neural Network for tourism demand forecasting. Fernandes et. al. [9] and Claveria [10] have used ANN for forecasting tourism demand. From the results of these research, Neural Network, K-NN, and SVM models yield mixed performance results.

To obtain high forecasting accuracy rate using ARIMA, k-NN, SVM and neural network techniques individually is very difficult. Various types of hybrid models have been implemented for forecasting in the various fields. A hybrid models that combine linear and nonlinear techniques can be expected to produce high predictive accuracy [11]. Purwanto [12] has proposed adaptive hybrid models, and uses ARIMA and Neural models or Neural Networks with ARIMA for health data, and the accuracy of the resulting Hybrid models is

better than individual models. A hybrid models that combine exponential smoothing method and neural network model have been applied to predict financial time series [13. Zhang [14] has used a hybrid model by combining ARIMA and Neural Network and producing better accuracy. All hybrid models have better performance than individual models. Zhang [14] used a hybrid model with a fixed Neural Network configuration. Predictive accuracy in the Neural Network is determined by parameters such as the activation function, number of neuron in input layer and number of neuron in hidden layer.

In this study, we propose a hybrid model that combines ARIMA and Neural Network for forecasting tourist arrivals. The ARIMA model uses the best parameters and the Neural Network model uses the best configuration with optimum number of neuron in input layer, optimum number of neuron in hidden layer, optimum of activation function to improve the performance of accuracy for forecasting tourist arrivals.

## 2. METHODOLOGY

### 2.1  Data Set Used

To evaluate the model, we use the time series tourist arrivals data in Semarang regency, Central Java, Indonesia for the period January 1991 to December 2013. Before using the model, the normalization of the data using Minmax normalization is shown in equation (1).

$$w_i = \frac{x_i - \min(\mathrm{x})}{\max(x) - \min(x)} \qquad (1)$$

where $x= (x1, x2,...,xn)$ is actual data, and $w_i$ is normalized data ($i= 1, 2, ...., n$).

Time series tourist arrivals data after normalization is shown in Table 1 as follow.

*Table 1. Normalized data of the Time series tourist arrivals*

### 2.2  Proposed Method

A hybrid model combining ARIMA and Neural Netwok models, ARIMA, k-NN, Neural Network, and SVM models are implemented for forecasting tourist arrivals. The proposed hybrid model is ilustrated in Figure 1.

*Figure 1. Proposed Method for forecasting tourist arrivals*

The proposed method is used for forecasting tourist arrivals. Step by step of the proposed method can be described as follows:

STEP 1 : In this step, the time series data of the tourist arrivals is collected as show in Tabel 1. We use  univariate time series data that have  two attributes i.e. time (in monthly) and number of tourist arrivals.

STEP 2 : In step 2, the experiments using time series prediction model, namely ARIMA (p, d, q) with different value of p, d, and q. We make pre-processing the time series data for k-NN, Neural Network and SVM models,. Pre-processing time series data is performed as follow:

From time series data of the tourist arrivals data $\{x1, x2, x3, ..., xm\}$, the data is divided into input (independent variables) and target (dependent variable) data [15]. The first pattern consists of $\{x1, x2, ..., xp\}$ as the independent variables and xp+1 as the target.

The second pattern comprises $\{x2, x3, ..., xp+1\}$ as the independent variables and $xp+2$ as the target. The third pattern comprises  $\{x3, x4, ..., xp+2\}$ as the  independent variables  and xp+3 as the target, etc.

STEP 3 : To evaluate the model, we compare performance of models  to obtain the best ARIMA model, Neural Network, k-NN and SVM using the best kernel. We also compute the hybrid model combining the best ARIMA and the best Neural Network configuration with optimum number of neuron in input layer, optimum number of neuron in hidden layer, optimum of activation function. ARIMA model is used to handle linear part of data and Neural Network model is used to handle non-linear part of the data. ARIMA model  uses actual data as input and Neural Network uses residual data as input. Results of hybrid model are total of summing prediction using ARIMA and prediction using Neural Network. In this step, the authors select the parameters p, d, q from ARIMA in determining the best parameters. Whereas in Neural Network, the authors use the Neural Network model with the optimum number of neurons in the input layer, the optimum number of neurons in hidden layer, optimum of activation function.

