Using Machine Learning to forecast Solar Power in Ismailia

Gehan H. Allam[*], Basem E. Elnaghi[**], M. N. Abdelwahab[**], Reham H. Mohammed[**]

[*] Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt
[**] Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt

DOI: 10.29322/IJSRP.11.12.2021.p12033
http://dx.doi.org/10.29322/IJSRP.11.12.2021.p12033

Abstract- The main goal of this study is the forecasting of Photovoltaic (PV) power production in Ismailia, Egypt. For the aim of this, a Photovoltaic system was chosen to produce electric power as a clean power. The forecasting process depends on the weather data downloaded from the database of the European Commission's science and knowledge service for the region of Ismailia city, Egypt. The data had been collected for the period of 7 years from 2010 to 2016 and the forecasted period had been chosen to be 12 months for the future prediction. Three Machine Learning algorithms have been developed and tested to forecast the PV power production: Facebook Prophet, Random Forest, and Long Short-Term Memory Networks. It was observed from the results that Facebook Prophet acts more accurate than the others. The preparation procedure of the dataset and the development of the ML models had been built by python programming language to reduce the running time.

Index Terms- Machine Learning, Python, PV, Solar Energy, Weather Parameters.

I. INTRODUCTION

Nowadays, Due to the rapid development of Renewable energy such as PV, it has been considered as an alternative to conventional energy. But this transition would never be possible unless the process of forecasting is utilized properly. The Technical Summary Version 2 (TS-V2) of the Intergovernmental Panel on Climate Change (IPCC) is a special report on climate change highlights the various impacts of climate change on the earth natural systems. It mainly focused on the interactions between the atmosphere and the land surface. Human activities have a significant impact on the climate and global warming. According to a study conducted by the IPCC, human activities have increased the global warming by 1.530 degrees Celsius since the pre-industrial period [1]. To reduce global warming and improve the efficiency of solar and PV systems, they should be integrated with Information Technology (IT). This will enable them to provide stable and predictable energy systems [2]. The stability of PV systems can be achieved by using accurate forecasts and the continuous monitoring of weather data collected by sensors [3]. Intelligent systems have been created to improve the performance of the solar energy prediction process by utilizing weather data. According on historical data, several of these systems have been classified [4,5]. Das, U.K *et. al.* [4] classified solar prediction methods into different categories and experimented with various approaches and procedures until they discovered that Machine Learning methods for predicting solar energy are the most accurate. Even though the prediction method requires a vast data set. ML is a brilliant method and a cutting-edge methodology. Some ML models are required a specific structure for the input data and projected output data. That is, if the machine learning algorithms are fed unstructured input data, the model will not fit the output forecasted data correctly [5]. Although, there will be missing or distorted data during the gathering of data from sensors or metrological stations due to sensor failure, several researchers have proposed an approach to interpret the data flow [6,7]. Forecasting techniques and methodologies might be mathematical, statistical, or numerical in nature. While the first two approaches are widely employed (mathematical or statistical models such as Autoregressive Integrated Moving Average (ARIMA) or Linear Regression (LR), the third way relies on data self-information, and the Artificial Neural Network (ANN) is the most effective method [7]. The ARIMA model have the popularity stems mostly from its ability to extract statistical traits and use of the Box-Jenkins method [8]. To forecast PV power production, the ARIMA model can be used alone or in combination with other models. The ARIMA model drawback is that the time series data it uses must be stable [9]. Both weather and solar power forecasting have been developed using numerical models. However, this requires a powerful processing machine [10,11]. There are two basic types of solar energy forecasting models: cloud imaging with a physical model and machine learning models. Each model has a varied level of accuracy depending on the input data [4]. Researchers discovered that forecasting models can be divided into two types: direct forecasting models and indirect forecasting models. The solar energy was predicted directly by previous meteorological data in the direct models. The solar irradiation has been predicted in multiple time scales using indirect methods such as numerical, statistical, image processing, and ANN models [12,13]. Random Forest is a perfect technique for selecting the most effective factors on output to eliminate training time, according to R. Genuer *et al.* [14]. Some researchers used Photovoltaic-Electrolytic (PV-E) to produce hydrogen using several machine learning algorithms and found that Facebook Prophet was the most accurate technique, with an MAE of 3.7 percent for a forecasting period of three months, and the first element forecasted

