

Comparative analysis of wind speed prediction: enhancing accuracy using PCA and linear regression vs. GPR, SVR, and RNN

Somasundaram Deepa¹, Jayanthi Arumugam², Raguraman Purushothaman³, D. Nageswari⁴,
L. Rajasekhara Babu⁵

¹Department of Electrical and Electronics Engineering, Panimalar Engineering College, Chennai, India

²Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India

³Department of Computer Science & Engineering - (Artificial Intelligence), Madanapalle Institute of Technology & Science, Madanapalle, India

⁴Department of Science and Humanities (General Engineering-EEE), R.M.K. College of Engineering and Technology, Thiruvallur, India

⁵Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

Article Info

Article history:

Received Jul 10, 2024

Revised Oct 9, 2024

Accepted Oct 23, 2024

Keywords:

Energy management

Machine learning

Performance metrics

Regression model

Renewable energy

Wind system

ABSTRACT

For power systems with significant wind power integration to operate in an efficient and dependable manner, wind speed prediction accuracy is crucial. Factors such as temperature, humidity, air pressure, and wind intensity heavily influence wind speed, adding complexity to the prediction process. This paper introduces a method for wind speed forecasting that utilizes principal component analysis (PCA) to reduce dimensionality and linear regression for the prediction model. PCA is employed to identify key features from the extensive meteorological data, which are subsequently used as inputs for the Linear Regression model to estimate wind speed. The proposed approach is tested using publicly available meteorological data, focusing on variables such as temperature, air pressure, and humidity. Popular models like recurrent neural networks (RNN), support vector regression (SVR), and Gaussian process regression (GPR) are used to compare its performance. Evaluation metrics such as root mean square error (RMSE) and R^2 are used to measure effectiveness. Results show that the PCA combined with Linear Regression model yields more accurate predictions, with an RMSE of 94.11 and R^2 of 0.9755, surpassing the GPR, SVR, and RNN models.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Somasundaram Deepa

Department of Electrical and Electronics Engineering, Panimalar Engineering College

Chennai, India

Email: dee_soms123@yahoo.co.in

1. INTRODUCTION

Wind power has become a crucial source of renewable energy, experiencing significant growth worldwide. By 2018, global installed wind power capacity had reached 592 GW, with expectations to surpass 800 GW by 2021 [1]-[5]. A key challenge for wind energy production is the variability and unpredictability of wind speed, which directly impacts the efficiency and integration of wind energy into power grids. Accurate wind speed forecasting is essential for optimizing power systems, maintaining grid stability, and maximizing the economic benefits of wind energy [6]-[8]. However, the unpredictable nature of wind speed makes reliable forecasting a complex task, requiring advanced methods to ensure accuracy [9], [10].

Several forecasting approaches have been developed, each with its advantages and limitations. Physical models, such as numerical weather prediction (NWP), utilize complex mathematical equations to simulate atmospheric dynamics and predict wind patterns. While these models can provide high accuracy, they demand significant computational resources and high-quality data, making them less suitable for short-term or real-time predictions. Additionally, NWP models may perform poorly in areas with unpredictable weather patterns [11]-[14]. On the other hand, statistical methods, including autoregressive integrated moving average or ARIMA, Kalman filters, and Gaussian process regression (GPR), use historical data to make short-term forecasts efficiently. However, they often struggle to model nonlinear relationships between variables, which limits their long-term forecasting capabilities [15]-[17].

Machine learning approaches, such as artificial neural networks (ANNs), backpropagation neural networks (BPNNs), and recurrent neural networks (RNNs), have gained popularity for predicting wind speed due to their capacity to learn complicated, nonlinear correlations from huge datasets [18]-[21]. RNNs, in particular, have demonstrated effectiveness in identifying temporal dependencies in wind data, hence enhancing forecast accuracy. However, these models are sometimes regarded as "black boxes," providing little interpretability into how particular meteorological conditions influence wind speed. Furthermore, machine learning models require significant computational resources and huge datasets, which might be difficult for real-time or resource-constrained applications.

