FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 550-556 | e-ISSN: 2347-470X

Regression Based Predictive Machine Learning Model for Pervasive Data Analysis in Power Systems

Dr. K. Sasikala[1], Dr. J. Jayakumar[2], Dr. A. Senthil Kumar[3], Dr. Shanty Chacko[4], Dr. Hephzibah Jose Queen[5]

[1]Vels Institute of Science and Technology, Chennai, India
[2,4,5]Karunya Institute of Technology and Sciences, Coimbatore, India
[3]Dilla University, Ethiopia

***Correspondence:** Dr. J. Jayakumar; jayakumar@karunya.edu

ABSTRACT- The main aim of this paper is to highlight the benefits of Machine Learning in the power system applications. The regression-based machine learning model is used in this paper for predicting the power system analysis and Economic analysis results. In this paper, Predictive ML models for two modified IEEE 14-bus and IEEE-30 bus systems, integrated with renewable energy sources and reactive power compensative devices are proposed and developed with features that include an hour of the day, solar irradiation, wind velocity, dynamic grid price, and system load. An hour-wise input database for the model development is generated from monthly average data and hour-wise daily curves with normally distributed standard deviations. A very significant Validation technique (K Fold cross validation technique) is explained. Correlation between Input and output variable using spearman's correlation analysis using Heat maps. Followed by the Multiple Linear Regression based Training and testing of the Modified IEEE 14 and IEEE30 Bus systems for base load case, 10% and 20% load increment with the 5-fold cross validation is also presented. Comparative analysis is performed to find the best fit ML Model for our research.

Keywords: Regression-based Machine Learning, Correlation Analysis, Power system Analysis, Voltage Stability, Cost analysis.

ARTICLE INFORMATION
Author(s): Dr. K. Sasikala, Dr. J. Jayakumar, Dr. A. Senthil Kumar, Dr. Shanty Chacko, Dr. Hephzibah Jose Queen
Received: 11/07/2022; **Accepted**: 03/09/2022; **Published**: 07/09/2022;
e-ISSN: 2347-470X;
Paper Id: IJEER 1106-07;
Citation: 10.37391/IJEER.100324
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100324.html

crossref member
CROSSREF.ORG
THE CITATION LINKING BACKBONE

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The process of transferring expertise into a skill or knowledge is defined as Learning. Humans are flawed when attempting to create a correlation between several variables and assessments. Machine learning techniques can solve some of these problems with higher accuracy and robustness [1]. This paper clearly explains Machine Learning Techniques, Data selection, and Feature Selection using Correlation and Machine Learning Model Development. In the Existing Techniques, the Programming concept is trained on the Input data and the program can get the output for the particular concept. While in Machine Learning, the Algorithm is trained with the Input data and its output data. The machine Learns the concept behind the Input and output through all the hidden layers and develops its logic from the Input and output data sets provided to the Machine. Now the Machine Learned logic is applied to the test system to check the effectiveness of the Machine Learning algorithm developed for our trained data [2-3]. When the Machine Learning techniques rely on the labeled information sets then it can be known as Supervised Learning. When the computational output referred to as dependent parameters are reliant up on the independent parameters, then the Prediction can be achieved using the Supervised Machine Learning Techniques. The algorithm is built on the training information data sets and after many iterations the algorithm becomes efficient. Regression and Classification are the two main types of Supervised Learning techniques [4-5].

By utilizing the training data set, Regression -Supervised Learning algorithm is used to forecast single value output using the training data set. The output value is always called the dependent variable, while the inputs are the independent variable. We have different types of regression in Supervised Learning, they are Linear Regression, Multiple Regression, and Polynomial Regression. In Linear regression, to predict the outcome only one input variable is applied. In the case of the Multiple Regression model, the prediction of output is done based on many inputs. In the case of Polynomial regression, the relationship between the input and the output variables is portrayed through graphical representation [6]. For example, the relationship between solar radiation and the time of the day. Both Linear and Polynomial Regression based Machine Learning Techniques are used. It is concluded that polynomial regression of order 4 was best suited for our problem. Regression based ML was opted over other artificial intelligence techniques like Neural Networks etc., because it provides a better control and leads to a better analysis of the correlation between the features and better feature selection.