was Global Horizontal Irradiance (GHI), followed by hydrogen [15]. Because of the ability to determine the relationship between the input and the output without prior knowledge of the physical characteristics, machine learning techniques are excellent for time series data in forecasting [6]. Five forecasting strategies were compared by Pedro H *et al.* [16]: The Persistent model, K-Nearest Neighbor (KNNs), ARIMA, ANNs, and ANNs optimized using Genetic Algorithms (GAs/ANN) were all found to be more successful than the time series models. The historical data is divided into two groups: training data (almost 70% of total data) is preferable for learning the model to forecast output power and testing data, the remaining 30% is prepared to validate the PV power predicting model [6]. Other researchers divided the data into three sections: 60% training, 20% test, and 20% validation [17]. The dataset was segmented into 88 percent for train and 12 percent for test by Sayed Altan Haider *et al.* [15]. A specific structure for the input data and projected output data is required by some machine learning approaches. That is, if the machine learning algorithms are fed unstructured input data, the model will not fit the output forecasted data correctly. Cleaning data can be obtained by several steps in different methodologies to improve the accuracy of the forecasting output [18]. Observations may not always be in the desired data format. Numeric and categorical data types are two examples of data types that are desirable for use in machine learning models. Numerical figures are required for ML-Algorithms of the regression type. These can take the form of a float or an integer value. In contrast, a categorical ML-Model, which can be separated into nominal, ordinal, or Boolean values, can be used. The nominal and ordinal values denote some sort of name for the ML-end-state.

II. METHODOLOGY

The machine learning algorithms utilized in prediction process had been presented in this section. The data of the weather parameters used and, from which database downloaded. The goal of the study is to forecast PV power generation. The technique of forecasting is referred to as supervised learning. As a result, time series regression is proposed in the ML models have been developed for the forecasting process.

A. Characteristics of PV System

By entering the PV technology, module efficiency, slope, and PV maximum power into the database, the power generated by the PV system computed using the Eq (1) [15,19]:

$$E_{PV} = G \times \eta_{PC} \times \eta_{PV} \quad (1)$$

Where:

G : the global solar irradiance
η_{PC} : Power Conditioning (PC) efficiency
η_{PV} : the reference efficiency of the PV module.

The global solar irradiance calculated from the summation of direct and diffuse irradiance by using Eq. (2) [20] as follows:

$$G_i = D_i + B_i \times \cos z \quad (2)$$

Where:

D_i: direct solar irradiance.
B_i: diffuse solar irradiance.
z: the zenith angle at the measuring time.

For the aim of this paper, the PV technology had been chosen was Crystalline silicone with system losses 14%, two axis mounting type and, 17% efficiency [15,21]. Table (1) contains the specifications of the PV system assumed to be used.

Table 1: The specifications of PV system.

Characteristic		Description
Electrical Parameter	Power Max(W)	252
	Short Circuit Current (A)	8.95
	Maximum Voltage (V)	29.95
	Open circuit Voltage (V)	37.25
	System Voltage (V)	1000
Mechanical Specification	Type	Monocrystalline Panel
	No. of Cells in Series	60
	Frame Type	Aluminum
	Y – Axis Mounting Hole	946
	X – Axis Mounting Hole	609
	Junction Box Cable 4 mm	4 mm
	Glass Type	Tempered 4 mm

Figure (1) shows the general steps in the prediction process starting from downloading weather parameters, data preparation, features selection, testing all models and selecting the best one to validate accuracy.