To overcome these issues, a hybrid wind speed prediction method using principal component analysis (PCA) and linear regression is presented. PCA reduces the dimensionality of meteorological data by eliminating redundant features, focusing on the most critical factors influencing wind speed. Using this streamlined dataset, linear regression efficiently predicts wind speed while maintaining accuracy [22], [23]. This approach is computationally efficient, interpretable, and capable of managing multicollinearity in the input data. Compared to more complex models like GPR, support vector regression (SVR), and RNN, the PCA-linear regression model offers a balanced solution, combining efficiency, accuracy, and interpretability, making it a practical choice for wind speed forecasting.

2. METHODOLOGY

The suggested wind speed prediction approach is divided into four major stages: data preprocessing, dimensionality reduction using PCA, wind speed prediction using a linear regression model, and model evaluation. Figure 1 depicts the overall procedure, while each stage is discussed in detail below.

2.1. Data pre-processing

- Data collection and data normalization:

The dataset consists of historical meteorological data, including air pressure, humidity, temperature and wind speed. These factors significantly influence wind speed, making them critical for accurate wind speed prediction. To ensure that all features are on a comparable scale, the data is normalized using the MinMaxScaler, which transforms each feature into the range [0, 1]. Normalization helps the model learn the relationships between the variables by preventing features with broader ranges from controlling the learning process [24], [25]. This is how the normalization is carried out, as in (1).

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \quad (1)$$

Where X is the original data, Xmin and Xmax are the minimum and maximum values of X, respectively.

2.2. Dimensionality reduction

- PCA

The normalized meteorological data's dimensionality is decreased by the use of PCA. PCA removes unnecessary information from the original data while preserving the majority of its variation by converting it into a set of uncorrelated principal components. This step reduces the computational complexity of the model while preserving essential information for wind speed prediction. The transformation is defined as in (2):

$$Z = X_{\text{norm}} \cdot W \quad (2)$$

The major component matrix is Z, the normalized data matrix is Xnorm, and the eigenvector matrix is W. The number of principle components preserved is determined by the total explained variance. Typically, components accounting for 95% of the total variance are chosen to strike a balance between dimensionality reduction and information retention.

2.3. Wind speed prediction

- Linear regression model

The reduced feature set obtained from PCA is used as input for the linear regression model. Linear regression establishes a linear relationship between the principal components and the wind speed, which is expressed by the (3):

$$y = \beta_0 + \sum_{i=1}^n \beta_i Z_i \quad (3)$$

where y is the predicted wind speed Z_i are the principal components, and β_i are the regression coefficients.

2.4. Model evaluation

The dataset is split 75-25, with 75% utilized for model training and 25% for testing. Cross-validation methods such as k-fold cross-validation are used to evaluate model robustness and reduce overfitting. Metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R^2 are used to assess model performance, including prediction accuracy and fit. MSE measures the average of squared errors, whereas R^2 evaluates the model's ability to explain variance in data. The following measures examine the effectiveness of regression models, and the following metrics were used to evaluate performance.

Figure 1 depicts the block diagram of the proposed method. The wind speed prediction process entails gathering historical meteorological data (temperature, humidity, air pressure, and wind speed) and normalizing it with MinMaxScaler to assure feature comparability. The data is then divided into training and test sets. PCA is used to reduce dimensionality by retaining the most important features and removing redundant information. A linear regression model is trained on the PCA-transformed features to forecast wind speed. The model's performance on the testing set is assessed using the RMSE and R-squared metrics.