STEP 4 : Comparison accuracies of forecasting models, namely the best ARIMA, the best SVM, Neural Network and Hybrid model combining the best ARIMA and the best Neural Network configuration is performed. The best model that has the smallest values of RMSE and MSE is used to forecast the tourist arrivals.

## 3. RESULTS AND DISCUSSION

We use the four prediction models for forecasting tourist arrivals. The models are based on ARIMA, Neural Network, *k*-NN, SVM and Hybrid model combining ARIMA and Neural Network. The results obtained are discussed below.

### 3.1 ARIMA model for forecasting tourist arrivals

ARIMA (p, d, q) models with different values of parameters are implemented for forecasting tourist arrivals. The results of the forecasting tourist arrivals using ARIMA models is shown in Table 2.

*Table 2.  Performances of  RMSE  using ARIMA models*

| NO | MODELS | MSE | RMSE |
|----|--------|-----|------|
| 1 | ARIMA (1,1,1) | 0.0161 | 0.127 |
| 2 | ARIMA (12,2,3) | 0.0137 | 0.117 |
| 3 | ARIMA (2,1,2) | 0.0164 | 0.128 |
| 4 | ARIMA (12,1,10) | 0.0144 | 0.120 |
| 5 | **ARIMA (12,1,2)** | **0.0130** | **0.114** |
| 6 | ARIMA (3,1,1) | 0.0161 | 0.127 |
| 7 | ARIMA (5,1,3) | 0.0156 | 0.125 |
| 8 | ARIMA (10,1,5) | 0.0135 | 0.116 |
| 9 | ARIMA (3,1,5) | 0.0161 | 0.127 |
| 10 | ARIMA (10,1,3) | 0.0137 | 0.117 |

From Table 2, it is seen that ARIMA (12, 1, 2) model has the smallest values of RMSE and MSE, so ARIMA (2, 1, 2) model is the best model.

### 3.2 Neural Network for forecasting tourist arrivals

Architecture configurations of Neural Network model with different numbers of neuron in input layer and different numbers of neuron in hidden layer and activation functions are implemented for forecasting tourist arrivals. The experimental using the neural network model with 2, 5, and 7 neurons in input layer and different numbers of neurons in hidden layer were performed to find the smallest value of the RMSE. The results of the MSE and RMSE for forecasting tourist arrivals using the Neural Network model are shown in Table 3.

*Table 3. Results of RMSE using Neural Network model*

Table 3 shown that the smallest value of RMSE and MSE are obtained for the Neural Network model using Hyperbolic Tangent activation function, 5 neuron in input layer and 10 neuron in hidden layer. So, we can conclude that NN (5,10,1) model is the best Neural network configuration.

### 3.3 K-Nearest Neighbor method for forecasting tourist arrivals

The K-Nearest Neighbor method is used to predict tourist arrivals forecasting. We conduct an experiment by specifying the *k* parameter for K-NN. In determining the *k* parameter for K-NN, the input data used is by $x_{t-1}$, $x_{t-2}$, $x_{t-3}$, $x_{t-4}$, and $x_{t-5}$, as inputs, and $x_t$ as target. This test is done to get the best k value for k-NN by looking at the smallest RMSE value. In the experimental process, the neighboring values are used for KNN from k = 1 to 18. Predictive performance using k-NN is shown in Table 4 below.