The rest of the paper is organized in the following manner. *Section 2* describes the implementation of proposed predictive Machine Learning Model on IEEE 14 bus and IEEE 30 bus voltage stability analysis. This section also describes training, testing and K Fold validating techniques, correlation between input and output and heat maps. *Section 3* exhibits the results attained by Multiple Linear Regression based Training and testing of the Modified IEEE 14 and IEEE30 Bus systems for various load conditions and in *Section 4* we draw the conclusion.

2. SYSTEM MODEL

In the proposed method, modified test systems (IEEE 14 bus and IEEE 30 bus) are taken into consideration for the forecasting of Technical Analysis (Voltage stability) and Economic analysis (Price of power received from the grid to meet the load demand of the Modified test system. Figure1 depicts the basics of the Machine Learning techniques which comprises three stages.

- **Stage1:** Data Set Generation, is the most consequential portion of developing the ML Model.
- **Stage2:** Training and Validating ML Model, the developed data-based is used for training the ML model, and also validation is done
- **Stage3:** The effectiveness of the developed ML model is tested in this stage for the particular application using the test data available.

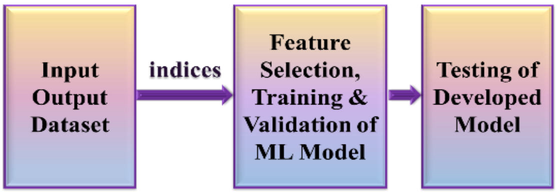
Input Output Dataset
indices
Feature Selection, Training & Validation of ML Model
Testing of Developed Model

Figure 1: Basic Flow of ML Model Development

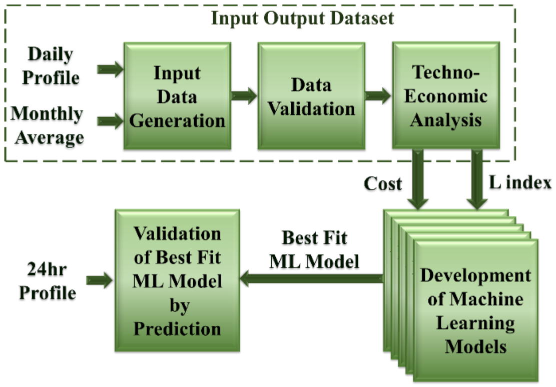

Figure 2: Block Diagram Predictive Technical and Economic Analysis

Figure 2 portrays the Block diagram of the predictive technical (Voltage stability) and Economic (Price of the power received from the grid to meet the load demand) analysis.

Stage 1: Data Set Generation, is the most consequential portion of developing the ML Model. This stage 1 is sub-divided into three small divisions.

- **Data generation:** The Input data are solar irradiation, Wind velocity, Hour of the day, Dynamic Load, and Electricity price.
- **Data validation:** The data are validated and they should be within the feasible limits
- **Technical and Economic analysis:** This analysis is done in the MATLAB platform and the output variables are obtained. The output data are the L-Index value for each load bus and the Price of power received from the grid.

Stage 2: While selecting the data for modeling the Machine Learning Model, correlation analysis must be done. Correlation analysis is used to select the feature for the ML model development. In some cases, the two independents are highly correlated so in such cases, we can neglect one of the independent variables. Here we are choosing spearman's correlation over Pearson's correlation as our variables are non-linear. Now the Input data to Model our ML are independent variables (Solar Irradiation, wind velocity, Load, Electricity, Hour of the day) and the target /Dependent variables (L-Index, Price of power received from the grid to meet the load demand). When providing these data to the ML the model develops an equation that gives the relationship between the dependent and independent variables.