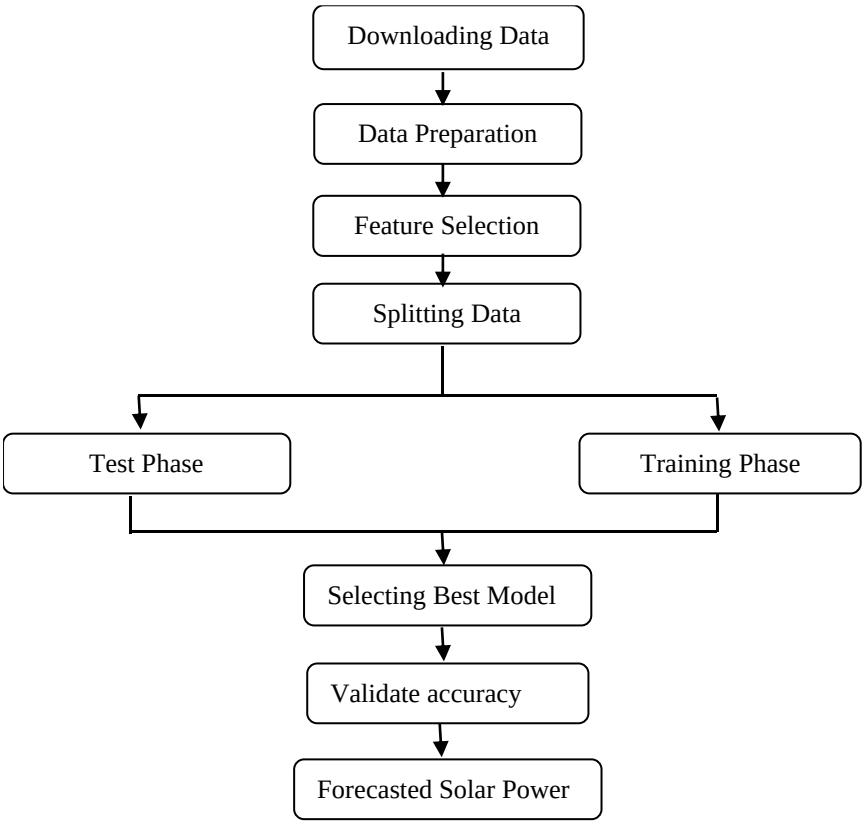

Figure 1: General steps to Forecast Solar Power.

B. Data

Solar irradiance components (direct irradiance, diffuse irradiance, reflected irradiance), wind speed at 10m, temperature recorded at 2m, and sun height all these features are included in the dataset. Outliers and missing values were studied in the dataset. For discovering outliers, clean and prepare the dataset using Python programming language tools and libraries such as *Pandas* and *Numpy*. Spearman and Pearson Correlation techniques had been developed to choose the most characteristics has a positive impact

on the prediction output. After extracting all the features required for the forecasting process using multiple ways, all the parameters required for the forecasting process were chosen.

C. Machine Learning Models

The aim of this paper is getting the best results of the prediction of output solar power, three forecasting models are presented in the next sections.

- *Random Forest (RF)*
  RF is one of the most often used machine learning algorithms. It is well-suited to machine learning challenges, but it may also be used to select features. This algorithm is thought to be a useful tool for determining which features have the biggest impact on the output variable. In fact, the more inputs ML algorithms have, the longer running time. As a result, decreasing the number of features before running the ML model is critical. The preparation of the dataset is the first step in the RF algorithm's feature selection. The *Sklearn* function *train-test-split()* was used to produce subsets from the dataset by splitting the data into two portions for the training section. The selected features are represented in the first component, which includes direct irradiance, reflected irradiance, diffuse irradiance, temperature, sun height, wind speed, time of day, and day of the year. The output parameter, which is the output power, is represented by the other half. Each portion is divided into two sections, one for testing and the other for training. The next step in the prediction process is running the model for training and fit to the forecasted output.

- *Facebook Prophet*
  Prophet is an open-source framework developed by Facebook researchers to handle and forecast time series data that is affected by daily seasonality, weekly, annual, and holiday fluctuations. Prophet is a decomposable time series model, which means that it is made up of multiple smaller models. The method is sensitive to missing data and can effectively deal with outliers. The formula of Prophet is made up of the following elements as described by Eq. (3):

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (3)$$

  Where:

  $g(t)$: the trend function non-periodic changes.
  $s(t)$: are the periodic changes, such as weekly or yearly changes within the data.
  $h(t)$: describes the effects of the holidays throughout the year.
  ϵ_t: represents the error which cannot be found from other components of this model and assumed to be normal distributed. For this study the yearly seasonality and daily seasonality set to true.