*Table 4.  Performances of RMSE using K-Nearest Neighbor*

| NO | MODELS | MSE | RMSE |
|----|--------|-----|------|
| 1 | 1- Nearest Neighbor | 0.0331 | 0.182 |
| 2 | 2- Nearest Neighbor | 0.0266 | 0.163 |
| 3 | 4- Nearest Neighbor | 0.0246 | 0.157 |
| 4 | 5- Nearest Neighbor | 0.0237 | 0.154 |
| 5 | 7- Nearest Neighbor | 0.0228 | 0.151 |
| 6 | 8- Nearest Neighbor | 0.0225 | 0.150 |
| 7 | 11- Nearest Neighbor | 0.0222 | 0.149 |
| 8 | 16- Nearest Neighbor | 0.0225 | 0.150 |
| 9 | 17- Nearest Neighbor | 0.0228 | 0.151 |
| 10 | 18- Nearest Neighbor | 0.0228 | 0.151 |

The best k value for k-NN is based on the smallest root mean square error (RMSE) value. Then the best k-NN is 11-Nearest Neighbor with RMSE value of 0.149

### 3.4 Support Vector Machine model for forecasting tourist arrivals

Forecasting tourist arrivals using the SVM model in this study uses kernel dot, radial and polynomial. The parameter value C on SVM uses values of 0.1, 0.2, and 0.3. In the preprocessing process,

univariate time series data is divided by $x_{t-1}$, $x_{t-2}$, $x_{t-3}$, $x_{t-4}$, and $x_{t-5}$, as inputs, and $x_t$ as targets. The following are the results of experiments conducted using the kernel and C parameters on SVM

*Tabel 5. Results of RMSE using SVM model*

From Table 5, it is seen that the smallest value of RMSE is obtained for the SVM using Radial kernel type and C=0.1. We can conclude that the best SVM model is SVM using Radial kernel type and C=0.1.

### 3.5 Hybrid model combining ARIMA and Neural Network

From Table 2, it is seen that ARIMA (12,1,2) has the smallest values of RMSE and MSE. From Table 3, it is shown that NN (5,10,1) with Hyperbolic Tangent activation function has the smallest value of RMSE and MSE. Then, the hybrid model combining ARIMA (12,1,2) and NN (5,10,1) is used for forecasting tourist arrivals. The hybrid model combining ARIMA and Neural Network is shown in Figure 2.

*Figure 2. Hybrid model combining ARIMA and Neural Network*

The RMSE and MSE values of the forecasting tourist arrivals using the hybrid model yields 0.0911 and 0.0083, respectively.

The comparison of actual values and predicted values using proposed hybrid model combining ARIMA(12,1,2) and NN (5,10,1) is shown in figure 3 as follow:

*Figure 3. Comparison actual and forecasting tourist arrivals values*

From figure of comparison actual and forecasting tourist arrivals values, it is seen that the forecasting tourist arrivals values using the hybrid model combining ARIMA (2, 1, 2) and NN (7, 12, 1) are very close to the actual values

### 3.6 Comparison of models and Discussion

We have conducted experiments using ARIMA, k-NN, SVM, Neural Network and Hybrid models to predict tourist arrivals. Based on Table 2, Table 3, Table 4, Table 5 and the performance results of hybrid model, we can make comparison of models performance as shown in figure 4 as follow:

*Figure 4. Comparison of Performances*

From Figure 4, the RMSE and MSE values of the proposed hybrid model are smallest. It is seen that the performance of hybrid model is better than performance of ARIMA, *k*-NN, SVM and Neural Network. So that the best model for forecasting tourist arrivals is the hybrid model combining ARIMA (12, 1, 2) and Neural Network (5,10,1) with Hyperbolic Tangent activation function. Zhang [14] used hybrid model combining ARIMA and Neural Network with NN(4,4,1), NN(7,5,1) and NN(7,6,1). Based on Table 3, Neural Network (5,10,1) with Hyperbolic Tangent activation function yields the best performances result compared to NN(4,4,1), NN(7,5,1), NN(7,6,1) and the other configuration of Neural Network. So, the proposed hybrid model uses ARIMA and Neural Network (5,10,1) with Hyperbolic Tangent activation function for forecasting tourist arrivals.