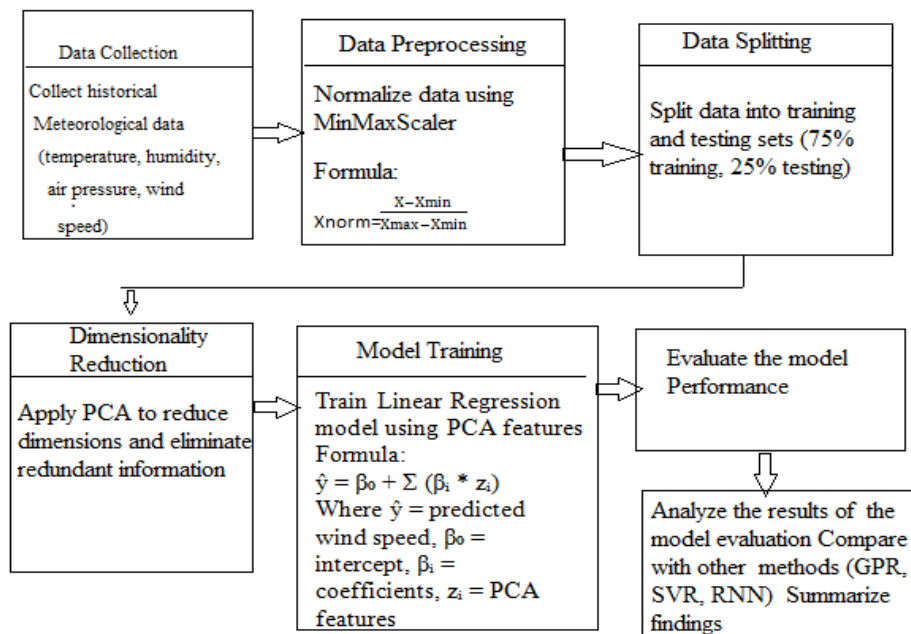


Figure 1. Block diagram of proposed method

2.5. Comparison with other methods

To assess the performance of the proposed PCA-linear regression approach, it is compared with several widely adopted wind speed prediction methods, such as GPR, SVR, and RNN.

- Gaussian process regression (GPR): GPR is a probabilistic model that provides a flexible, non-linear regression approach based on Gaussian distributions. It is known for its ability to quantify uncertainty in predictions but can be computationally intensive.

- Support vector regression (SVR): SVR constructs a hyperplane in a high-dimensional space to model the relationship between input features and wind speed. It is effective in capturing non-linear relationships but requires careful tuning of hyperparameters.
- Recurrent neural networks (RNN): Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) models, are highly effective for time-series prediction due to their capability to learn and retain temporal patterns within sequential data. However, they can be resource-intensive and often require substantial datasets to achieve optimal performance. To achieve a fair comparison, all approaches' performance is tested with the same dataset and metrics (RMSE and R^2). The findings show that the PCA-linear regression strategy surpasses the other methods in terms of both accuracy and computing economy, especially when dealing with high-dimensional meteorological data.

3. RESULTS AND DISCUSSION

The results from the linear regression model, including predicted and actual active power values for both the training and testing data, are visualized in the plot shown in Figure 2. Training data: the actual active power (shown in blue) and predicted active power (orange) are plotted over time. The model fits the training data well, indicating that the regression model has learned the underlying relationship between meteorological features and the target variable. Testing data the actual active power (green) and predicted active power (red) are plotted for the testing period. The predictions closely follow the actual data, demonstrating the model's ability to generalize to new, unseen data.

