Solar Irradiance Prediction Based on Weather Patterns Using Bagging-Based Ensemble Learners with Principal Component Analysis

Engr. Justin D. de Guia
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
justin_deguia@dlsu.edu.ph

Engr. Ronnie S. Concepcion II
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
ronnie_concepcionii@dlsu.edu.ph

Engr. Hilario A. Calinao Jr.
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
hilario_calinao@dlsu.edu.ph

Engr. Rogelio Ruzcko Tobias
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
rogelio_tobias@dlsu.edu.ph

Dr. Elmer P. Dadios
Manufacturing Engineering and Management Department
De La Salle University
Manila, Philippines
elmer.dadios@dlsu.edu.ph

Dr. Argel A. Bandala
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
argel.bandala@dlsu.edu.ph

***Abstract*— Energy production of photovoltaic (PV) system depends on the amount of solar irradiance present on a certain location. Accurate prediction of solar irradiance ensures economic integration of PV system to grid and leads to optimal dispatching of available energy resources. Weather conditions has strong correlation with solar irradiance, and its erratic nature causes fluctuation to energy production. Therefore, it is difficult to achieve consistent optimal energy production and reliable prediction of solar irradiance. In the study, a bagging-based ensemble learning system was used to predict solar irradiance based on weather patterns. Previous researches confirmed that ensemble learners produced unbiased and more accurate results compared to single learners. A pre-processed stacked long-short term memory model (stacked LSTM) was used as base learner in ensemble learning since it has good performance in handling time series sequences. A plot that compares the performance between single learner and ensemble learners was provided. From the plot, it shows that at some iteration, ensemble learners get consistent at providing more accurate predictions compared to single learners. Metrics used in the study include explained variance score, maximum residual error, mean absolute error, mean squared error, and regression score function.**

Keywords— bagging, energy management, ensemble learning, solar irradiance prediction, stacked long-short term memory

I. INTRODUCTION

Energy production of photovoltaic (PV) system depends on the amount of solar irradiance. Solar irradiance heavily depends on weather patterns; hence it is erratic [1][2]. The erratic nature of weather conditions results to fluctuating characteristics of solar power therefore, it is difficult to obtain consistent optimal energy production. Also, it is difficult to achieve reliable prediction of solar energy. Factors such as cloud presence, aerosols, and dust can affect the consistency of accurate prediction of solar irradiance.

Accurate solar irradiance forecasting ensures economic integration of PV system in the smart grid. Prediction of energy production and end-user demand can achieve optimal dispatching of available energy resources due to ahead knowledge of energy production and end-user demands [3][4]. It also allows users to be flexible in making decisions with regards to load scheduling. Lastly, users can also enjoy the benefits of profiting from energy production such as profitable bilateral contracts where it allows selling of excess energy production to electric distribution company.

Research involving solar irradiance or PV system forecast include implementing machine learning techniques by using methods such as fitting regression models, decision trees, and random forest. However, there are regression models that encounters underfitting and overfitting problems which may affect its prediction performance [5]. There is an attention to developing an algorithm that allows combining of predictions made from models or ensemble learners is used to improve accuracy during classification or regression problems. Linear combination of methods for time series data proved to produce unbiased and more accurate results compared to single learners that make up the ensemble. Types of ensemble learning for classification and regression problems include boosting and bagging. In the study, bagging is applied in predicting solar irradiance.

In this paper, a long-short term memory (LSTM) model was used as a base learner for bagging-based ensemble learning system. The ensemble learner consists of 20 LSTM models, where each model is preprocessed by z-score normalization and principal component analysis. The performance of ensemble learner is quantified by using metrics such as regression score function, maximum residual error, mean squared error, mean absolute error, and explained variance score. Comparison of base learner and ensemble learners is made by observing the trend of plot with the said metrics on the y-axis.

II. BAGGING BASED ENSEMBLE LEARNING

Error from learning model consist of three parts, namely; bias, variance, and irreducible error. It is expressed mathematically as;

$$Err(x) = (E[f(x)] - f(x))^2 + E[f(x) - E[f(x)]^2 + {\delta_e}^2 \quad (1)$$

Bias error is used in quantifying the difference of the average of predicted values to the actual values. High bias error indicates under performance of the model because when it misses out significant trends in input sequence. Variance refers to the consistency of the prediction made. High variance overfits training dataset and will perform poorly on datasets excluded on model training.