### 4. CONCLUSION

This paper has discussed the proposed hybrid model for forecasting tourist arrivals. The hybrid model combines ARIMA and Neural Network with using the best configuration with optimum number of neuron in input layer, optimum number of neuron in hidden layer, optimum of activation function. The hybrid model is evaluated using time series tourist arrivals data. RMSE and MSE values have been employed for comparison of the models performance. From the results, it is found that the hybrid model combining ARIMA (12,1,2) and Neural Network (5, 10,1) with Hyperbolic Tangent activation function yields the best forecasting result compared to ARIMA, Neural Network, k-NN, SVM models.

**REFRENCES:**

[1] Choden and Suntaree Unhapipat, ARIMA model to forecast international tourist visit in Bumthang, Bhutan , *IOP Conf. Series: Journal of Physics: Conf. Series 1039*, 2018.

[2] Loganathan, Nanthakumar and Yahaya Ibrahim, Forecasting International Tourism Demand in Malaysia Using Box Jenkins SARIMA Application, *South Asian Journal of*

*Tourism and Heritage*, Vol. 3, Number 2, 2010.

[3]  Chia-Lin Changa, Songsak Sriboonchitta, Aree Wiboonpongse, Modelling and forecasting tourism from East Asia to Thailand under temporal and spatial aggregation, *Mathematics and Computers in Simulation*,79, 2009, pp. 1730–1744.

[4]   Marcos álvarez Díaz, Josep Mateu-Sbert, Forecasting Daily Air Arrivals in Mallorca Island Using Nearest Neighbour Methods, *Tourism Economics*, (2011), Volume: 17 issue: 1, 2011, pp. 191-208

[5]  Elena Olmedo , Comparison of near neighbour and neural network in travel forecasting, *Journal of Forecast*. 35, 2016, pp. 217–223

[6]  Rafidah Ali and Ani Shabri, Modelling Singapore Tourist Arrivals to Malaysia by Using SVM and ANN, *SCIREA Journal of Mathematics*, 2017

[7]  Yosep Oktavianus Sitohang, Yudhie Andriyana, Anna Chadidjah, The Forecasting Technique Using SSA-SVM Applied to Foreign Tourist Arrivals to Bali, *TELKOMNIKA*, 16 (4), 2018,

[8]  Han-Chen Huang and Cheng-I Hou, Tourism Demand Forecasting Model using Neural Network, *International Journal of Computer Science & Information Technology (IJCSIT)* Vol 9, No 2, 2017.

[9]  Paula Odete Fernandes, João Paulo Teixeira, João Matos Ferreira, and Susana Garrido Azevedo, Forecasting Tourism Demand With Artificial Neural Networks, *International Conference On Tourism & Management Studies –Algarve*, 2011

[10] Oscar Claveria, and Salvador Torra, Forecasting tourism demand to Catalonia: Neural networks vs. time series models, *Economic Modelling*, Volume 36, January 2014, pp.220-228

[11] Durdu Ömer Faruk, A Hybrid neural network and ARIMA model for water quality time series prediction, *Engineering Applications of Artificial Intelligence*, 23, 2010, pp. 586–594

[12] Purwanto, Eswaran, C., and Logeswaran, R. (2010). Adaptive Hybrid Algorithm for Time Series Prediction in Healthcare. *Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation* (CIMSiM, Bali, Indonesia (IEEE), 2010, pp. 21-26.

