Machine Learning based Forecasting Systems for Worldwide International Tourists Arrival

Ram Krishn Mishra[1]
Department of Computer Science
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates 345055

Siddhaling Urolagin[2]
Department of Computer Science
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates 345055

J. Angel Arul Jothi[3]
Department of Computer Science
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates 345055

Nishad Nawaz[4]
Department of Business Management,
College of Business Administration
Kingdom University, Riffa, Kingdom of Bahrain

Haywantee Ramkissoon[5]
College of Business, Law, and Social Science,
Derby Business School,
University of Derby, Derby, United Kingdom

Abstract—**The international tourist movement has overgrown in recent decades, and travelers are considered a significant source of income to the tourism economy. When tourists visit a place, they spend considerable money on their enjoyment, travel, and hotel accommodations. In this research, tourist data from 2010 to 2020 have been extracted and extended with depth analysis of different dimensions to identify valuable features. This research attempts to use machine learning regression techniques such as Support Vector Regression (SVR) and Random Forest Regression (RFR) to forecast and predict worldwide international tourist arrivals and achieved forecasting accuracy using SVR is 99.4% and using RFR is 84.7%. The study also analyzed the forecasting deadlock condition after covid-19 in the sudden drop of international visitors due to lockdown enforcement by all countries.**

Keywords*—*Tourists; forecasting; machine learning; Covid-19

I. Introduction

The tourism industry plays a significant role in economic development, with several countries focusing on building the best possible policies for international travelers. Tourism is playing a significant role in contributing to multi-dimensional economic growth [1]. Multiple business sector economies across the globe rely on tourism to create employment opportunities, improve infrastructure, and foster cultural interchange between visitors and residents. Tourism can reap more benefits through a multi-stakeholder engagement approach[2]. Tourists rely on local transportation, accommodation, food and beverage, entertainment, and very importantly, visitors may want to buy new things which are not available in their local places. Such transactions contribute to mobilization of the local economy. hence contributing to the local economy. According to a World Tourism Organization (WTO) study in 2020, the percentage of people who travel for enjoyment as family and solo trips have increased from 50% in 2000 to 55% in 2019 [3]. The revenue generated from international tourists' arrivals can help the Country's economy and significantly contribute to balance payment of downgraded sectors such as unemployment, transportation, and healthcare [4].

One of the main motives for tourists to travel is to visit a new place to escape the monotony of boring routine life. The solo or family trip helps ease stress and get a unique environment for a happier and healthier experience. The host country aims to provide the best possible facilities to tourists even when there are high-demand referrals. The forecasting system can help host countries prepare for the tourist requirements well in advance. Forecasting is a technique for creating accurate and optimize predictions[5] based on previous data. Fig. 1 shows the process of forecasting Systems. Many business stakeholders adopt the forecasting for various variables, including projecting future costs, quantity, or planning the budget. The major problems researchers face for developing a forecasting system is to collect the actual data. Two sources are there to gather the data, the first primary source contains first-hand information gathered directly by the organization. The data is generally collected by various surveys, focus groups, or interviews and direct methods of obtaining data make it more reliable and accurate to build the systems. Second, secondary sources are data that has already been collected and processed by a third party. The forecasting process is sped up by receiving data in a well-organized and compiled format.

A tourism forecasting system helps administration in planning and arranging essential things for tourists. With rapid infrastructure, economy, and politics changes, forecasting systems help to get things done on prior deadlines. Government organization and associated stakeholders which are involved in tourism planning required highly accurate forecasting system. With the help of forecasting system, they can adopt the required changes in much better and faster way. When there is no availability of highly accurate forecasting systems, these organization face difficulties[6]. In simple words, the meaning is, to minimize the possibility of the decision failing to attain the coveted goals. Hence an accurate prediction is very essential to the government.

Machine learning methods have attracted significant attention in tourism research [7] for better results than traditional approaches. Some machine learning methods like Neural Networks (NN) and SVR play a big role in forecasting time. Most of the techniques applied in prediction and tourism modelling are categorized into four categories: time series model, econometrics model, Artificial Intelligent (AI) techniques, and qualitative methods. AI techniques have been applied across

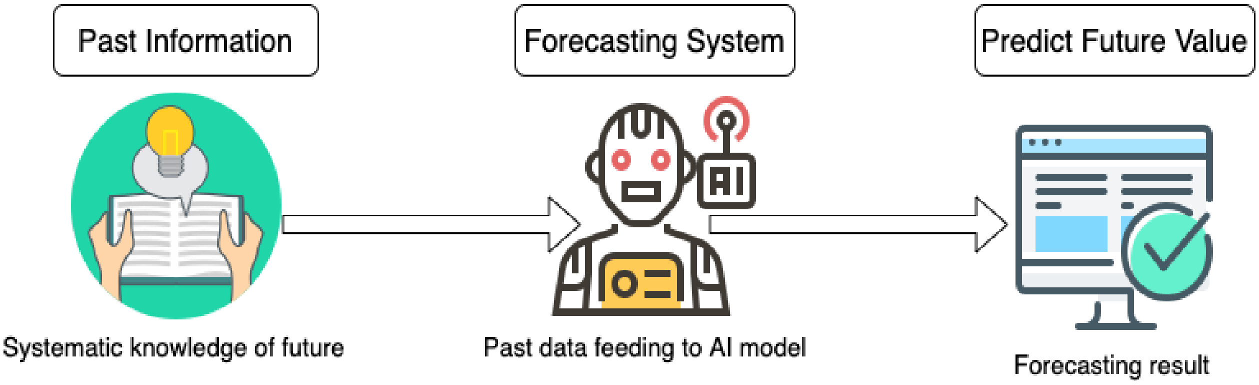


Fig. 1. Forecasting Systems.

several domains and in a variety of data structures.

In the current study, the ten-years tourists arrival data have been collected for developing forecasting systems. Many countries can use the proposed methods for tourist arrival forecasting for arranging the required facilities. This can inform tourism policy by forecasting tourism revenue. The forecast can assist the government in creating temporary job opportunities in the tourism sector for a particular period in that year. The advantage is that it can promote seasonal work in tourism and assist those whose livelihoods depend on tourism. Most of the work in tourism has been focused on domestic tourism forecasting [8]–[10]. Due to the rare availability of data, there are existing challenges in developing proper tourist forecasting systems. This study develops a worldwide tourist forecasting system by applying machine learning techniques such as SVR and RFR. The machine learning methods [11], [12] have been tested with different kinds of feature selection techniques and clear identification of attributes before feeding into the model. Developed tourist forecasting systems will help to analyze the flow of tourists internationally and in the host country. This system will also help to identify the transport traffic and facilitate the number of flights between two countries, arranging or extending local transport systems and analyzing the required number of rooms in hotels. The COVID-19 pandemic had sudden travel restrictions across borders. The year 2020 was the worst in tourism history in international tourist arrivals. The SARS-COV-2 virus has led to a setback in the current forecasting Systems. In this study, we retrieve the forecasting data and compare the results with the current covid-19 scenario.