Stage3: Now the effectiveness of the Best fit model is validated using 24hr test data by allowing it to predict the Technical and Economic values.

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable (target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables

2.1 Predictive Techno Economic Analysis

The Input Output data set is generated by gathering the data from the NASA website, Analysis of Voltage Stability and Operating price in MATLAB coding environment, etc., The Input data set generated is checked for its feasibility using the load flow computations and the power rating of the RES devices. The ML model based on regression is selected for our analysis as the Regression model gives accurate statistical results. The multiple linear regression and polynomial regression (3rd, 4th, and 5th order) are compared with the help of the K-Fold cross-validation technique. The best fit model is shortlisted and the predictive analysis is carried out for the 24 hrs. test data [7-10].

All-inclusive numerical information set for forecasting is developed in stages of three.

1. Random Input-Output Data generation
2. Data Feasibility Test
3. Final Input-Output database.

The algorithm that is proposed is based on the Zaragoza region in Spain case study. The wind velocity and solar irradiation data of that particular location are taken into consideration for input data implementation [11-12].

- **Load data:** The load demand of the test system for one day is taken from calibrated data provided by (Li et al., 2017). Load data for each hour of one day (24hr) is devised and the same load data is produced for 365 days and 8760 hours by a normally distributed random deviation formula.
- **Solar irradiation data:** The average solar irradiation data provided by Jayakumar et.al [13-14]. For each month for one year. This data is acquired from the National Aeronautics and Space Administration. Then we convert the monthly average data to an 8760hrdataset.
- **Wind velocity data:** The monthly average wind velocity data is represented and they are acquired from the National Aeronautics and Space Administration. Then we convert the monthly average data to hourlydatafor8760hr.
- **Electricity Price data:** The electricity prices are in accordance with the prices published by the European Energy Exchange on the 1st of September 2017.

The Load demand, dynamic price, wind velocity, solar irradiation, and hour of the data are the Inputs taken into consideration for the prediction of voltage stability analysis and Operating cost of IEEE 14 and 30 bus data using ML techniques.

The obtained input database, which includes daily profile information of load demand, solar irradiation, power prices, and wind speed for the period 8760hr, is checked for feasibility to ensure the information is well within the variable's feasible range.

The data viability assessment of the electric grid load is evaluated by comparing the energy output with the rated capacity of the implemented farm and evaluating the convergence of power flow and renewable inputs. To estimate the system's price and voltage stability, a machine learning-based algorithm is constructed. The 2 output factors with in scrutiny are calculated for every hour for input information. As ML models are formed by training and testing a basic model with a large input-output dataset. One of the output parameters is the L-Index for Voltage stability Index and the other output parameter price of power consumed from the electric grid.

2.2 Development of Predictive ML Model

Two independent algorithms are created to attain maximum levels of accuracy in the forecasting of the two different output responses. **Spearman's Correlation Coefficient**: Spearman's correlation which can be also called "Spearman's Rank-Order Correlation" is applied when there is a lack of normally distributed nature within the variable set.

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2 - 1)} \tag{1}$$

A heat map provides a way to present data in a statistical graph format as an attractive and informative medium to impart information. Sea born is a Python library and it is based on mat plot lib. It is used for data visualization. A heat map is one of the items supported by seaborne where variation in related data is depicted with the help of a color palette.

Figures 3-6 illustrate the heat maps that were created. Absolute correlation is represented by a value of +/-1, whereas no association is represented by a value of 0. From the heat maps plotted from the correlation analysis of Input features, it is seen that no 2 variables have a good correlation between them. As a result, all five recommended variables are taken into account while training and testing Machine Learning algorithms for predictive modeling. Considering a day's data, the hour of day constantly increases from 1 to 24, while both solar irradiation and wind speed increase and decrease during the period.