[13] Kin Keung Lai, Lean Yu, Shouyang Wang, and Wei Huang, Hybridizing exponential smoothing and neural network for financial time series predication. *Proceedings of the 6th International Conference Computational Science*, 2006,pp. 493 – 500

[14] G. Peter Zhang, Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 2003, pp. 159-175

[15] Purwanto, C. Eswaran and R. Logeswaran, Improved Adaptive Neuro-Fuzzy Inference System for HIV/AIDS Time Series Prediction. In: *Informatics Engineering and Information Science*, 253, Springer-Verlag Berlin Heidelberg , 2011, pp. 1-13

*Table 1. Normalized data of the Time series tourist arrivals*

| No | Month | Year | | | | | |
|----|-------|------|------|------|------|-----|------|
| | | 1991 | 1992 | 1993 | 1994 | ... | 2013 |
| 1 | January | 0.33789 | 0.30510 | 0.22587 | 0.15938 | ... | 0.25319 |
| 2 | February | 0.09199 | 0.27140 | 0.14663 | 0.11384 | ... | 0.14208 |
| 3 | March | 0.07650 | 0.19490 | 0.17395 | 0.06466 | ... | 0.22769 |
| 4 | April | 0.07286 | 0.30419 | 0.31421 | 0.24954 | ... | 0.15938 |
| 5 | May | 0.05464 | 0.28233 | 0.39162 | 0.34608 | ... | 0.17213 |
| 6 | June | 0.11020 | 0.39891 | 0.40710 | 0.26321 | ... | 0.26594 |
| 7 | July | 0.31330 | 0.50273 | 0.59927 | 0.55647 | ... | 0.29326 |
| 8 | August | 0.24226 | 0.56011 | 0.43260 | 0.58379 | ... | 0.38251 |
| 9 | September | 0.23588 | 0.51002 | 0.39617 | 0.32696 | ... | 0.33607 |
| 10 | October | 0.23406 | 0.38798 | 0.31148 | 0.28871 | ... | 0.22222 |
| 11 | November | 0.18944 | 0.29872 | 0.21585 | 0.11840 | ... | 0.28689 |
| 12 | December | 0.30146 | 0.30783 | 0.17031 | 0.12113 | ... | 0.31785 |
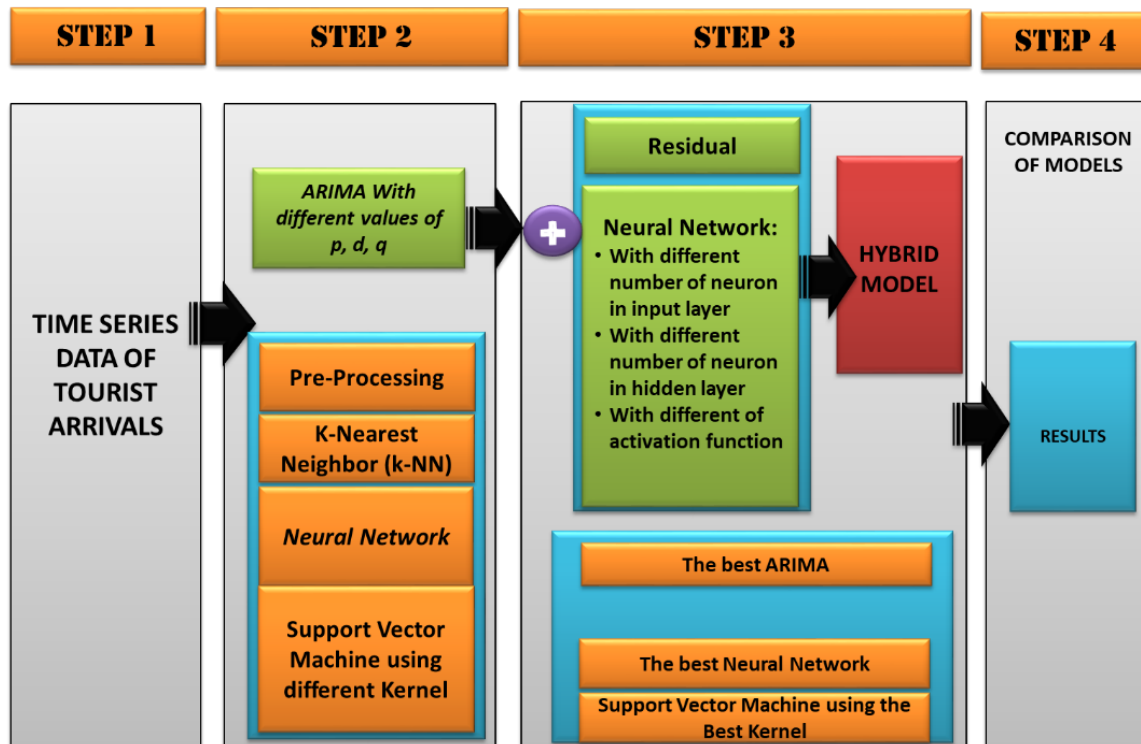