A tourism forecasting system helps to plan and arrange essential things for tourists. With rapid infrastructure, economy, and politics changes, forecasting systems help to get things done on prior deadlines. Government organization and associated stakeholders which are involved in tourism planning required highly accurate forecasting system. With the help of forecasting system, they can adopt the required changes in much better and faster way. When there is no availability of highly accurate forecasting systems, these organization face difficulties[6]. In simple words, the meaning is, to minimize the possibility of the decision failing to attain the coveted goals. Hence an accurate prediction is very essential to the government.

Machine learning methods have attracted significant attention in tourism research [7] for better results than traditional approaches. Some machine learning methods like Neural Networks (NN) and SVR play a big role in forecasting time. Most of the techniques applied in prediction and tourism modelling are categorized into four categories: time series model, econometrics model, Artificial Intelligent (AI) techniques, and qualitative methods. AI techniques have been applied across several domains and in a variety of data structures.

In the current study, we have collected ten-year prior visitor history data to develop forecasting systems. Many countries can use the proposed methods for tourist arrival forecasting for arranging the required facilities. This can inform tourism policy by forecasting tourism revenue. The forecast can assist the government in creating temporary job opportunities in the tourism sector for a particular period in that year. The advantage is that it can promote seasonal work in tourism and assist those whose livelihoods depend on tourism. Most of the work in tourism has been focused on domestic tourism forecasting [8]–[10]. Due to the rare availability of data, there are existing challenges in developing proper tourist forecasting systems. This study develops a worldwide tourist forecasting system by applying machine learning techniques such as SVR and RFR. The machine learning methods [11], [12] have been tested with different kinds of feature selection techniques and clear identification of attributes before feeding into the model. Developed tourist forecasting systems will help to analyze the flow of tourists internationally and in the host country. This system will also help to identify the transport traffic and facilitate the number of flights between two countries, arranging or extending local transport systems and analyzing the required number of rooms in hotels. The COVID-19 pandemic had sudden travel restrictions across borders. The year 2020 was the worst in tourism history in international tourist arrivals. The SARS-COV-2 virus has led to a setback in the current forecasting Systems. In this study, we retrieve the forecasting data and compare the results with the current Covid-19 scenario.

II. Literature Survey

A forecasting system for tourism will provide direct and indirect benefits to the government, society, people, business, services, and economy of the country. Tourism contributes

TABLE I. COMPARISON OF MODELS AND USED REGIONS

Reference Number	Region Focused	Research Objects	Data Frequency	Methodologies	Performance Measure	Variables
[13]	Las Vegas	Tourism Demand	Monthly	Logistic Growth Regression	MAPE, RMSPE, DM.	Tourist Arrivals
[14]	Taiwan	Inbound Tourism	Half Yearly,Annually	SARIMA- GARCH	MAPE, MAD, RMSE.	Tourist Arrivals
[15]	Hong Kong	Inbound Tourists	Monthly	Sparse GPR	MAE, MAPE, MSE.	Tourist Arrivals
[16]	Taiwan	Outbound Tourism	Monthly	SARIMA	MAPE	Tourist Arrivals
[17]	South Tyrol	Tourism Demand	Monthly	SARIMA.	MAPE, R.M.S.E.,M.S.E., MAD	Tourist Arrivals

to GDP, employment, visa services, and tourism-related businesses. Given the significant positive impacts of tourism, performing the prediction on the number of tourist visitors, the time, when tourists visit the places, the duration of tourist's visits will provide crucial information to the government. Researchers are keen to develop an accurate forecasting system and to find a novel approach to deal with different sizes of data datasets. The seasonal ARIMA, v-Support Vector Regression and Multi-Layer Perceptron (MLP) Neural Networks models were applied on monthly data for the tourist arrival in Turkey and proposed an approach to select the model in a given time series [11]. Combined techniques have been discussed to predict tourism demand [18] . The authors combined ACF, NN, and Genetic Algorithms (GA) to perform the classification. A framework has been suggested based on the Generalized Dynamic Factor Model (GDFM) to generate the composite search index[19]. It has improved the forecast accuracy as compared to the traditional time series model and Principal Component Analysis (PCA) model. Decomposition based on eigen were used to reduce the dimensions [20] in time series data prediction. Wang Jun et.al. [21] have proposed the forecasting model by combining ANN and a clustering algorithm and compared this model with other ANN-based and ARIMA model; this model performed better than other related methods. For the multisource data and passenger flow volume, authors have proposed a new algorithm by merging the non-linear, genetic algorithm and S.V.R. Karo Solat, et al.[22], have used elliptically symmetric principal components for predicting exchange rates. Forecasting data belongs to the regression category; researchers have applied the methods such as regression, the Delphi method, moving average models, ARIMA, MLP, GRNN, radial bias function (RBF) among others. Shaolong Sun et.al. have developed a tourist arrival forecasting model. One of the most widely used time series forecasting models is the ARIMA. However, the latter does not perform better with multi-source data [23]. The authors proposed the Kernel Extreme Learning Machine (KELM) models to improve the forecasting accuracy and robustness analysis on the Baidu Index and Google Index data. According to the authors in [24], the most used time series analysis model for the prediction of tourist arrivals is ARIMA and was used extensively in the last few years. Authors used Seasonal Autoregressive Integrated Moving Average (SARIMA) with Generalised Autoregressive Conditional Heteroskedasticity (GARCH) to forecast tourist arrivals in Taiwan[25]. In [26], the authors have used SARIMA to predict the demand for traveling by air. Hence, all these studies, research, and work done demonstrated that enhanced ARIMA models lead to better predictions.

Accurate tourist forecasts are essential because they provide crucial information to tourism practitioners and academics when making decisions about resource allocation, priority, and risk assessment. Based on an extant review of the literature[27], prediction methods in tourist arrivals can be categorized into Machine Learning (ML) models and techniques of time series analysis. With reference to model building, tourism demand prediction studies depend strongly on variables that are input to the model[28]. These variables are supposed to be strongly connected to tourism demand, with no missing or incorrect values. Tourism demand prediction components can be defined in several ways using various parameters. They can be classified into indicators and determinants, depending on the relationship with tourism demand establish ML methods for estimating the number of tourists coming to Turkey. In their work, Linear Regression and NN-MLP are implemented to create multivariate tourism predictions for Turkey. They compare performances of the predictions in the context of Relative Absolute Error (RAE), Root Relative Squared Mean (RRSE) and Correlation Coefficient (R) measurements depicting MLP for regression produces enhanced performance. Extensively used ML models consist of Artificial Neural Networks (ANN) and SVR Authors in [29] have used the method SVR and ”Fly Optimization Algorithm (FOA)” together for predicting tourism arrivals. In [13], a prediction model has been suggested, which amalgamates ”Back-Propagation Neural Network (BPNN)” and ”Empirical Mode Decomposition (EMD)”. This model foretells how many tourists will visit the place. Li et al. [14] enhanced BPNN by incorporating the PCA and DE (ADE) algorithm to predict how many tourists are willing to visit the place in the future. Outcomes of the work in [13] and [14] revealed that enhanced BPNN was outperforming ARIMA Authors in [30] created a new structural NN model, forecasting the number of tourists willing to visit the place in the future. The outcomes demonstrated that none of the models were superior in any of the situations.