A powerful way to improve the performance of the model is to introduce ensemble learning. Ensemble method combines several model in order to produce one optimal predictive model.

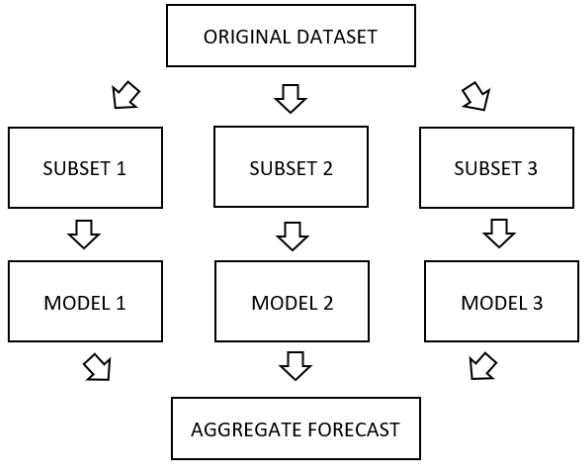

Fig. 1. Bagging Illustration

In the study, bagging-based ensemble learning was adopted in predicting solar irradiance. Figure 1 shows the general method for bagging. "Bagging" or bootstrap aggregation is an algorithm that combines all predictions made by base learners. The goal is to reduce the variance error of the estimates by averaging the estimates from multiple sources [6][7]. For instance, subsets of the data can be used to train neural networks and compute the ensemble expressed mathematically as

$$f(x) = \frac{1}{M}\sum_{1}^{M} f_m(x) \quad (2)$$

Bootstrap sampling is used to obtain different subsets from input sequence with replacement, through resampling methods. A base learner is trained for each random subset, and works in parallel. The learners are independent to one another. The final prediction of ensemble learning is determined from the combination of all predictions from the base models. Predictions are made by taking the average predictions made by each base learner [8].

III. Dataset Exploration

Fig. 2. Weather Station

Figure 2 shows the weather station deployed in Morong, Rizal where the on-grid PV system is located. Data gathering of weather parameters was conducted between September 2019 up to February 2020. Recorded measurements are saved as comma separated values (.csv) format. Table 1 shows the statistics of the parameters chosen as features during modeling.

The output variable in the study is the solar irradiance. Solar irradiance is the intensity coming from the sun in the form of electromagnetic radiation. It is measured in terms of watt per square meter (W/m^2). Pyranometer is an instrument for measuring solar irradiance on planar surface. A photodiode calibrated based on its spectral property was used as a pyranometer [9][10]. The photodiode is enclosed with Teflon diffuser to enable proper reading operations. Weather parameters considered as features include windspeed, ambient light, humidity, ambient and station temperature, station altitude, and absolute and sea level pressure.

TABLE I. Statistics of Feature Parameters

Feature	*Average*	*Standard Deviation*	*Minimum Value*	*Maximum Value*
Windspeed	15.9235	2.6095	0.5065	26.6640
Illuminance	1131.1892	710.1353	0.0245	4541.8745
Humidity	74.1667	19.1923	31.8	100
Ambient Temperature	31.8207	12.1710	16.8200	63.94
Station Temperature	32.6449	4.9210	23.98	45.8
Sea Level Pressure	1020.3179	2.3139	1012.4520	1027.7080
Station Altitude	69.77	19.1071	8.8	135
Absolute Pressure	1004.8927	2.2789	997.1420	1012.1680