*Figure 1. Proposed Method for forecasting tourist arrivals*

*Table 3. Results of RMSE and MSE  using Neural Network model*

| No | Models | Activation Function | MSE | RMSE |
|---|---|---|---|---|
| 1 | NN(2,5,1) | Hyperbolic Tangent | 0.01537 | 0.12398 |
| 2 | NN(2,5,1) | Bipolar Sigmoid | 0.01625 | 0.12748 |
| 3 | NN(2,6,1) | Hyperbolic Tangent | 0.01641 | 0.12810 |
| 4 | NN(2,6,1) | Bipolar Sigmoid | 0.01615 | 0.12708 |
| 5 | NN(2,7,1) | Hyperbolic Tangent | 0.01657 | 0.12872 |
| 6 | NN(2,7,1) | Bipolar Sigmoid | 0.01622 | 0.12736 |
| 7 | NN(2,10,1) | Hyperbolic Tangent | 0.01664 | 0.12900 |
| 8 | NN(2,10,1) | Bipolar Sigmoid | 0.01596 | 0.12633 |
| 9 | NN(4,4,1) | Hyperbolic Tangent | 0.01603 | 0.12661 |
| 10 | NN(2,15,1) | Bipolar Sigmoid | 0.01583 | 0.12582 |
| 11 | NN(5,5,1) | Hyperbolic Tangent | 0.01183 | 0.10877 |
| 12 | NN(5,5,1) | Bipolar Sigmoid | 0.01188 | 0.10900 |
| 13 | NN(5,6,1) | Hyperbolic Tangent | 0.01114 | 0.10555 |
| 14 | NN(5,6,1) | Bipolar Sigmoid | 0.01172 | 0.10826 |
| 15 | NN(5,7,1) | Hyperbolic Tangent | 0.01128 | 0.10621 |
| 16 | NN(5,7,1) | Bipolar Sigmoid | 0.01240 | 0.11136 |
| **17** | **NN(5,10,1)** | **Hyperbolic Tangent** | **0.01101** | **0.10493** |
| 18 | NN(5,10,1) | Bipolar Sigmoid | 0.01143 | 0.10691 |
| 19 | NN(7,5,1) | Hyperbolic Tangent | 0.01524 | 0.12345 |
| 20 | NN(7,6,1) | Hyperbolic Tangent | 0.01455 | 0.12062 |