Fernandes et al. state that Artificial Intelligence (AI) has played an essential role in attaining outstanding applications in predicting the demand of tourists in the region. Despite that, many of the AI methods used till now are not deep architectures. They have little ability for researching greater non-linearities, especially when data is big-scale and vague patterns [31]. The authors have come up with a new deep learning technique called the ”Stacked Autoencoder” with ”Echo-State Regression (SAEN) which helped in predicting demand for tourism [32]. SAEN is employed in four different tourism situations and the outcome of the prediction reveals that SAEN is better than the standard models. A big data based system for tourism forecasting is proposed [33]. The authors have included leading indicators such as price index which improved the performance of the model. In [34], the authors present the Real-Value Genetic Algorithm (RGA) to specify the available parameter of SVR, called GA-SVR. It optimizes each parameter of SVR at the same time from training data. Afterward, they forecast tourism demand in China. Moreover, they carried out a comparison between BPNN and

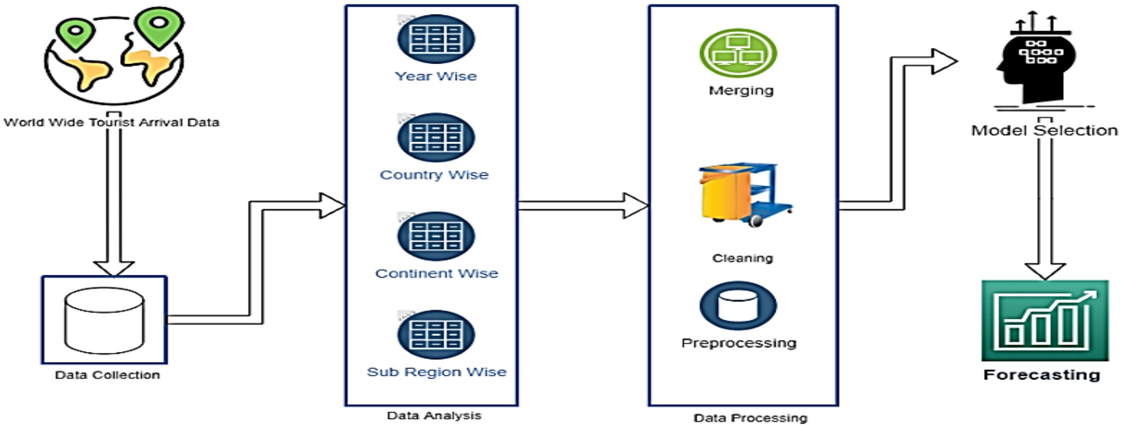


Fig. 2. Tourist Forecasting Systems.

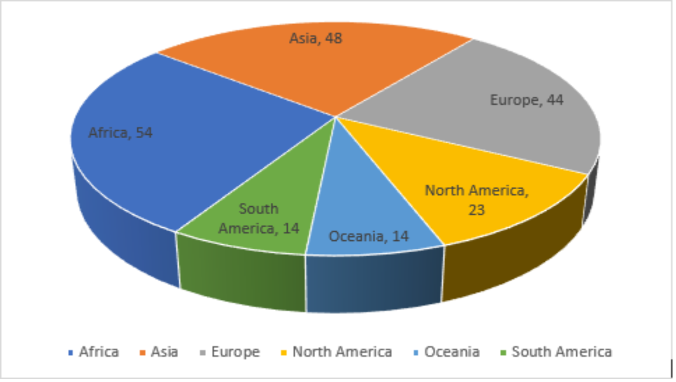


Fig. 3. Continent and Associated Number of Countries.

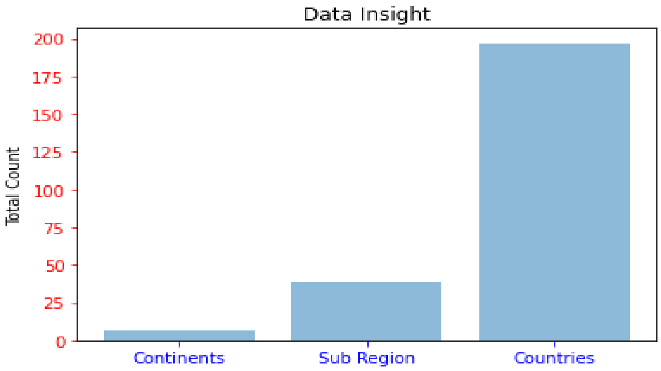


Fig. 4. Data Insights.

time series models. This comparison helped them know that SVR has a good predicting capability. Moreover, the authors have mentioned eight sections. One section presents studies associated with SVR, while another section summarizes the current methods to the option of hyperparameters. Another section details the GA-SVR technique and the rest of the sections deal with the analysis of outcomes, the origin of the data, etc. Various NN models were developed in [34] on cross learning to predict time series data. The principal components of prediction are ”Determinants”. Traditional economic ideas, like ”consumption behavior theory” and ”utility theory” indicate that factors, for example, cost, earning, and publicising affect the demand of tourist arrivals. It can be seen [1], [16] [16][29] to have a complete examination of tourism demand prediction studies. All these studies notice that the functioning of predicting models differs based on various considerations, for example, the data's frequency, prediction horizon's length, the source countries, and the destination. What made us concentrate on the data-driven methods in accordance with ML, is the dearth of agreement about the most correct model to predict tourism demand. In [35], the authors notice that the technique of SVM based is much signified to deal with traits of the tourist's data. They contrast and differentiate the predicting accuracy of various ML models to ARMA using month-wise tourist visits from 13 countries coming to Hong Kong. They considered the years from 1985 to 2008, and from their work, they obtained the best correct results with ML techniques. The requirement for further correct predictions had given rise to more dependency on ML models to get better-sophisticated forecasts of tourists.

In [29], the authors have employed a ”Rough Sets Approach” to predict demand for tourism in Hong Kong from the US and UK Gaussian Process Regression (GPR) has been used in past years for prediction. It is a supervised learning approach followed by generalized linear regression to forecast data locally. To bridge the gap, the authors in [13], [19] have designed a prediction test to contrast GPR to NN and SVR. Their primary objective behind this study was to examine the relative advancement of ML techniques' prediction accuracy through a linear stochastic procedure employing two substitute approaches. First is the direct one, which predicts the aggregate series. The second approach is using the same models to predict the particular series for the regions one by one. Finally, the predicting performance of both methods was compared. Weather forecasting can be applied using Deep Learning (DL) techniques also. DL can be used in various fields like entertainment, visual recognition, and including forecasting such as tourism forecasting, forecasting stock prices, etc. The authors have compared the performance of the prediction of ”Recurrent Neural Network (RNN)”, ”Conditional Restricted

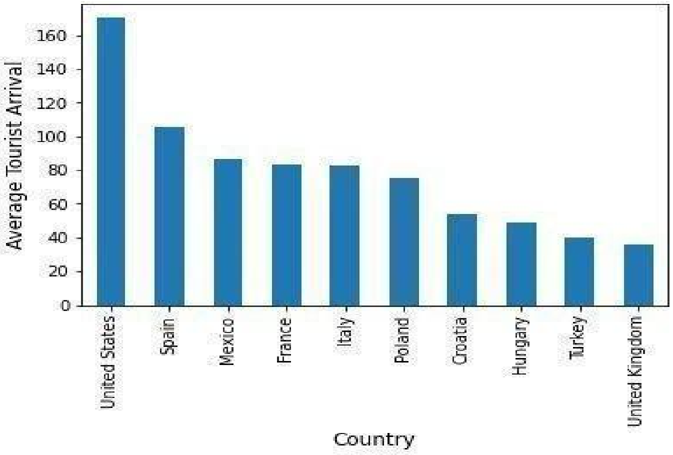


Fig. 5. Average Tourist Arrival in Top 10 Most Visited Countries.