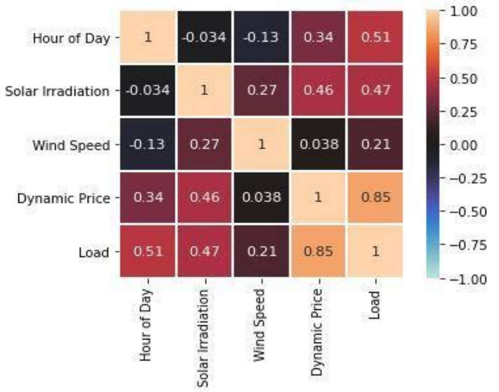

Figure 3: Heatmap of correlation among the input variables for 14 Bus

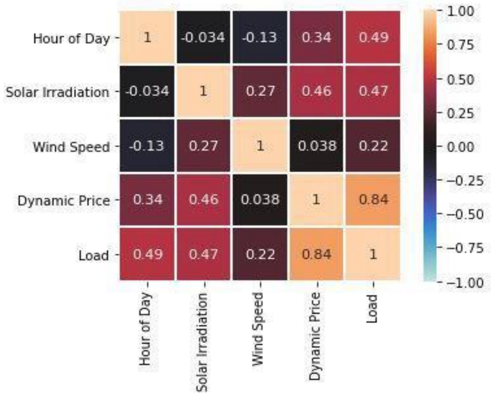

Figure 4: Heat map of correlation among the input variables for 30 Bus

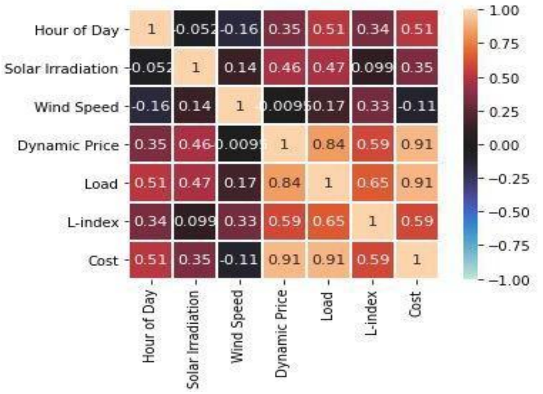

Figure 5: Heat map of Spearman correlation –Voltage stability and Cost for 14 Bus

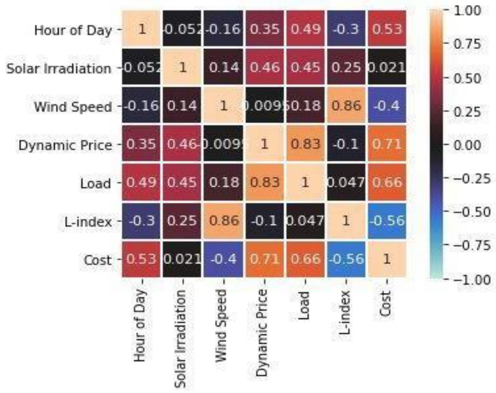

Figure 6: Heat map of Spear man correlation–Voltage stability and Cost for 30 Bus

2.3 Training and Testing of ML Model Development

Using the acquired dataset, an ML model is constructed through attaining phase and testing. ML techniques for system design may be divided into two types: supervised learning and unsupervised learning algorithms. R Squared (R^2) denotes the proportion of the variance for the dependent variable y that's described by the independent variable X. R^2 explains the how much extent the variance of one variable can explain the variance of the other variable. Hence, when the R^2 of a model is 0.75, then the model's features can explain approximately 75% of the observed variation. R^2 is computed by taking one minus the sum of squares of residuals divided by the total sum of squares.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \qquad (2)$$

R^2 compares the compatibility of the proposed solution to a straight axis that serves as a reference. Once the proposed model works worse than a horizontal axis, R^2 is negative. Although the "square" is included, the R^2 formula allows it to get a negative number without breaking any math laws. R^2 is negative whenever the algorithm doesn't reflect the data trend and fits worse than a base line.