  Prophet requires two columns in the input dataset, one labeled (ds) for date or timestamp data and the other named (y) for numeric values (the forecasted variables). All other weather variables were disregarded and discarded. Using the Panda library, the dataset was converted to a Data Frame. In this model, the holidays have no effect on the data.

- *Long Short-Term Memory Network (LSTM)*
  Are a form of artificial neural network. specially of Recurrent Neural Network in particular (RNN). The LSTM features recurrent connections, unlike a standard MLP (Multilayer Perceptron), which has multiple layers with neurons and the input data is propagated through the network itself. This means that the output is additionally contextualized by the state of earlier activations. However, unlike traditional RNNs, the LSTM Network's design allows it to avoid the problem of vanishing or exploding gradients. This indicates that the weight update mechanism rapidly alters the weights in one way or the other, graduating to zero or infinity. For longer sequences, these phenomena render the neural network unusable [22,23].

The results of ML models compared due to several performance metrics, such as:

- Mean Squared Error (MSE) is defined as the summation of the squared errors between the actual variable and the forecasted one. Eq. (4) described the calculations of MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (4)$$

  Where: Y is the actual variable and $\hat{Y}$ is the predicted variable.

- Mean Absolute Error (MAE) is the average of the errors between the set of actual observations and the prediction values of the same group. Eq. (5) described the MAE:

$$MAE = \frac{\sum_{i=1}^{N} |Y_i - \hat{Y}_i|}{N} \quad (5)$$

- R-squared represents the goodness of the prediction model and measures the strength of the relation between the model and the input data. The measurements of the R-squared is in the scale of 0 - 100% and can be calculated by Eq. (6):

$$R^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \quad (6)$$

Where the actual variable is Y, $\hat{Y}$ is the predicted variable and, $\bar{Y}$ is the mean of the actual variable.

III. RESULTS AND DISSCUTION

R2, MAE, and MSE were used as performance indicators to compare all the models. The selected algorithms were developed using the impact of meteorological data acquired from the PVGIS database to forecast the output power of a PV system. The collected data is timeseries data in one hour apart. The forecasting procedure began after the preparation of the dataset for each algorithm individually. Table (2) compares all models due to the error metrics. The three models act well for the chosen dataset, but Facebook Prophet performed more accurate than the others. Figure (2) represents a comparison between all models due to their mean squared error for different horizon days. from Figure (6), it can be seen that in some horizons, LSTM and RF have lower MSE than Prophet, but Prophet learns the data better.

Table2: The Performance Metrics between Models

Model	R^2	MSE	MAE
Random Forest	0.82	34.26	12.832
Prophet	0.93	10.76	8.765
LSTM	0.91	23.87	11.7

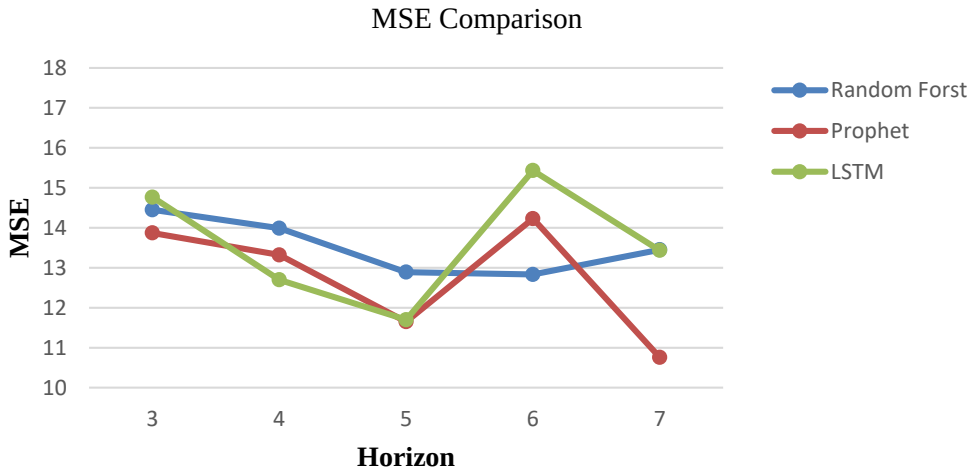

Figure 2: MSE comparison between models over different forecast horizon.