Boltzmann Machines (CRBM)" and "Convolution Neural Network (CNN)" [15]. Authors in [13] give an introduction to principles of ANN and they also provide a stage-by-stage guide to methods they have applied for building a NN for predicting tourist arrivals. They have involved many rules and have included some points of discussion among the authors to apply ANN effectively. A comparison of various forecasting methods is shown in Table I.

III. Methodology

This research proposes tourist forecasting systems in Fig. 2. To predict international tourist arrivals, the methods adopted are data collection from globally trusted sources, followed by data analysis, data processing, and the creation of a machine learning model. The machine learning techniques include SVR and RFR.

1) Data Collection and Analysis: This research draws on historical data to tackle the forecasting challenges and develop the predictive model. A substantial amount of data gathered by the government or other public entities is made available. These data sets are referred to as public data since they do not require specific authorization to use them. The data is gathered from reliable online sources and official tourist websites of countries. This dataset contains tourist arrivals for most of the countries between 2010 to 2020. Since data for nearly 13 countries are not available on the internet, those countries are not included in proposed forecasting system. In addition, the tourism industry suffered greatly because of Covid-19 in terms of visitor arrivals. As a result, data for 2020 is not available for most of the countries. Whatever data have been found for year 2020 been used for Covid-19 analysis in respective to international tourist arrivals.

Fig. 3 shows the number of countries on a particular continent. Fig. 4 depicts the data insights of number of continents, sub-regions and countries.Africa has the highest number of countries, i.e. 48 and South America has the lowest, i.e. 12 countries. Continent wise number of countries are Africa = 48, Asia = 45, Europe = 43, North America = 23, Oceania = 13, South America = 12.

Fig. 5 shows a graphical representation of average tourist arrivals in the top 10 countries. The United States has the highest number of tourist arrivals while the United Kingdom comes in number 10. These top 10 countries decide the flow of international tourists and make common global tourism policies.

Year vs average analysis essentially explains the distribution of arrival data as well as the year-by-year data association of annual arrivals. In most cases the year-wise data is matching with average data, there are not many changes in tourist arrivals as depicted in Fig. 6. scatter plot depicts the annual data points distribution and it can be seen Fig. 7(a) and 7(b) that data is not equally aligning in years 2011 and 2012.

Correlation matrix shows the relationship between two variables which is shown in Fig. 8. If the variables are highly correlated then the value will be closer to 1. In Table the data is ranging from year 2010 to 2019 and average.

A. Data Preprocessing

The significance of preprocessing data must be comprehended first before moving on to developing forecasting system. It has the potential to make or ruin forecasting. The self-lag differencing method have been used to preprocess the data where the previous 3 years had been used for training and 4th year for forecasting.

Table II shows the originally annual collected data for 10 countries. This data is in the form of raw data which cannot be directly fed to the forecasting model, so before moving ahead preprocessing steps have been applied. The previous day's, month's, and year's data are very important to make the prediction. In other words, the value at time t-1 has a significant impact on the value at time t. Lags are the past values, therefore t-1 is lag 1, t-2 is lag 2, and so on. The lag features-based data preparation techniques have been used and after the process the result that have been found is shown in Table III.

Table III depicts the data preparation of county United States after removing the null values from Table IV. For the year 2013, International tourist's arrival counted 179.31 million and lag 3 values is T-3 which is 162.28 million, lag 3 is 147.27 million, and lag 1 is 162.28. The data is now ready for the next step to develop the machine learning based forecasting systems.

What are the variations in the top 5 arrivals countries have been shown in Fig. 9, which shows that the United States is at the top and Spain is in second place, but the growth of tourist arrivals are growing year by year. France having variations every year means ups and down in arrivals year by year.

B. Machine Learning Models

Machine learning technique has inspired due to the wide variety of applications in multiple domains. Machine learning has proven to perform better on complicated data and tasks, and this is a reason for draining it for adopting into the forecasting systems. Below are the models with different parameters that have been applied in this research:

a) Support Vector Regression

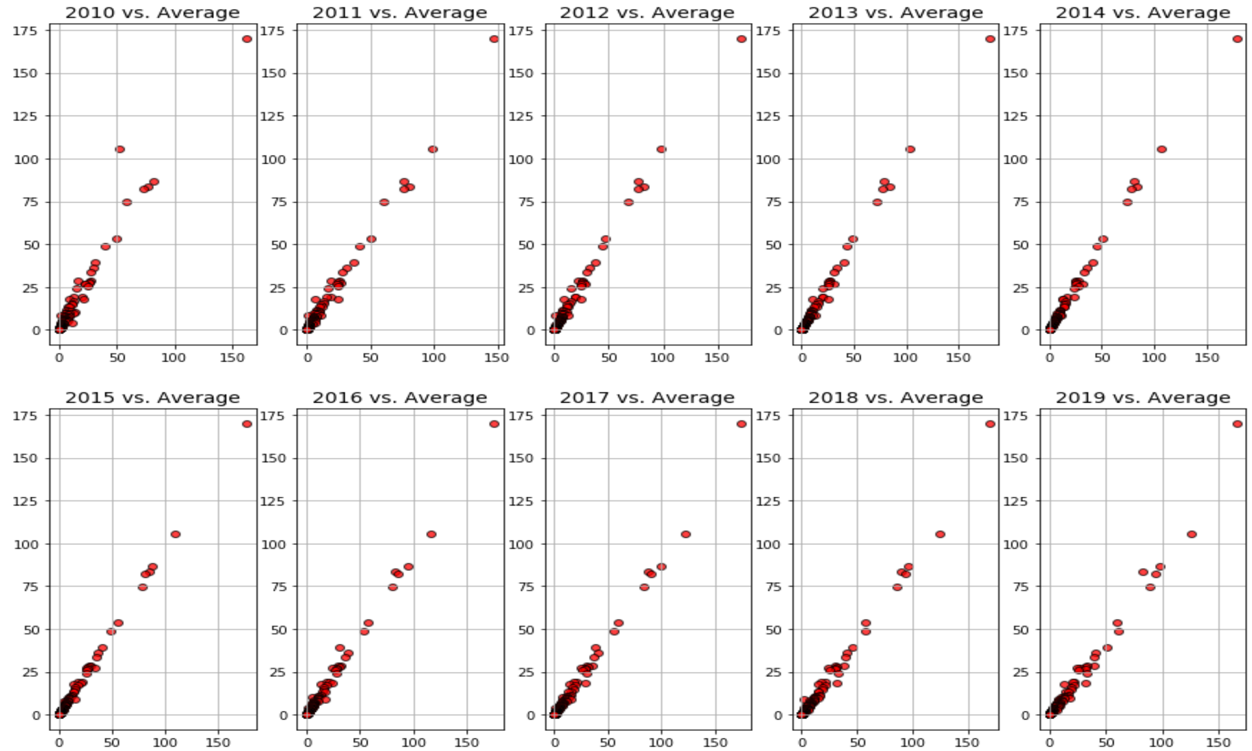
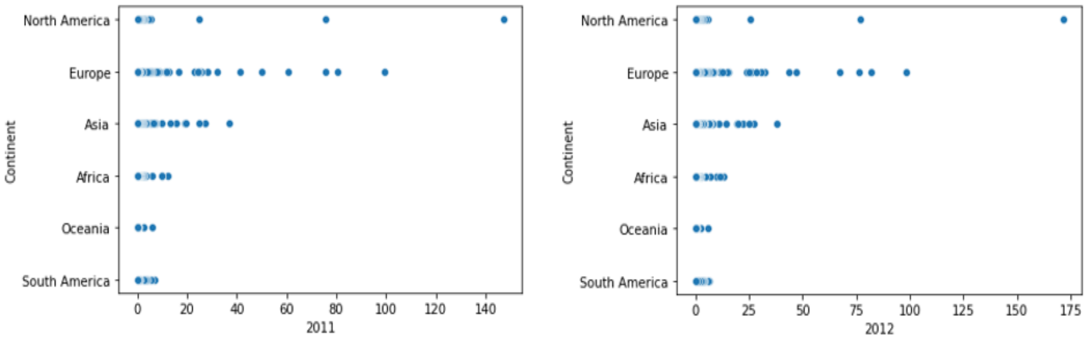