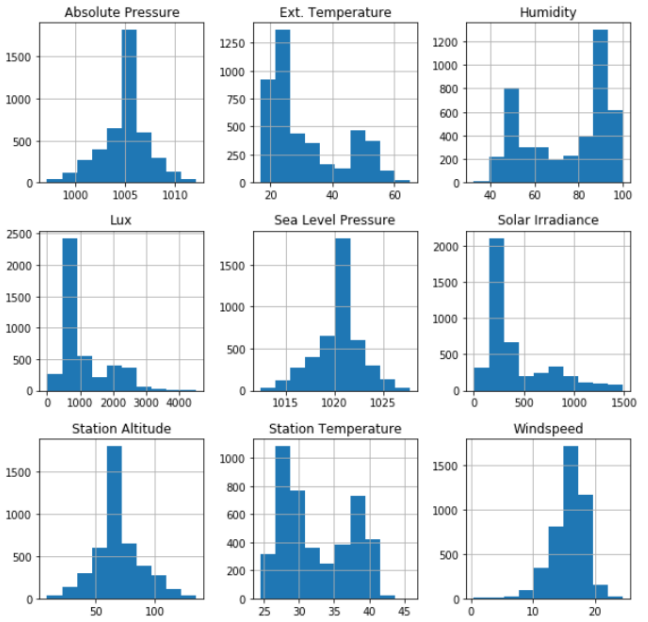

Fig. 3. Histogram of Variables found on the Dataset

Figure 3 shows the histogram of the variables found in the dataset. Histogram is useful for checking the normality of the dataset by analyzing the shape of dataset distribution. The plot shows that some variables don't have normal distribution like appearance such as humidity and temperatures. Features such as windspeed, solar irradiance, and illuminance shows skewness either to the left or right. This implies that using linear regression model for model fitting is not valid because of the assumption that all variables must possess normality. Neural network is more appropriate to use because it doesn't have any assumption on data or error. It uses exhaustive search for assuming any function, even if it has complex patterns.

Also, neural network has activation function to handle non-linearity found in input dataset.

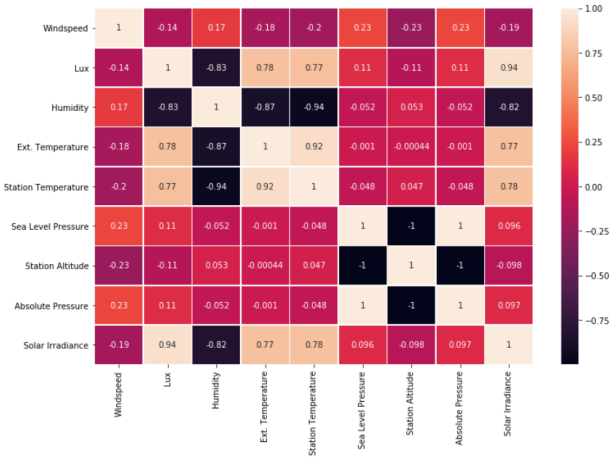

Fig. 4. Correlation Heatmap of the Variables

Figure 4 shows the correlation heatmap of the dataset. The darkest color indicates strong inverse relationship, otherwise it possesses strong direct relationship between two variables. It is reported that solar irradiance correlates strongly with ambient light and ambient temperature, having a regression coefficient of 0.94 and 0.77, respectively. It implies that as the temperature goes warmer, the sunlight intensifies as well as the ambient light. Also, solar irradiance shows strong negative correlation with humidity, having a regression coefficient of -0.82. Humidity is associated with rainfall and high value of it is due to evaporation. Rainfall results to reduced sunlight intensity and ambient temperature. Lastly, solar irradiance has weak negative correlation with windspeed, having a regression coefficient of -0.19. Windspeed has weak direct relationship with humidity, sea level pressure and absolute pressure having regression coefficient values of 0.1836, 0.2161, and 0.2159, respectively. It means that there's a chance of having reduced sunlight intensity if there is a strong windspeed due to high atmospheric pressure and humidity.