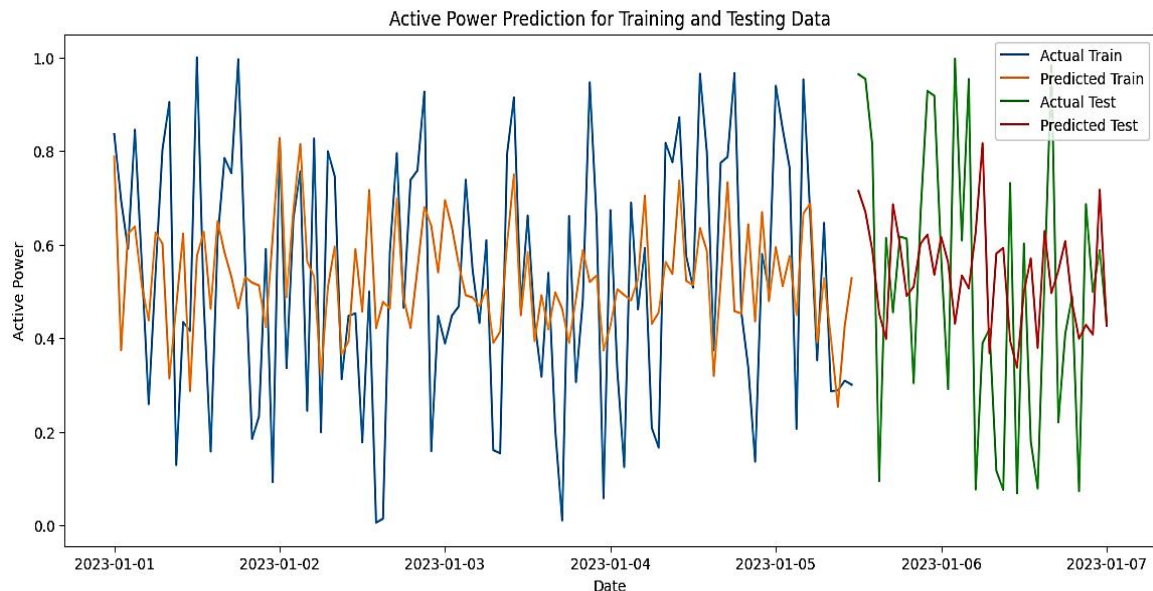


Figure 2. Shows the active power prediction for training and testing data

The updated code includes visualizations for active power prediction, enhancing understanding of model performance. Figure 3 shows both actual and predicted active power values for training and testing data, with confidence intervals indicated by shaded areas. Figure 4 illustrates the distribution of residuals, comparing training and testing residuals with Kernel density estimation or KDE. These plots provide insights into prediction accuracy and the residuals' distribution, highlighting the model's reliability and areas where it might be improved. In this study, we investigated the performance of four distinct approaches for predicting wind speed: PCA + linear regression, GPR, SVR, and recurrent neural networks. The training and testing datasets were compared using the performance measures RMSE and coefficient of determination (R^2). The findings are summarized in Table 1. Figure 5 displays a comparison of different algorithms.

Table 1. Shows the comparison of different algorithm

| Method | RMSE (Training Set) | R^2 (Training Set) | RMSE (Testing Set) | R^2 (Testing Set) |
|-----------------------------------|---------------------|----------------------|--------------------|---------------------|
| PCA + linear regression | 95.59 | 0.9745 | 94.11 | 0.9755 |
| Gaussian process regression (GPR) | 96.65 | 0.9730 | 96.40 | 0.9739 |
| Support vector regression (SVR) | 94.52 | 0.9681 | 99.02 | 0.9692 |
| Recurrent neural network (RNN) | 94.87 | 0.9754 | 92.12 | 0.9752 |