Fig. 6. Year Wise vs Average Data Analysis.



(a)

(b)

Fig. 7. Continent Wise Visitors.

The Support Vector Regression (SVR) is adopted from support Vector Machine (SVM) for the regression type data to predict the value. While dealing with real number data, the SVM changes its variant as regression. The output for real type data has infinite possibilities, and researchers have to see all possible solutions to decide the final prediction. While dealing with real time data, the primary idea is to minimize equation 1 and in case, if problem is linear then support vector regression is represented by equation 2 and error minimization has given in equation 3:

$$y = x\prime\beta + b \tag{1}$$

$$y = \sum_{n=1}^{N} (\alpha - \alpha\prime_i)(x_i, x) + b \tag{2}$$

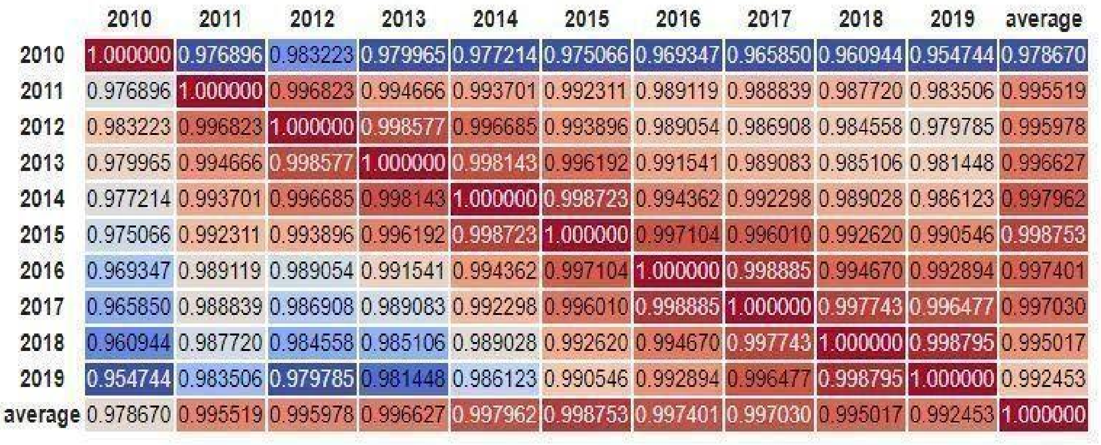
$$\frac{1}{2}||w||2 + c.i = 0n(-'i) \tag{3}$$

below constraints need to be taken care with linear support vector regression.

$$y_i \ - wx_i + b \leq (\epsilon + \epsilon_i)$$

$$y_i \ - wx_i + b \leq (\epsilon + \epsilon\prime_i)$$

$$\epsilon + \epsilon\prime_i \geq 0$$



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	average
2010	1.000000	0.976896	0.983223	0.979965	0.977214	0.975066	0.969347	0.965850	0.960944	0.954744	0.978670
2011	0.976896	1.000000	0.996823	0.994666	0.993701	0.992311	0.989119	0.988839	0.987720	0.983506	0.995519
2012	0.983223	0.996823	1.000000	0.998577	0.996685	0.993896	0.989054	0.986908	0.984558	0.979785	0.995978
2013	0.979965	0.994666	0.998577	1.000000	0.998143	0.996192	0.991541	0.989083	0.985106	0.981448	0.996627
2014	0.977214	0.993701	0.996685	0.998143	1.000000	0.998723	0.994362	0.992298	0.989028	0.986123	0.997962
2015	0.975066	0.992311	0.993896	0.996192	0.998723	1.000000	0.997104	0.996010	0.992620	0.990546	0.998753
2016	0.969347	0.989119	0.989054	0.991541	0.994362	0.997104	1.000000	0.998885	0.994670	0.992894	0.997401
2017	0.965850	0.988839	0.986908	0.989083	0.992298	0.996010	0.998885	1.000000	0.997743	0.996477	0.997030
2018	0.960944	0.987720	0.984558	0.985106	0.989028	0.992620	0.994670	0.997743	1.000000	0.998795	0.995017
2019	0.954744	0.983506	0.979785	0.981448	0.986123	0.990546	0.992894	0.996477	0.998795	1.000000	0.992453
average	0.978670	0.995519	0.995978	0.996627	0.997962	0.998753	0.997401	0.997030	0.995017	0.992453	1.000000

Fig. 8. Correlation Matrix of Tourist Arrivals.