IV. Flowchart

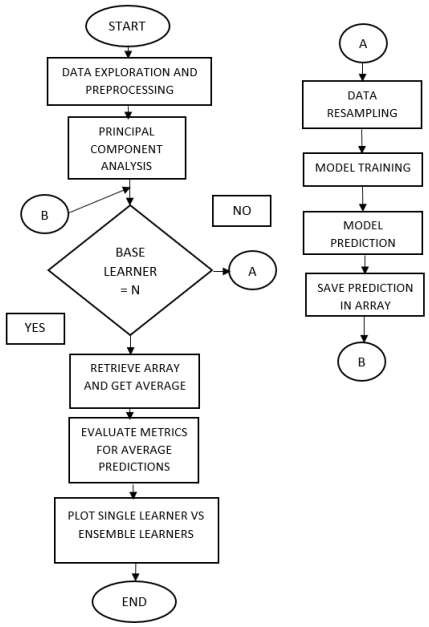

Fig. 5. Flowchart of the Bagging-based Ensemble Learning

Figure 5 shows the flowchart of bagging-based ensemble learning adopted in the study. The input dataset will first undergo preprocessing and data analysis to explore the relationship between variables. Preprocessing involves Bayesian ridge regression for handling missing values in rows and z-score normalization for handling outliers. To determine relevant features in input dataset, principal component analysis is applied. A for-loop method was used in implementing the bagging method by setting a control number equal to the number of base learners. In the study, there are 20 base learners in an ensemble. Inside the iteration, the input dataset is resampled randomly and is later used for model training. The prediction made by the fitted model is saved on an array and the iteration will stop until it reached 20th iteration. The predictions made by ensemble learners in each iteration is averaged and is evaluated by metrics applicable in regression problems. A plot of performance comparison between ensemble learners and single learner is provided for further analysis.

V. Data Preprocessing

A. *Bayesian Ridge Regression*

Bayesian Ridge Regression was used to handle the missing values on rows by estimating its value based on the correlation found on rows. The values are estimated using probability distribution, specifically from Gaussian distribution. The model for Bayesian Linear Regression with the output generated from probability distribution is mathematically expressed as

$$y \sim N(\beta^T X, \sigma^2 I) \tag{3}$$

The output is generated in terms of mean and variance. The mean is expressed as the product of the transpose of weight matrix and predictor matrix. Variance, however, is expressed in terms of the product of square of standard deviation and identity matrix. Bayesian ridge regression works on a principle based on finding the posterior distribution of model parameters assuming it is from Gaussian distribution. The posterior probability of the model is a conditional probability, in terms of training and output sequences mathematically expressed as

$$P(\beta|y,X) = \frac{P(y|\beta,X)*P(\beta|X)}{P(y|X)} \tag{4}$$

The posterior probability refers to the ratio of the product of likelihood of data and prior probability of parameters to normalization. The advantage of Bayesian ridge regression include adaptability to the given dataset and use of regularization parameters during estimation. Making model inferences, however, requires high computational time.

B. *Z-score Normalization*

Normalization helps in improving model training because of faster convergence of data. In the study, z-score was used over min-max normalization because of handling outliers. However, this does not guarantee that the data have the exact same scale. The formula for z-score normalization is

$$z = \frac{x-\mu}{\sigma} \tag{5}$$

Where x is any point in dataset, μ is the mean value of the feature, and σ is its standard deviation.

C. *Principal Component Analysis*

In the study, principal component analysis was applied in order to reduce the size of input dataset without compromising its information. This also reduce the likelihood of overfitting the data, as well as minimizing the computational time requirement during model training [11][12]. Figure 6 shows the first 3 principal components projected in 3d plane after applying the principal component analysis.

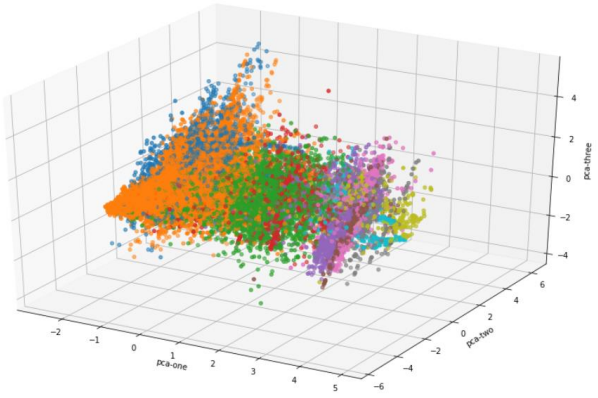

Fig. 6. 3D Plot of First 3 Principal Components

Figure 7 shows the plot of cumulative explained variance versus number of principal components. The plot determines the amount of variability retained when arbitrarily choosing the number of principal components. A cumulative explained variance value close to 1 indicates higher variability of data. In the study, principal component analysis determined that having 4 principal components makes the projected dataset close to 100 percent variability.