The first most important phase of ML algorithm construction is training, which involves feeding both input and result data to the ML algorithm to train and develop the system. The complete dataset is divided into two halves for training and testing, with the model being trained with 4 parts of the data and tested with the remaining 1 part of the data. The training data is used to create line are regression models and polynomial regression model so f orders 3,4 and 5. Transformation (normalization and polynomial transformation) and prediction are used to create polynomial regression models (using linear regression). The stream mechanism can be used to connect every one of these phases, reducing the computational burden.

Testing is the phase of ML model construction where the created model is validated or evaluated. As per the model structure, the remainder 20% of the testing data sources are being used to predict and estimate the relevant out comes. The root-mean-square error (RMSE) and R^2 scores are two metrics that can be used to evaluate the created models. The R^2 values represent the percentage of correct predictions that the developed model is capable of. A number of 1 indicates that the fit is flawless. As a result, R^2 values are used to assess the model's fitness using the testing data. Because it ensures that every observation from the original dataset has the chance of appearing in training and test set. This is one among the best approach if we have a limited input data.

2.4 Validation of the ML Model

Without checking the accuracy of model output, putting it to use and then relying on the results can be devastating. One may analyze and validate forecasts utilizing multiple strategies via ML model evaluation services. A cross-Validation is a crucial approach that experts employ frequently. The problem concerning machine learning techniques is that people won't understand how great they are until they are evaluated on a large dataset. Validation helps in estimating the performance of our model. K-Fold Cross Validation is one of the few types of cross-validation.

1. K-Fold Cross Validation is a common form of validation set used in machine learning. The procedures for performing K-fold cross-validation are as follows:
2. The entire training data set is divided in to k equal subsets, each of which is referred to as a fold. Let's call the folds f_1, f_2,..., and f_k.
3. From1to k, i=1to i=k

- The remaining k-1 folds are preserved in the Cross-validation training data set, whereas fold fi is kept in the Testing set.
- Our machine learning model is trained using a cross-validation training set, and the model's accuracy is determined by verifying the anticipated out comes against the validation set.

In the k-fold validation data approach, all of the items in the original training data set are used for training and validation. Furthermore, each item is validated just once. The entire dataset was split into training and testing data in the ratio 4:1 (i.e., 80% Training and 20% Testing) for linear and polynomial regression without cross-validation. A similar ratio was intended to be maintained to ensure better comparative analysis and validation of ML models. Hence, the dataset was divided into 5 folds – 4 of which will be for training and 1 will be for testing.

The total data is divided into 5-folds in this validation approach, and various sets of 4:1 train-test data are picked from the folds. In cross-validation, the combo that provides the greatest match is chosen. The final model is selected based on best fit in both single hold-out train-test split and cross-validation. In the next part, we'll go through the results and how we choose the best-fit model among the produced models.

3. RESULTS AND DISCUSSION

The 8760hr database created for the research is used to construct all of the Machine learning for the prediction analysis. Various training methods such as polynomial regression, 5-fold cross-validation, and multi-variable linear regression are used and their respective R^2 scores are calculated. For the forecast of voltage stability, we consider L-index as one of the output variables to be forecasted and analyzed. The model developed for this analysis considered the L-index of the entire system which is the maximum of the L-indices computed for every load bus. ML Model Validation for Voltage Stability Index and for Cost of Energy Purchased is mentioned in *table 1* and *table 2*.