TABLE II. SAMPLES OF COLLECTED DATA

Country	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Finland	3.67	4.19	4.23	2.8	2.73	2.62	2.79	3.18	3.22	3.29	0.0988
Paraguay	3.17	3.37	3.66	3.54	3.46	4.1	4.32	4.74	4.18	4.37	0.308
Netherlands	10.88	11.3	11.68	12.78	13.93	15.01	15.83	17.92	18.78	20.13	0.47
Qatar	1.7	2.06	2.32	2.61	2.84	2.94	2.94	2.26	1.82	2.14	0.551
Croatia	49.01	49.97	47.19	48.35	51.17	55.86	57.59	59.24	57.67	60.02	1.48
United States	162.28	147.27	171.63	179.31	178.31	176.86	175.26	174.29	169.32	166.01	2.68
Hungary	39.9	41.3	43.57	43.61	45.98	48.35	52.89	54.96	57.67	61.4	3.686
Ukraine	21.2	24.54	25.06	26.03	13.23	13.03	13.73	14.58	14.34	13.71	3.965

TABLE III. PREPROCESSING OF COLLECTED DATA

Country	Year	T-3	T-2	T-3	Arrival
United States	2010	NaN	NaN	NaN	162.28
United States	2011	NaN	NaN	162.28	147.27
United States	2012	NaN	162.28	147.27	171.63
United States	2013	162.28	147.27	171.63	179.31
United States	2014	147.27	171.63	179.31	178.31
United States	2015	171.63	179.31	178.31	176.86
United States	2016	179.31	178.31	176.86	175.26
United States	2017	178.31	176.86	175.26	174.29
United States	2018	176.86	175.26	174.29	169.32
United States	2019	175.26	174.29	169.32	166.01

TABLE IV. PREPROCESSED DATA AFTER REMOVAL OF NAN VALUES

Country	Year	T-3	T-2	T-3	Arrival
United States	2013	162.28	147.27	171.63	179.31
United States	2014	147.27	171.63	179.31	178.31
United States	2015	171.63	179.31	178.31	176.86
United States	2016	179.31	178.31	176.86	175.26
United States	2017	178.31	176.86	175.26	174.29
United States	2018	176.86	175.26	174.29	169.32
United States	2019	175.26	174.29	169.32	166.01

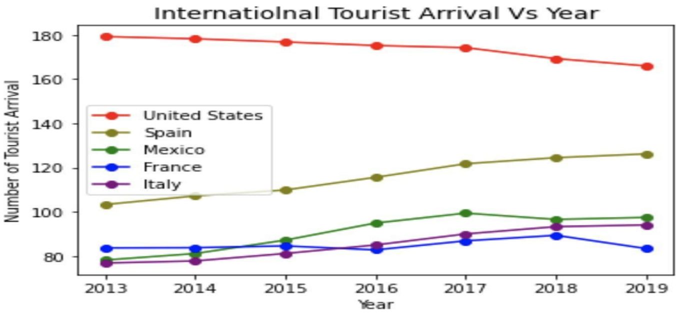


Fig. 9. Top 5 Country Tourist Variations Comparisons.

b) Random Forest Regressor A tree structure of data arrangement gives an actual estimator. Random forest follows the pattern of the decision tree, where each data node will be split into daughter nodes. While splitting the data nodes, a split criterion is being chosen to be appropriately partitioned. All the data nodes at the bottom are terminal. In the case of regression data, the predicted value at a node is the average response variable for all observers in the nodes. Splitting criteria for regression is chosen by equation (4).

$$RSS = \sum_{left}(y_i - y_l^*) + \sum_{right}(y_i - y_r^*) \tag{4}$$

Where

$$y_l^* = mean\ y\ value\ for\ left\ node$$

$$y_r^* = mean\ y\ vlaue\ for\ right\ node$$

Sometimes dealing with classified data where the predicted class is the most common class in the node, which is also known as the majority vote. So far classification tree estimated probability calculated members of each class.

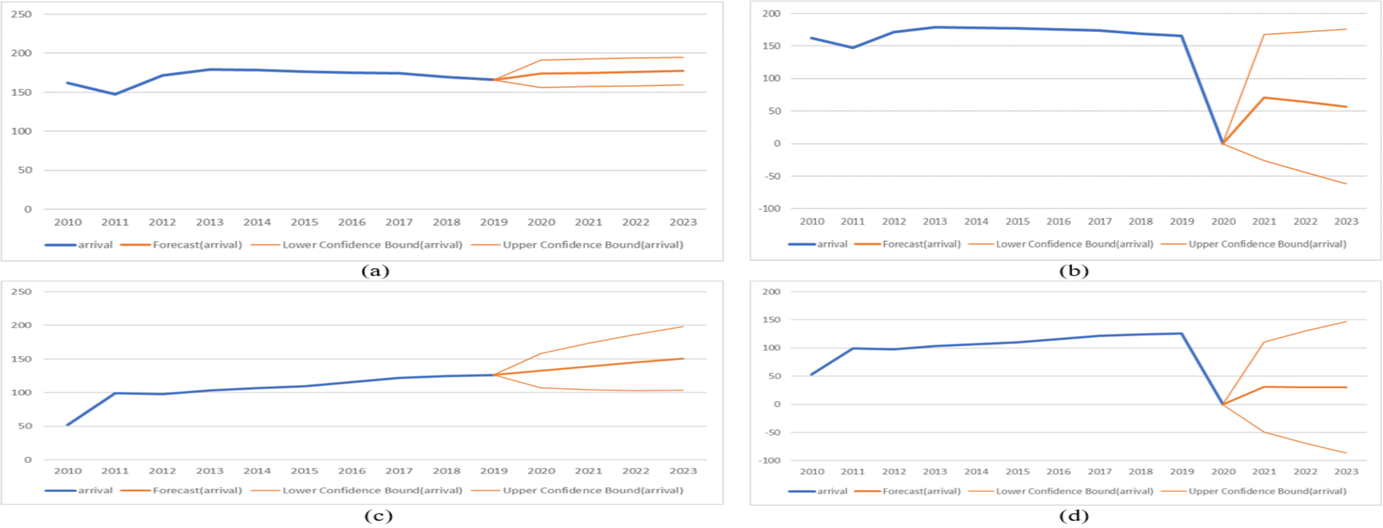


Fig. 10. Forecasting Trend before and after Covid for USA and Spain.

Splitting criteria for classified data is given by Gini index, which is shown in equation (5).

$$Gini = N_l \sum_{k=0}^{n} k = 1, ..KP_{kl}(1 - P_{kl}) + N_r \sum k = 1, ..kP_{kr}(1 - P_{kr}) \tag{5}$$

Where

$$P_{kl} = \text{proportion of class k in left node.}$$

$$P_{kr} = \text{proportion of class k in right node.}$$

A random forest is a meta estimator that uses averaging to increase predictive precision and control over-fitting by fitting several classifying decision trees on different sub-samples of the dataset. Although the sub-sample size is the same as the initial input sample size, the samples are drawn with substitution. For classification tasks, the Decision Tree and Random Forest models are often used. However, the concept of Random Forest as a regularizing meta-estimator over a single decision tree is better illustrated by extending it to regression problems. In this way, it can be shown that a single decision tree is vulnerable to overfitting and learning false associations in the face of random noise. At the same time, an adequately built Random Forest model is more resistant to such overfitting.

IV. Experimental Result

The experimental results are collected using the following setup. Dataset used contained tourist arrivals for mostly all global countries. Python 3.7 was used along with scikit-learn, NLTK and NumPy libraries for each learning algorithm used, the regression techniques, and the confusion matrix. First, baseline results have been obtained using SVR, and then RFR model have been used to train data. The number of features is 11, and data partitioning between training of 2/3 and testing is 1/3.