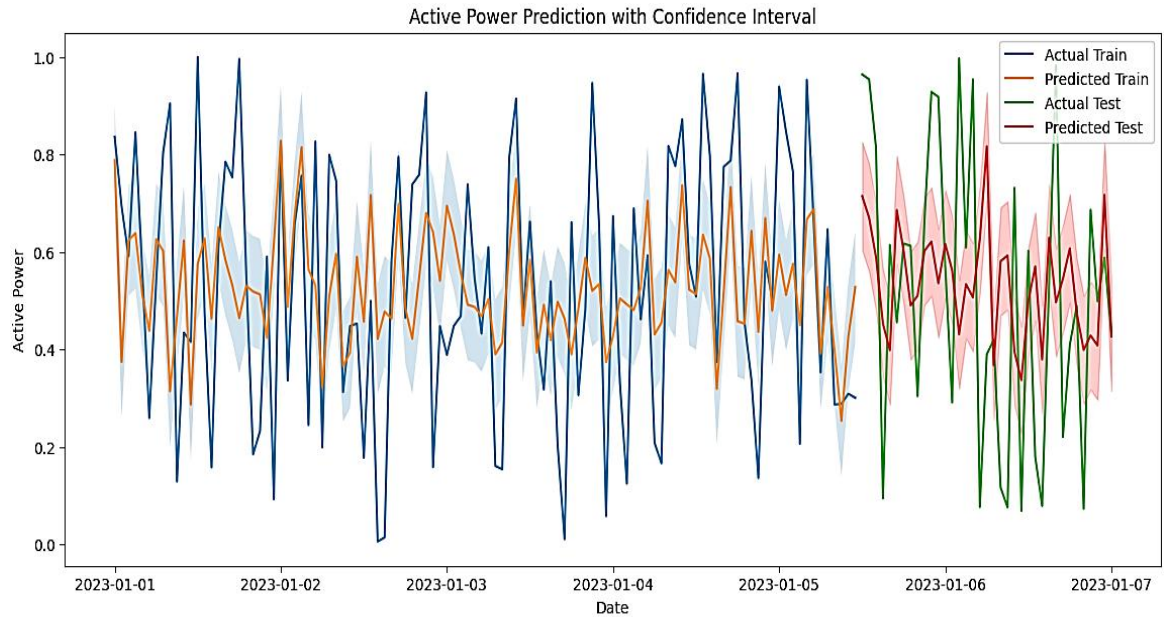


Figure 3. Shows the actual power prediction with confidence level

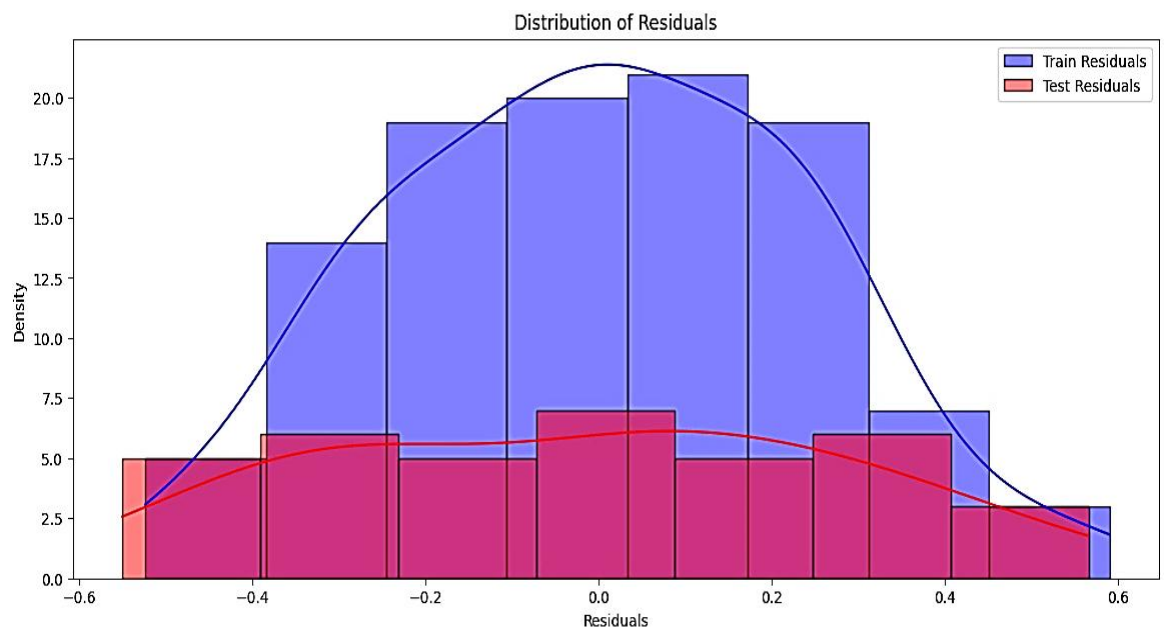


Figure 4. Shows the distribution of residuals

- PCA + linear regression:

The combined method of PCA and linear regression demonstrated superior overall performance, with an RMSE of 95.59 on the training data and 94.11 on the test data. Additionally, the R^2 values were 0.9745 for training and 0.9755 for testing, indicating a high level of accuracy and strong generalization to new data. The integration of PCA effectively reduced input dimensionality, enhancing both prediction accuracy and computational efficiency.

- GPR

GPR performed reasonably well, yielding an RMSE of 96.65 for training and 96.40 for testing, along with R^2 values of 0.9730 and 0.9739, respectively. While GPR displayed solid predictive ability, its accuracy fell slightly short of the PCA + linear regression model. The marginally higher RMSE values suggest that GPR may not capture data intricacies as effectively.