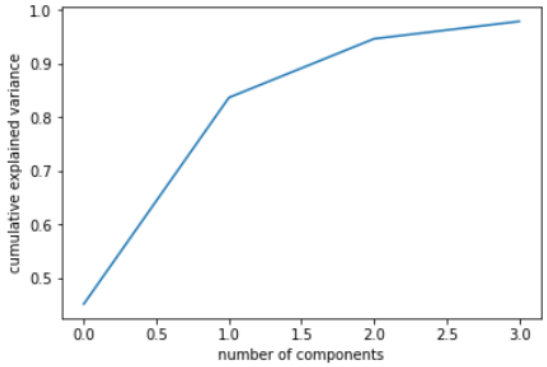

Fig. 7. Plot of Cumulative Explained Variance vs Number of Principal Components

Table 2 shows the explained variance ratio of each principal components. When the number of principal components is 4, 97% of variability was retained. The table also shows the percentage of variability of each principal component holds. Principal component 1 holds most of the variance, having a percentage of 45.18%.

TABLE II. EXPLAINED VARIANCE RATIO OF EACH PRINCIPAL COMPONENTS

Principal Component	***Explained Variance Ratio***
1	0.4518
2	0.3851
3	0.1096
4	0.0324

VI. BASE LEARNER MODELING

A. *Stacked Long-Short Term Memory*

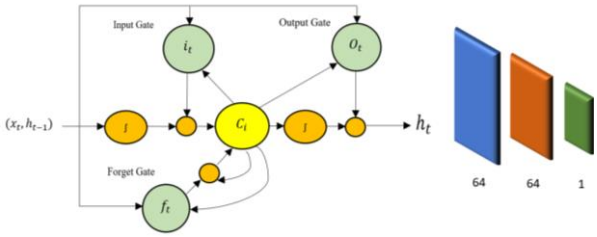

Fig. 8. Left: Long Short Term Memory Architecture Right: Stacked Long Short Term Memory

Figure 8 shows the architecture of long short-term memory (LSTM) and the structure of stacked long-short term memory used in the study. Long short-term memory is suitable for handling time series data because it is capable to store information at any arbitrary period[13]. It is also faster to converge compared to recurrent neural network (RNN) [14][15]. LSTM consists of input gate i_t, output gate o_t and forget gate f_t which manage the information coming in and out of memory cell, and memory cell C_t which is responsible for handling vanishing gradient problem. Each gate possesses activation function which receives the same input as input neurons. Rectified linear unit (ReLu) was used as an activation function because it is simpler to use compared to sigmoid function and produces better results. Stacked LSTM uses 2 LSTM layers with 64 neurons that serve as memory units. Equations (6) to (10) is used to update output values for each time step.

$$f_t = g(W_f x_t + U_f h_{t-1} + b_f) \quad (6)$$

$$i_t = g(W_i x_i + U_i h_{t-1} + b_i) \quad (7)$$

$$c_t = i_t k_t + f_t c_{t-1} \quad (8)$$

$$o_t = g(W_o x_t + U_o h_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \tanh(c_t) \quad (10)$$

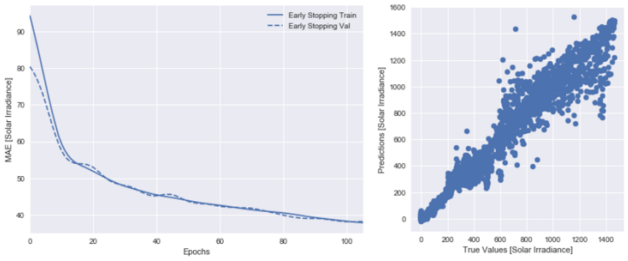

Fig. 9. Training History and Predicted Values of Stacked Long short-term Memory

Figure 9 shows the training history of a single LSTM model with mean absolute error in Y-axis and the plot of its predicted values versus true values of testing set. As observed from the training history plot of a single LSTM model, the validation and training error are close to one another. The validation error is slightly higher than the training error, implying that the model does not overfit. It took about 100 epochs for the model to become stable in learning the input sequence and averages a mean absolute error of 35 units. There are some outliers when observing the plot of predicted values versus true values, however the plot maintains a diagonal like appearance which indicates that most predicted values are close to the real values of testing set.

VII. PERFORMANCE METRICS

This section presents various performance metrics used in assessing the model. The performance metrics considered is usually applied for regression problems.