Table 1: ML Model Validation for Voltage Stability Index

System	Increase in Load	R^2 values				
		Linear Regression	Polynomial Regression deg (3)	Polynomial Regression deg (4)	Polynomial Regression deg (5)	5-fold Cross Validation (Mean)
IEEE 14 bus	0 %	0.9638	0.9948	0.9976	0.9989	0.978 (p3) 0.946 (p4) -5.199 (p5)
	10 %	0.9679	0.9956	0.99	0.9989	0.973 (p3) 0.962 (p4) -2.438 (p5)
	20 %	0.971	0.9962	0.9982	0.9989	0.972 (p3) 0.964 (p4) -1.873 (p5)
IEEE 30 bus	0 %	0.9154	0.979	0.9932	0.9985	0.924 (p3) 0.845 (p4) -10.46 (p5)
	10 %	0.922	0.98	0.9861	0.998	0.942 (p3) 0.862 (p4) -7.745 (p5)
	20 %	0.9305	0.9815	0.9942	0.9986	0.935 (p3) 0.893 (p4) -6.376 (p5)

The machine learning model for the modified 6 bus system did not converge, i.e., the 1-year data taken for 6 bus system was insufficient to develop a reliable ML model. It will be considered in the future works. In *table 3*, R^2 values of the models developed for forecasting the voltage stability are given. The best models elected by comparing R^2 scores can be used to forecast the system's voltage stability for any futuristic input data. For forecasting cost of energy bought from the grid for the best-fit model is picked by making a comparison of the R^2 values of the developed models (multi-variable linear regression, polynomial regression, and 5-foldcross-validation).

In *table 3*, the R^2 values of the models created for forecasting the cost are displayed. The operating cost of any futuristic input data can be predicted using the selected model. Both cost and voltage stability prediction require a best-fit ML model. As the models of different loading conditions show similar R^2 values, the base load case is considered topic k the best fit model. It is also observed from *table 3* and *table 5* that the 4th-degree polynomial regression model exhibits a better fit among the other polynomial models. The comparison of R^2 scores obtained in the linear regression, 4th-degree polynomial regression, and 5-fold cross-validation of the 4th-degree polynomial regression for base load case is portrayed in *table 5*. It can be understood from the scores that the polynomial regression models of degree 4 are considered most suitable, for the forecast of both voltage stability and cost in a power system. The machine learning model for the modified 6 bus system did not converge, i.e., the 1-year data taken for 6 bus system was insufficient to develop a reliable ML model. This model can be applied for any larger power network with the availability of dataset (for considered network).

Table 2: ML Model Validation for Cost of Energy Purchased

System	Increase in Load	R^2 values				
		Linear Regression	Polynomial Regression deg(3)	Polynomial Regression deg(4)	Polynomial Regression deg(5)	5-foldCrossValidation (Mean)

IEEE 14bus	0 %	0.9648	0.9958	0.9988	0.9999	0.988(p3)
						0.966(p4)
						-5.181(p5)

	10 %	0.9689	0.9966	0.991	0.9999	0.989(p3)
						0.979(p4)
						-2.42(p5)

	20 %	0.972	0.9972	0.9992	0.9999	0.99 (p3) 0.982 (p4) -1.853 (p5)
IEEE 30 bus	0 %	0.9164	0.98	0.9949	0.9995	0.945 (p3) 0.865 (p4) -11.56 (p5)
	10 %	0.923	0.981	0.9871	0.999	0.947 (p3) 0.886 (p4) -8.745 (p5)
	20 %	0.9315	0.9825	0.9952	0.9996	0.958 (p3) 0.912 (p4) -5.368 (p5)

Table 3: Comparison of Developed ML Models (Base load)

Analysis	**ML Model**	**R^2values**	
		IEEE 14-bus	**IEEE 30- bus**
Voltage Stability	Linear Reg.	0.9638	0.9154
	4th deg. Polynomial Reg.	0.9976	0.9932
	5-fold CV	0.946	0.845
Cost	Linear Reg.	0.9648	0.9164
	4th deg. Polynomial Reg.	0.9988	0.9949
	5-fold CV	0.966	0.865