Figure (3) demonstrates the production of forecasted power of PV system in one hour interval in any selected day of the forecasted period. It can be seen that the forecasted power different from day to other, because the production of the power depends on the weather parameters especially the solar irradiance.

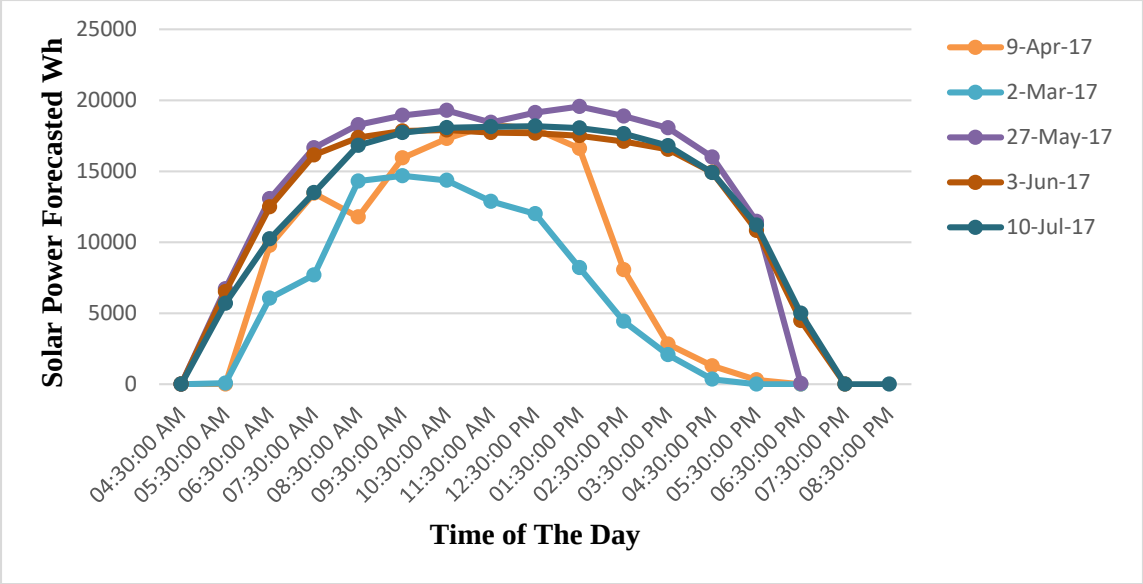

Figure 3: Solar Power Production in one hour interval.

IV. CONCLUSION

The major goal of this research is to use Machine Learning Algorithms to predict the amount of hydrogen produced by a PV system in Ismailia, Egypt. As a result, weather parameters were acquired from the PVGIS-ERA5 for the interval time from 2010 to 2016 at Latitude 30.601 and Longitude of 32.278. The weather data contains several features that had to be investigated in terms of their significance to the forecasting process output. For the prediction of the PV system energy output, each attribute was compared to its specific informative benefit. Only six features from the PVGIS dataset were chosen in general. The solar irradiation components (Direct – Diffuse – Reflected), temperature at 2 meters, wind speed at 10 meters, sun height, and PV maximum power assumption from a solar system producing 252 kwh were all included in the dataset. All the meteorological features were produced for the models by analyzing the Pearson and Spearman correlation coefficients for each of the two variables. Three machine learning algorithms were developed for the PV outcome power forecasting process: Random Forest, Facebook Prophet, and LSTM. Except for the Facebook Prophet, which works well with PV data solely as input data, each algorithm was trained and tested with a variety of datasets and PV power information. All the models were developed with the goal of forecasting solar energy power. To improve performance for the optimum state, Random Forest and LSTM experimented with Hyper Parameter Tuning. Facebook Prophet was the most accurate model with MSE of 10.76 and R^2 of 93%.

REFERENCES

[1] Shukla, P. R., et al. "Technical summary, 2019." Climate change and land: An IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems, 2019.

[2] Ameur, Arechkik, et al. "Forecast modeling and performance assessment of solar PV systems." Journal of Cleaner Production 267 (2020): 122167.