A. *Explained Variance Score*

Explained variance score (EVS) refers to the the dispersion of a certain dataset, measured in terms of variance. It uses biased variance for explaining the fraction of variance. It is mathematically expressed as

$$explained\ variance(y, \hat{y}) = 1 - \frac{Var(y-\hat{y})}{Var(y)} \quad (11)$$

B. *Maximum Residual Error*

Maximum residual error captures the difference between predicted values and true values. A perfect fitted model has a maximum residual error of 0. It is mathematically expressed as

$$Max\ Error\ (y, \hat{y}) = \max\ (|y_i - \hat{y}_i|) \quad (12)$$

Where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is its corresponding true value.

C. *Mean Absolute Error*

Mean absolute error (MAE) refers to the absolute error. It measures the average absolute difference or magnitude between predicted and actual values. Given that $\hat{y}_i$ is the predicted value of the i-th sample and y_i is its corresponding true value, it is mathematically expressed as

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \quad (13)$$

D. *Root Mean Squared Error*

Mean squared error (MSE) refers to the quadratic error. It measures the average squared difference between the predicted and actual values. Consider that $\hat{y}_i$ is the predicted value of the i-th sample and y_i is its corresponding true value, it is mathematically expressed as

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (14)$$

Root mean squared error is taken by taking the square root of mean squared error.

E. *Regression Score Function*

Regression score function R^2 computes the coefficient of determination. It refers to the goodness of fit and performance of model for predicting samples not included in the training set. A value of 1.0 indicates the best possible score for this metric. A negative value indicates inverse relationship between 2 variables. It uses raw sums of square for explaining the fraction of variance. Given that $\hat{y}_i$ is the predicted value of the i-th sample and y_i is its corresponding true value over n samples, it is mathematically defined as

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (15)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (16)$$

VIII. RESULTS

This section shows the comparison of performance between single learners and bagging-based ensemble learners. On the plot, the blue dot represents the performance of single learners, otherwise it is from bagging-based ensemble learners. The researchers arbitrarily choose 20 base stacked long-short term memory learners for bagging-based ensemble learning.

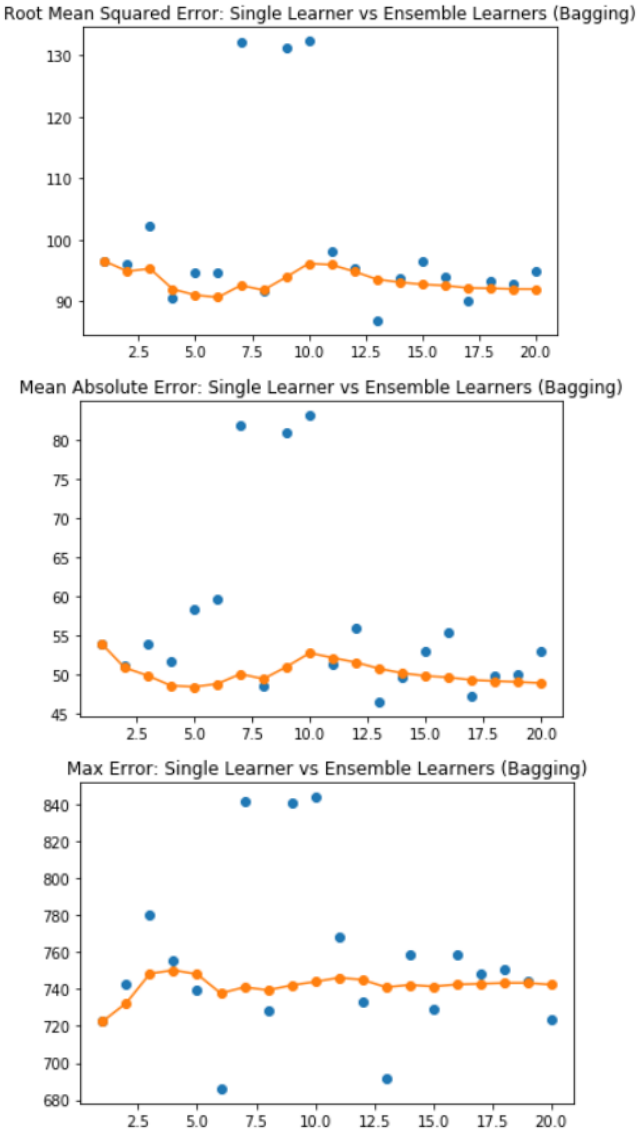

Fig. 10. Comparison of Error Metrics between Single Learners and Bagging Ensembles