*Tabel 5. Results of RMSE and MSE  using SVM model*

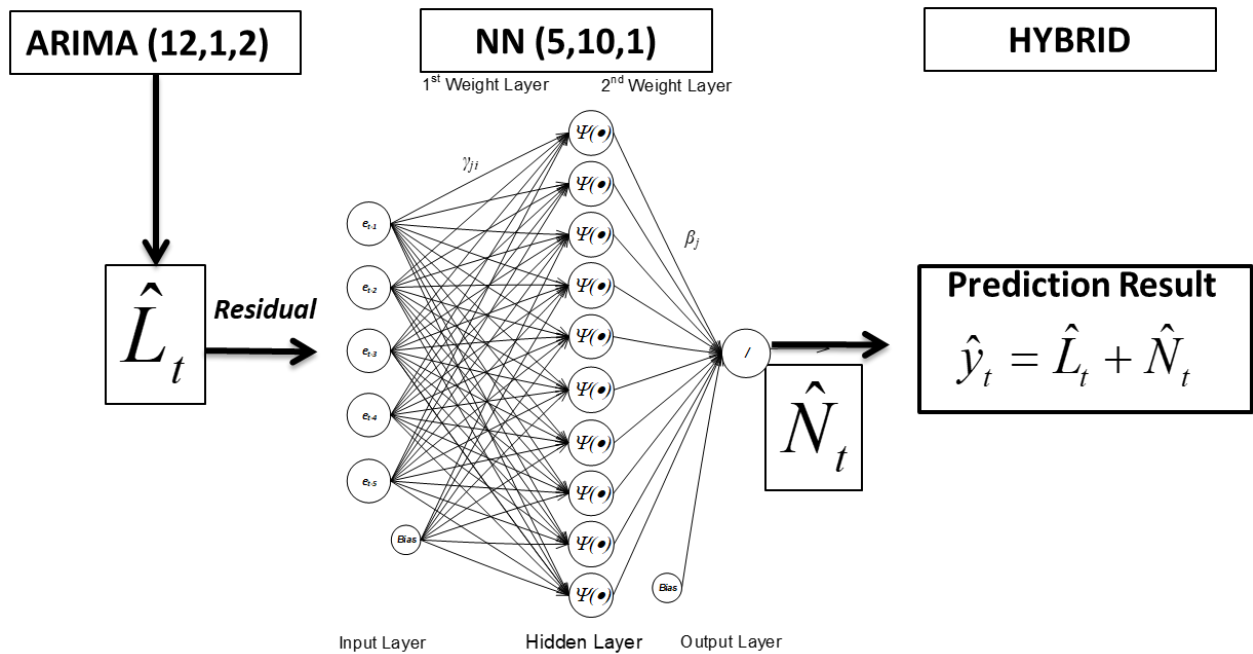| KERNEL TYPE | Paramaeter (C) | | | | | |
|---|---|---|---|---|---|---|
| | 0.1 | | 0.2 | | 0.3 | |
| | MSE | RMSE | MSE | RMSE | MSE | RMSE |
| Dot | 0.3249 | 0.570 | 1.2432 | 1.115 | 2.7324 | 1.653 |
| Radial | 0.0286 | 0.169 | 0.0289 | 0.170 | 0.0289 | 0.170 |
| Polynomial | 0.0595 | 0.244 | 0.0408 | 0.202 | 0.0408 | 0.202 |
| Neural | 5.3592 | 2.315 | 20.3221 | 4.508 | 43.9304 | 6.628 |