4. CONCLUSION

In this paper, a detailed explanation of the Machine Learning Technique is given. The step-by-step methodology of the Prediction by Machine learning model is described. The analysis and selection of the developed ML model are also discussed in this paper. A very significant Validation technique (K Fold cross-validation technique) is explained. Correlation between Input and output variable using spearman's correlation analysis using Heat maps. Followed by the Multiple Linear Regression based Training and testing of the Modified IEEE 14 and IEEE30 Bus systems for base load case, 10% and 20% loadincrementwiththe5-foldcross-validation is also presented. Comparative analysis is performed to find the best fit ML Model for our research. This model can be applied for any larger power network with the availability of dataset (for considered network). Both MATLAB and Machine Learning codes are generalized, i.e., they can interchangeably be used for any system by replacing the input data in the developed MATLAB codes for dataset generation replacing the modified dataset to train the ML models.

REFERENCES

[1] Kessel, P., &Glavitsch, H. (1986). Estimating the Voltage Stability of a Power System. IEEE Transactions on Power Delivery, 1(3).

[2] Kumar, S., Kumar, A., & Sharma, N. K. (2020). A novel method to investigate voltage stability of IEEE-14 bus wind integrated system using PSAT. Frontiers in Energy, 14(2).

[3] Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., & Yan, Z. (2009). A review on the forecasting of wind speed and generated power. Renewable and Sustainable Energy Reviews, 13(4).

[4] Leonardi, B., Ajjarapu, V., Djukanovic, M., & Zhang, P. (2010, August).Application of multi-linear regression models and machine learning techniques for online voltage stability margin estimation. 2010

IREP Symposium Bulk Power System Dynamics and Control - VIII (IREP).

[5] Musirin, I., & Abdul Rahman, T. K. (n.d.). Novel fast voltage stability index(FVSI) for voltage stability analysis in power transmission system. Student Conference on Research and Development.

[6] Parwaiz, A., Kumar Jain, V., Ansari, B., M., Jaipur, J., Professor, A.,(2008). Comparative Analysis of Load Flow Methods on Standard Bus System. International Research Journal of Engineering and Technology, 775

[7] Rahi, O. P., Yadav, A. K., Malik, H., Azeem, A., & Kr, B. (2012). Power System Voltage Stability Assessment through Artificial Neural Network. Procedia Engineering.

[8] Ren, C., Xu, Y., Zhang, Y., & Zhang, R. (2020). A Hybrid Randomized Learning System for Temporal-Adaptive Voltage Stability Assessment of Power Systems. IEEE Transactions on Industrial Informatics, 16(6)

[9] Ruisheng Diao, Kai Sun, Vittal, V., O'Keefe, R. J., Richardson, M. R., Bhatt,N., Stradford, D., &Sarawgi, S. K. (2009a). Decision Tree-Based Online Voltage Security Assessment Using PMU Measurements. IEEE Transactions on Power Systems, 24(2)

[10] Ruisheng Diao, Kai Sun, Vittal, V., O'Keefe, R. J., Richardson, M. R., Bhatt, N., Stradford, D., & Sarawgi, S. K. (2009b). Decision Tree-Based Online Voltage Security Assessment Using PMU Measurements. IEEE Transactions on Power Systems, 24(2)

[11] Jayakumar, Chitra, Shanty Chcacko Identification of Power Leakage and Protection of Over Voltage in Residential Buildings, International Journal of Electrical and Electronics Research, Volume 10, Issue 1, Pages 51 - 5630 March 2022

[12] J. Jayakumar and Honey Baby, Operating Cost Analysis of Microgrid Including Renewable Energy Sources and a Battery Under Dynamic Pricing, Lecture Notes in Electrical Engineering, 2022, 795, pp. 291–302

[13] J. Jayakumar and Hepsibah jose queen Comparative techno-economic analysis of power system with and without renewable energy sources and statcom Journal of Green Engineering, 2021, 11(2), pp. 1648–1667

[14] J. Jayakumar and Hepsibah jose queen Machine Learning-Based Predictive Techno-Economic Analysis of Power System IEEE Access, 2021, 9, pp. 123504–123516