Figure 10 shows the plot that shows comparison of the performance between single learners and ensemble learners using the metrics root mean squared error, mean absolute error, and maximum residual error. The plots for the said metrics are arranged from top to bottom. It can be observed from the plot the downward trend of ensemble learners at the 10^{th} iteration. It implies that as iteration increases, the value of error starts to decrease. This can also imply the improvement of accuracy in prediction in comparison with the individual learners. It is noticeable the consistency of the performance of ensemble learners compared to single learner Using the mean absolute error and root mean squared error as metric, the ensemble learners became stable and consistent at 10^{th} iteration. Meanwhile, by using the maximum residual error as y-axis, the ensemble learner starts to become consistent at around 6^{th} iteration.

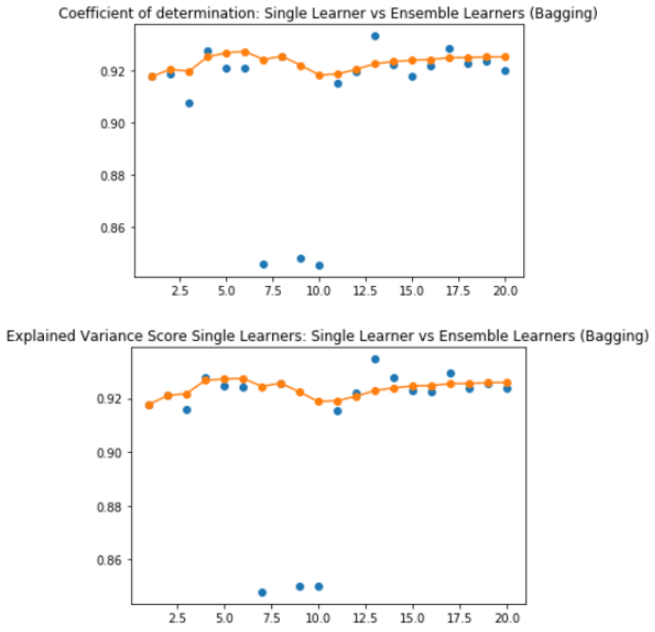

Fig. 11. Comparison of Regression Score and Explained Variance Score bet. Single Learner and Bagging Ensembles

Figure 11 shows the comparison of regression score and explained variance score between single learners and ensemble learners. Regression score and explained variance score are almost the same, except explained variance score can handle the skewness of the residuals. From the plot, it is noted that the ensemble learners start to become stable at 12th iteration. From the 12th iteration, the trend of ensemble learners starts to get higher. There are some points of iterations where single learner outperforms ensemble learners, however, its performance is not consistent. At average, the explained variance score and regression score of ensemble learners average around 0.92, which indicates high estimation of true values found in input sequence.

IX. Conclusion

In this paper, ensemble learning of PCA pre-processed stacked long-short term memory was used to predict solar irradiance using weather parameters as its features. Stacked long-short term memory was used as base learners since it is suitable for inputs such as time series sequence. It also overcome problems such as vanishing gradient problem. Based from the plots that shows the performance comparison of base learners and ensemble learners, it shows that at some iteration, ensemble learners get stable and consistently improves the prediction in terms of metrics used such as mean absolute error, root mean squared error, maximum residual error, explained variance score, and regression score function. It is recommended to use other ensemble learning methods such as stacking or parallel grid search algorithms for prediction problems. Also, it is recommended to add concentration of dust particles as features to consider possible power attenuation.

Acknowledgment

The researchers would like to thank Department of Science and Technology- Engineering and Research Development Technology (DOST-ERDT) and De La Salle University- Intelligent Systems Laboratory (DLSU-ISL) under the project “Smart Farm” for their continuous support.

References

[1] D. Solanki, U. Upadhyay, S. Patel, R. Chauhan, and S. Desai, “Solar Energy Prediction using Meteorological Variables,” *2018 Int. Conf. Recent Innov. Electr. Electron. Commun. Eng. ICRIEECE 2018*, no. 1, pp. 16–19, 2018, doi: 10.1109/ICRIEECE44171.2018.9009175.