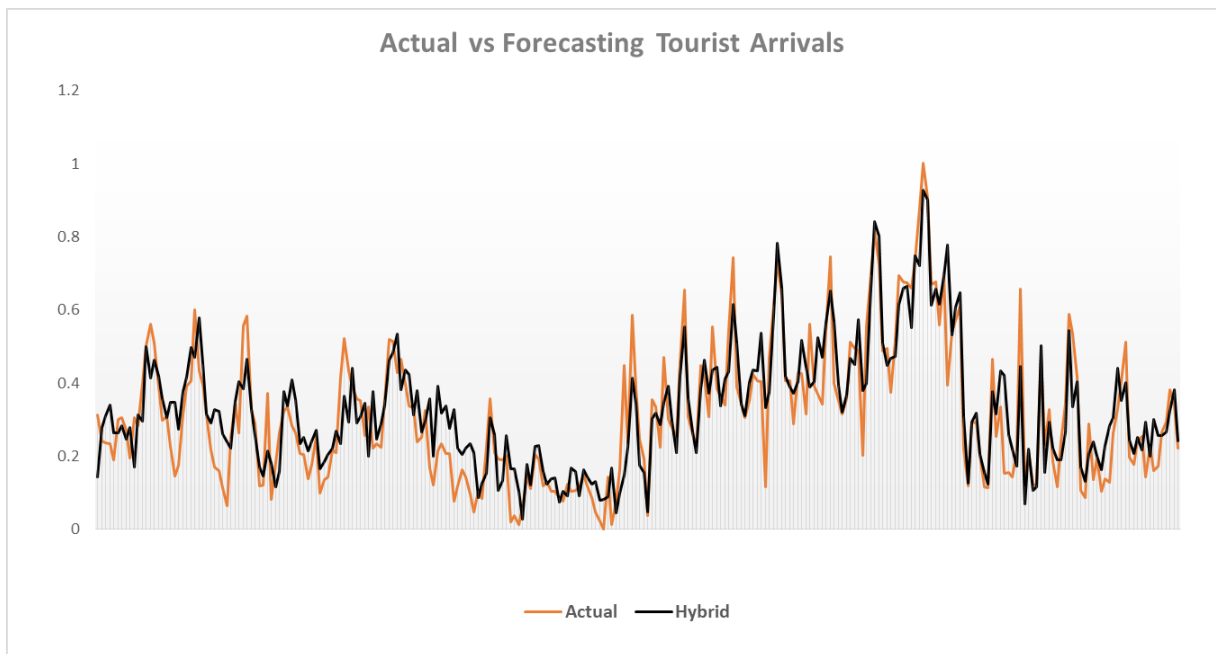*Figure 2. Hybrid model combining ARIMA and Neural Network*
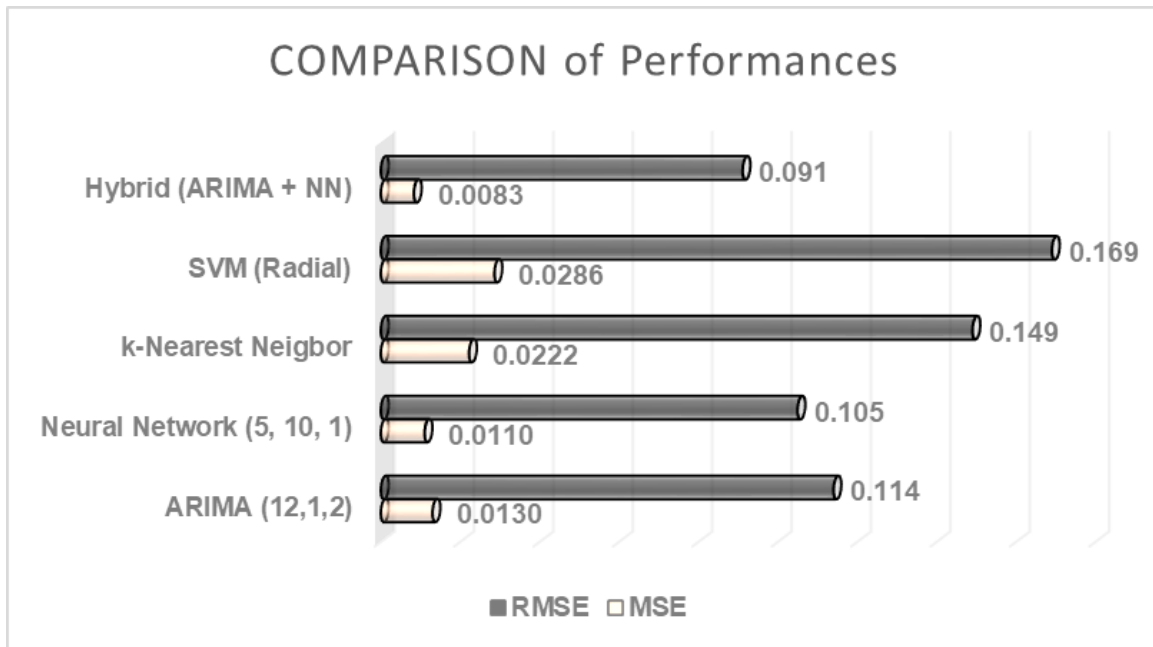


*Figure 3. Comparison Actual And Forecasting Tourist Arrivals Values*

*Figure 4. Comparison Of Performances*