- SVR

Among the tested methods, SVR had the lowest performance, with an RMSE of 94.52 for the training set and 99.02 for the testing set. The R^2 values were 0.9681 for training and 0.9692 for testing. Despite the low training RMSE, the significant increase in test error points to overfitting, indicating that SVR may struggle with generalization in this wind speed prediction task.

- RNN

RNN also achieved strong results, with an RMSE of 94.87 on the training set and 92.12 on the testing set. The R^2 values were 0.9754 and 0.9752, respectively. These values are close to those of PCA + linear regression, suggesting that RNN is highly competitive for wind speed prediction. Its slightly lower RMSE on the test set implies that RNN may capture certain patterns more effectively.

In summary, the comparison of these methods highlights PCA + linear regression as the best performer in terms of both accuracy and error minimization. While GPR and RNN also showed competitive results, SVR lagged behind, primarily due to overfitting issues. Overall, PCA + linear regression emerges as a reliable and efficient solution for handling high-dimensional meteorological data and delivering accurate wind speed forecasts.

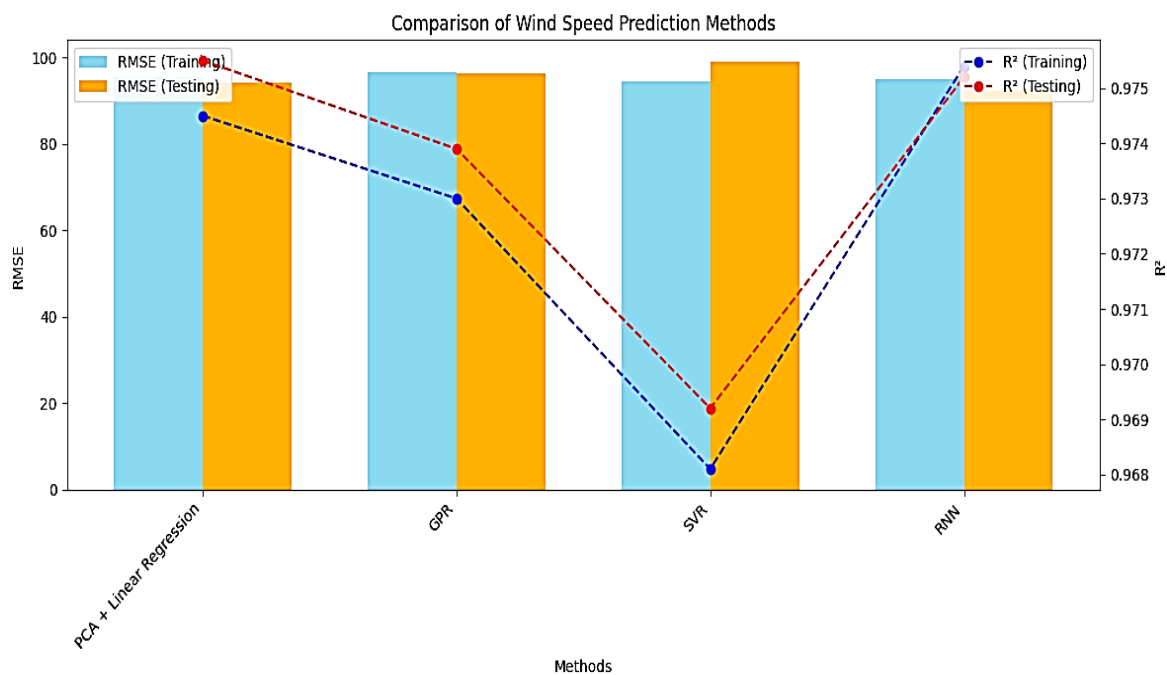


Figure 5. shows the comparison of different algorithm

4. CONCLUSION

This study showcases the effectiveness of combining PCA with linear regression for wind speed prediction. PCA's dimensionality reduction and feature selection via SelectKBest resulted in a model with an RMSE of 94.11 and an R^2 of 0.9755 on the testing set, outperforming GPR, SVR, and RNN in both accuracy and computational efficiency. The model's practical implications include improved wind energy integration and operational planning. However, limitations such as reliance on historical data and challenges with real-time adaptation are noted. Future research should explore incorporating additional variables, advanced machine learning techniques, and real-time application adaptations. This approach advances wind speed forecasting, benefiting energy providers and policymakers in optimizing wind power systems.