[2] A. K. Sahoo and S. K. Sahoo, “Energy forecasting for grid connected MW range solar PV system,” *India Int. Conf. Power Electron. IICPE*, vol. 2016-Novem, 2016, doi: 10.1109/IICPE.2016.8079388.

[3] S. Almazrouei, A. K. Hamid, and M. Shamsuzzaman, “Predictive energy management in large-scale grid connected PV-batteries system,” *5th Int. Conf. Renew. Energy Gener. Appl. ICREGA 2018*, vol. 2018-Janua, pp. 315–318, 2018, doi: 10.1109/ICREGA.2018.8337601.

[4] T. Niimura, S. Member, K. Ozawa, D. Yamashita, and S. Member, “Profiling Residential PV Output based on Weekly Weather Forecast for Home Energy Management System,” *2012 IEEE Power Energy Soc. Gen. Meet.*, pp. 1–5, 2012, doi: 10.1109/PESGM.2012.6345020.

[5] S. Cao, “Total daily solar irradiance prediction using recurrent neural networks with determinants,” *Asia-Pacific Power Energy Eng. Conf. APPEEC*, pp. 25–28, 2010, doi: 10.1109/APPEEC.2010.5448641.

[6] X. D. Zeng, S. Chao, and F. Wong, “Optimization of bagging classifiers based on SBCB algorithm,” *2010 Int. Conf. Mach. Learn. Cybern. ICMLC 2010*, vol. 1, no. July, pp. 262–267, 2010, doi: 10.1109/ICMLC.2010.5581054.

[7] Z. M. Omer and H. Shareef, “Adaptive boosting and bootstrapped aggregation based ensemble machine learning methods for photovoltaic systems output current prediction,” *2019 29th Australas. Univ. Power Eng. Conf. AUPEC 2019*, 2019, doi: 10.1109/AUPEC48547.2019.211856.

[8] D. K. Barrow and S. F. Crone, “Crogging (cross-validation aggregation) for forecasting - A novel algorithm of neural network ensembles on time series subsamples,” *Proc. Int. Jt. Conf. Neural Networks*, 2013, doi: 10.1109/IJCNN.2013.6706740.

[9] F. Vignola, J. Peterson, R. Kessler, M. Dooraghi, M. Sengupta, and F. Mavromatakis, “Evaluation of Photodiode-based Pyranometers and Reference Solar Cells on a Two-Axis Tracking System,” *2018 IEEE 7th World Conf. Photovolt. Energy Conversion, WCPEC 2018 - A Jt. Conf. 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC*, vol. 22, pp. 2376–2381, 2018, doi: 10.1109/PVSC.2018.8547299.

[10] A. Driesse, W. Zaaiman, D. Riley, N. Taylor, and J. S. Stein, “Investigation of pyranometer and photodiode calibrations under different conditions,” *2017 IEEE 44th Photovolt. Spec. Conf. PVSC 2017*, pp. 1–6, 2017, doi: 10.1109/PVSC.2017.8366307.

[11] P. Mair, “Principal Component Analysis and Extensions,” no. Icecs, pp. 179–210, 2018, doi: 10.1007/978-3-319-93177-7_6.

[12] R. Gordillo-orquera, L. M. Lopez-ramos, S. Muñoz-romero, P. Iglesias-casarrubios, D. A. Id, and A. G. M. Id, “Analyzing and Forecasting Electrical Load Consumption in Healthcare Buildings,” pp. 1–18, doi: 10.3390/en11030493.

[13] C. Li, Y. Zhang, and G. Zhao, “Deep Learning with Long Short-Term Memory Networks for Air Temperature Predictions,” *Proc. - 2019 Int. Conf. Artif. Intell. Adv. Manuf. AIAM 2019*, pp. 243–249, 2019, doi: 10.1109/AIAM48774.2019.00056.

[14] S. Anjana, K. Saruladha, and K. Sathyabama, “Bidirectional and stacked LSTM for sleep disorders prediction,” *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, no. Iccmc, pp. 912–916, 2019, doi: 10.1109/ICCMC.2019.8819736.

[15] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series,” *Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018*, no. April 2019, pp. 1394–1401, 2019, doi: 10.1109/ICMLA.2018.00227.