It is essential to evaluate the model using testing data once it has been trained. To verify the model's correctness, numerous evaluation matrices have been utilized. This study focuses mostly on R-Squared, a commonly used effectiveness accuracy metric as shwon in Table V. R-Square determines if the data is near the fitted regression line. The regression model, it's also known as the coefficient of determination or the value of multiple determination. R-squared is defined as the percentage of the variance in the response variable that is explained by a linear model.

TABLE V. Experimental Results

Model Name	R-Square
SVR (Kernel='linear')	0.994
SVR (Kernel='rbf')	0.863
RFR (Tree Model)	0.847

The R-squared value is always between 0 and 100%. In this research two models have been considered: SVR with different kernels and RFR with tree model. The partitioning of data between training and testing is 67% and 33% and found that the accuracy which is shown in Table V for SVR (kernel= linear) is 0.994, with kernel RBF is 0.863. The random forest regression works well for small size dataset and found R-Square result is 0.847.

V. Forecasting Before and After Covid-19

The graphs plot the forecasting regression percovid and post covid for the next 4 years and found that normal forecasting upper boundary line is going as usual however when Covid-19 enforced the restriction around the world then tourist arrival has drastically resulted null.

Fig. 10(a) shows the forecasting before covid-19 for the Country USA and Fig. 10(b) depicts the forecasting during the existence of Covid-19. As per the graph visualization upper bound forecasting is reviving in the year 2022. The same things can be seen in Fig. 10(c) without Covid-19 and Fig. 10(d) after Covid-19 for the Country Spain and differences can be observed as like USA.

VI. Discussion

The collected worldwide tourist arrival data from different trusted and official web portals are analysed to forecast future international tourist arrivals. Such analysis can mobilized the tourism industry. Table II has shown a different kind of collected data and time-frequency with applied methods by the researcher in [16], and the studies from [36]–[38] show that most of the collected data is focused on a specific region in a country. The focus is to align with the objective of the United Nations World Tourism Organization (UNWTO) to work on collective universal data and analyze the impact of tourism due to worldwide tourist movements.

The collected data has the frequency of yearly and building optimized machine learning models[39] of this variety of data having a lot of challenges. The actual data is from the year 2010 to 2019; in 2020 international travel was heavily impacted due to place confinement [37]. Whole word is gone through a very worst situations due to covid and many technological techniques have been used to analyze and predict the situations. This study emphasizes the comparative study before the Covid-19 pandemic of actual forecasting and how suddenly forecasting systems stopped predicting the correct values once the global health pandemic started. Handling the future pandemic situations and fulfilling the basic requirement for new arrivals, forecasting models will help not only to governing bodies but also to hospitality service provides such as hotels, restaurant, transportation, etc. The pandemic has also given a crises situation in healthcare industries and how basic medicine facilities can be provided to tourists who could not return to his/her country due to lockdown enforcement.

VII. Conclusion

Digitalization has made the whole world a village, it remains important to have collective forecasting of data that represents the whole globe. The UNWTO, and the World Travel and Tourism Council (WTTC) are working continuously on improving the global tourism facilities by analyzing the demand and increasing number of arrivals. This research focused on overall worldwide data with machine learning approaches such as support vector regression and random forest regression and the result shows that support vector regression has given better results as compared to random forest regression.

Since the number of vistors for any country is not exactly known, building the model with multiple techniques would give an analytic view for the comparative study. This is the reason for developing the model by using machine learning. Since the collected data is on annual frequency, it doesn't fit well with deep learning techniques so consideration for this work is machine learning techniques i.e., support vector regression and random forest regression. A future extension of this work would be a clustering-based forecasting system where the groups of data would be based on countries with most arrivals, mid arrival countries, and low arrival countries. The focus is to collect monthly data to forecast the season-wise and finding the most interesting month of a tourist visit.

Acknowledgment

The authors would like to thanks Intelligent Computing Lab, Department of Computer Science at BITS Pilani, Dubai Campus for infrastructure and computing facilities.

References

[1] A. Jelušić, "Modelling tourist consumption to achieve economic growth and external balance: Case of Croatia," Tourism and Hospitality Management, vol. 23, no. 1, pp. 87–104, 2017, doi: 10.20867/thm.23.1.5.

[2] H. Ramkissoon, "COVID-19 Place Confinement, Pro-Social, Pro-environmental Behaviors, and Residents' Wellbeing: A New Conceptual Framework," Frontiers in Psychology, vol. 11, no. September, pp. 1–11, 2020, doi: 10.3389/fpsyg.2020.02248.

[3] UN World Travel Organization, "International Tourism Highlights," Unwto, pp. 1–24, 2019.

[4] S. Aynalem, K. Birhanu, and S. Tesefay, "Employment Opportunities and Challenges in Tourism and Hospitality Sectors," Journal of Tourism & Hospitality, vol. 05, no. 06, 2016, doi: 10.4172/2167-0269.1000257.

[5] S. Pervaiz, Z. Ul-Qayyum, W. H. Bangyal, L. Gao, and J. Ahmad, "A Systematic Literature Review on Particle Swarm Optimization Techniques for Medical Diseases Detection," Computational and Mathematical Methods in Medicine, vol. 2021, 2021, doi: 10.1155/2021/5990999.

[6] G. González-Rivera, P. Loungani, and X. (Simon) Sheng, "Forecasting issues in developing economies," International Journal of Forecasting, vol. 35, no. 3, pp. 927–928, 2019, doi: 10.1016/j.ijforecast.2019.04.005.

[7] H. Rezapouraghdam, A. Akhshik, and H. Ramkissoon, "Application of machine learning to predict visitors' green behavior in marine protected areas: evidence from Cyprus," Journal of Sustainable Tourism, 2021, doi: 10.1080/09669582.2021.1887878.

[8] G. Athanasopoulos and R. J. Hyndman, "Modelling and forecasting Australian domestic tourism," Tourism Management, vol. 29, no. 1, pp. 19–31, 2008, doi: 10.1016/j.tourman.2007.04.009.

[9] T. Baldigara, "Modelling domestic tourism in Croatia," Turisticko poslovanje, vol. 43, no. 22, pp. 19–38, 2018, doi: 10.5937/turpos1822019b.

[10] J. Wu and Z. Ding, "Improved grey model by dragonfly algorithm for chinese tourism demand forecasting," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12144 LNAI, no. September, pp. 199–209, 2020, doi: 10.1007/978-3-030-55789-8 18.

[11] H. Drucker, C. J. C. Surges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," Advances in Neural Information Processing Systems, no. May 2018, pp. 155–161, 1997.

[12] M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression," Biostatistics, pp. 1 14, 2004, [Online]. Available: http://escholarship.org/uc/item/35x3v9t4.pdf

[13] Y. H. Liang, "Forecasting models for Taiwanese tourism demand after allowance for Mainland China tourists visiting Taiwan," Computers and Industrial Engineering, vol. 74, no. 1, pp. 111–119, 2014, doi: 10.1016/j.cie.2014.04.005.