[3] Carrera, Berny, and Kwanho Kim. "Comparison analysis of machine learning techniques for photovoltaic prediction using weather sensor data." Sensors 20.11 (2020): 3129.

[4] Das, Utpal Kumar, et al. "Forecasting of photovoltaic power generation and model optimization: A review." Renewable and Sustainable Energy Reviews 81 (2018): 912-928.

[5] Elsheikh, Ammar H., et al. "Modeling of solar energy systems using artificial neural network: A comprehensive review." Solar Energy 180 (2019): 622-639.

[6] Theocharides, Spyros, et al. "Machine learning algorithms for photovoltaic system power output prediction." 2018 IEEE International Energy Conference (ENERGYCON). IEEE, 2018.

[7] Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. "Statistical and Machine Learning forecasting methods: Concerns and ways forward." PloS one 13.3 (2018): e0194889.

[8] Boland, John. "Time series modelling of solar radiation." Modeling Solar Radiation at the Earth's Surface. Springer, Berlin, Heidelberg, 2008. 283-312.

[9] DIAGNE, Maimouna, et al. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renewable and Sustainable Energy Reviews, 2013, 27: 65-76.

[10] Perez, Richard, et al. "Forecasting solar radiation–Preliminary evaluation of an approach based upon the national forecast database." Solar Energy 81.6 (2007): 809-812.

[11] Dunlop, Ewan D., Lucien Wald, and Marcel Suri. Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale. Nova Science Publishers Inc., 2006.

[12] Tanaka, Kenichi, et al. "Optimal operation by controllable loads based on smart grid topology considering insolation forecasted error." IEEE transactions on smart grid 2.3 (2011): 438-444.

[13] Hocaoğlu, Fatih O., Ömer N. Gerek, and Mehmet Kurban. "Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks." Solar energy 82.8 (2008): 714-726.

[14] Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests." Pattern recognition letters 31.14 (2010): 2225-2236.

[15] Syed Altan Haider, Muhammad Sajid, Saeed Iqbal, Forecasting hydrogen production potential in islamabad from solar energy using water electrolysis, International Journal of Hydrogen Energy, Volume 46, Issue 2, 2021, Pages 1671-1681, ISSN 0360-3199.

[16] Pedro, Hugo TC, and Carlos FM Coimbra. "Assessment of forecasting techniques for solar power production with no exogenous inputs." Solar Energy 86.7 (2012): 2017-2028.

[17] Cousineau, Denis, and Sylvain Chartier. "Outliers detection and treatment: a review." International Journal of Psychological Research 3.1 (2010): 58-67.

[18] Brownlee, Jason. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery, 2020.

[19] Rahmouni S, et al. Prospects of hydrogen production potential from renewable resources in Algeria. Int J Hydrogen Energy 2017;42(2):1383e95.

[20] Carneiro, Tiago, et al. "Performance analysis of google colaboratory as a tool for accelerating deep learning applications." IEEE Access 6 (2018): 61677-61685.

[21] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, “Using the jupyter notebook as a tool for open science: An empirical study,” in ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2017, pp. 1–2.

[22] Touili, Samir, et al. "A technical and economical assessment of hydrogen production potential from solar energy in Morocco." international journal of hydrogen energy 43.51 (2018): 22777-22796.

[23] Brownlee, Jason. Long short-term memory networks with python: develop sequence prediction models with deep learning. Machine Learning Mastery, 2017.

AUTHORS

First Author – Gehan H. Allam, MSc. Student at Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt, gehan.allam@eng.suez.edu.eg.

Second Author – Basem E. Elnaghi, Associated Professor at Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt, basem.elhady@gmail.com

Third Author – M. N. Abdelwahab, Associated Professor at Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt, Mohamed.Nabil@eng.suez.edu.eg

Fourth Author – Reham. H. Mohamed, Assistant Professor at Department of Electrical Power and Machines, Suez Canal University, Ismailia, Egypt, riry4mody@yahoo.com.

Correspondence Author – Gehan H. Allam, geahn.allam@eng.suez.edu.eg, gehan.allam1@gmail.com, +201221721995.