[14] F. L. Chu, "Using a logistic growth regression model to forecast the demand for tourism in Las Vegas," Tourism Management Perspectives, vol. 12, pp. 62–67, 2014, doi: 10.1016/j.tmp.2014.08.003.

[15] A. G. Salman, B. Kanigoro, and Y. Heryadi, "Weather forecasting using deep learning techniques," ICACSIS 2015 - 2015 International Conference on Advanced Computer Science and Information Systems, Proceedings, pp. 281–285, 2016, doi: 10.1109/ICACSIS.2015.7415154.

[16] B. Petrevska, "Predicting tourism demand by A.R.I.M.A. models," Economic Research-Ekonomska Istrazivanja , vol. 30, no. 1, pp. 939–950, 2017, doi: 10.1080/1331677X.2017.1314822.

[17] Y. W. Chang and M. Y. Liao, "A seasonal ARIMA model of tourism forecasting: The case of Taiwan," Asia Pacific Journal of Tourism Research, vol. 15, no. 2, pp. 215–221, 2010, doi: 10.1080/10941661003630001.

[18] H. Zou and Y. Yang, "Combining time series models for forecasting," International Journal of Forecasting, vol. 20, no. 1, pp. 69–84, 2004, doi: 10.1016/S0169-2070(03)00004-9.

[19] J. Bowden, "A logistic regression analysis of the cross-cultural differences of the main destination choices of international tourists in China's main gateway cities," Tourism Geographies, vol. 8, no. 4, pp. 403–428, 2006, doi: 10.1080/14616680600922104.

[20] P. Nystrup, E. Lindström, J. K. Møller, and H. Madsen, "Dimensionality reduction in forecasting with temporal hierarchies," International Journal of Forecasting, vol. 37, no. 3, pp. 1127–1146, 2021, doi: 10.1016/j.ijforecast.2020.12.003.

[21] A. Aslanargun, M. Mammadov, B. Yazici, and S. Yolacan, "Comparison of ARIMA, neural networks and hybrid models in time series: Tourist arrival forecasting," Journal of Statistical Computation and Simulation, vol. 77, no. 1, pp. 29–53, 2007, doi: 10.1080/10629360600564874.

[22] K. Solat and K. P. Tsang, "Forecasting exchange rates with elliptically symmetric principal components," International Journal of Forecasting, vol. 37, no. 3, pp. 1085–1091, 2021, doi: 10.1016/j.ijforecast.2020.11.007.

[23] K. Y. Chen and C. H. Wang, "Support vector regression with genetic algorithms in forecasting tourism demand," Tourism Management, vol. 28, no. 1, pp. 215–226, 2007, doi: 10.1016/j.tourman.2005.12.018.

[24] C. Goh, R. Law, and H. M. K. Mok, "Analyzing and forecasting tourism demand: A rough sets approach," Journal of Travel Research, vol. 46, no. 3, pp. 327–338, 2008, doi: 10.1177/0047287506304047.

[25] N. Kamel and A. Atiya, "Tourism demand forecasting using machine learning methods," Aiml, no. January, 2008, [Online]. Available: http://infos2007.fci.cu.edu.eg/tourism/07184.pdf

[26] N. Kim and Z. Schwartz, "The accuracy of tourism forecasting and data characteristics: A meta-analytical approach," Journal of Hospitality Marketing and Management, vol. 22, no. 4, pp. 349–374, 2013, doi: 10.1080/19368623.2011.651196.

[27] H. Song and G. Li, "Tourism demand modelling and forecasting-A review of recent research," Tourism Management, vol. 29, no. 2, pp. 203–220, 2008, doi: 10.1016/j.tourman.2007.07.016.

[28] J. G. Brida and N. Garrido, "Tourism forecasting using SARIMA models in Chilean regions," International Journal of Leisure and Tourism Marketing, vol. 2, no. 2, p. 176, 2011, doi: 10.1504/ijltm.2011.038888.

[29] O. Claveria and S. Torra, "Forecasting tourism demand to Catalonia: Neural networks vs. time series models," Economic Modelling, vol. 36, pp. 220–228, 2014, doi: 10.1016/j.econmod.2013.09.024.

[30] W. Lijuan and C. Guohua, "Seasonal SVR with FOA algorithm for single-step and multi-step ahead forecasting in monthly inbound tourist flow," Knowledge-Based Systems, vol. 110, pp. 157–166, 2016, doi: 10.1016/j.knosys.2016.07.023.

[31] E. Noersasongko, F. T. Julfia, A. Syukur, , P., R. A. Pramunendar, and C. Supriyanto, "A Tourism Arrival Forecasting using Genetic Algorithm based Neural Network," Indian Journal of Science and Technology, vol. 9, no. 4, pp. 3–7, 2016, doi: 10.17485/ijst/2016/v9i4/78722.

[32] S. Sun, Y. Li, S. Wang, and J. e. Guo, "Tourism demand forecasting with tourist attention: An ensemble deep learning approach," arXiv, 2020.

[33] A. Guizzardi, F. M. E. Pons, G. Angelini, and E. Ranieri, "Big data from dynamic pricing: A smart approach to tourism demand forecasting," International Journal of Forecasting, vol. 37, no. 3, pp. 1049–1060, 2021, doi: 10.1016/j.ijforecast.2020.11.006.

[34] A. A. Semenoglou, E. Spiliotis, S. Makridakis, and V. Assimakopoulos, "Investigating the accuracy of cross-learning time series forecasting methods," International Journal of Forecasting, vol. 37, no. 3, pp. 1072–1084, 2021, doi: 10.1016/j.ijforecast.2020.11.009.

[35] S. Sun, Y. Wei, K. L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," Tourism Management, vol. 70, no. February 2018, pp. 1–10, 2019, doi: 10.1016/j.tourman.2018.07.010.

[36] J. G. Brida and W. A. Risso, "Research note: Tourism demand forecasting with sarima models - The case of south tyrol," Tourism Economics, vol. 17, no. 1, pp. 209–221, 2011, doi: 10.5367/te.2011.0030.

[37] Y. Yao et al., "A paired neural network model for tourist arrival forecasting," Expert Systems with Applications, vol. 114, pp. 588–614, 2018, doi: 10.1016/j.eswa.2018.08.025.

[38] C. F. Chen, M. C. Lai, and C. C. Yeh, "Forecasting tourism demand based on empirical mode decomposition and neural network," Knowledge-Based Systems, vol. 26, pp. 281–287, 2012, doi: 10.1016/j.knosys.2011.09.002.

[39] W. H. Bangyal, K. Nisar, A. A. B. A. Ibrahim, M. R. Haque, J. J. P. C. Rodrigues, and D. B. Rawat, "Comparative Analysis of Low Discrepancy Sequence-Based Initialization Approaches Using Population-Based Algorithms for Solving the Global Optimization Problems," Applied Sciences 2021, Vol. 11, Page 7591, vol. 11, no. 16, p. 7591, Aug. 2021, doi: 10.3390/APP11167591.