REFERENCES

- [1] A. L. Mahmood, A. M. Shakir, and B. A. Numan, "Design and performance analysis of stand-alone PV system at al-nahrain university, Baghdad, Iraq," *International Journal of Power Electronics and Drive Systems*, vol. 11, no. 2, pp. 921–930, 2020, doi: 10.11591/ijpeds.v11.i2.pp921-930.
- [2] G. Nguyen et al., "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.




- [3] S. Deepa, *et al.*, "Machine learning applications for predicting system production in renewable energy," *International Journal of Power Electronics and Drive Systems*, vol. 15, no. 3, pp. 1925–1933, 2024, doi: 10.11591/ijpeds.v15.i3.pp1925-1933.
- [4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
- [5] R. Ayop *et al.*, "The performances of partial shading adjuster for improving photovoltaic emulator," *International Journal of Power Electronics and Drive Systems*, vol. 13, no. 1, pp. 528–536, 2022, doi: 10.11591/ijpeds.v13.i1.pp528-536.
- [6] N. Priyadarshi, S. Padmanaban, M. S. Bhaskar, F. Blaabjerg, and J. B. Holm-Nielsen, "An improved hybrid PV-wind power system with MPPT for water pumping applications," *International Transactions on Electrical Energy Systems*, vol. 30, no. 2, p. e12210, 2020, doi: 10.1002/2050-7038.12210.
- [7] V. Maheswari, *et al.*, "Theoretical and simulation analysis of first generation DC-DC converters," *International Journal of Advanced Science and Technology*, vol. 28, no. 19, pp. 72–78, 2019.
- [8] S. Deepa, N. Anipriya, & R. Subbulakshmy, "Design of controllers for continuous stirred tank reactor". *International Journal of Power Electronics and Drive Systems*, vol. 5, no. 4, pp. 576–582, 2015, doi: 10.11591/ijpeds.v5.i4.pp576-582.
- [9] E. H. M. Ndiaye, A. Ndiaye, M. Faye, and S. Gueye, "Intelligent control of a photovoltaic generator for charging and discharging battery using adaptive neuro-fuzzy inference system," *International Journal of Photo Energy*, vol. 2020, 2020, doi: 10.1155/2020/8649868.
- [10] Y. E. García-Vera, R. Dufo-López, and J. L. Bernal-Agustín, "Optimization of isolated hybrid microgrids with renewable energy based on different battery models and technologies," *Energies*, vol. 13, no. 3, p. 581, 2020, doi: 10.3390/en13030581.
- [11] Y. Wu *et al.*, "Large scale incremental learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June, 2019, pp. 374–382, doi: 10.1109/CVPR.2019.00046.
- [12] A. Mosavi, S. Shamshirband, E. Salwana, K. wing Chau, and J. H. M. Tah, "Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 482–492, 2019, doi: 10.1080/19942060.2019.1613448.
- [13] Wang, Z., Liu, K., Li, J., Zhu, Y., & Zhang, Y., "Various Frameworks and Libraries of Machine Learning and Deep Learning: A Survey," *Archives of Computational Methods in Engineering*, vol. 31, pp. 1-24, 2024, doi: 10.1007/s11831-018-09312-w.
- [14] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [15] J. R. Saura, B. R. Herrera, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100–37108, 2019, doi: 10.1109/ACCESS.2019.2905301.
- [16] Deepa, S., Praba, S., Deepalakshmi, V., Jayaprakash, L., & Manimurugan, M., "A fuzzy GA based STATCOM for power quality improvement," *International Journal of Power Electronics and Drive Systems*, vol. 8, no. 1, pp. 483–491, 2017, doi: 10.11591/ijpeds.v8.i1.pp483-491.
- [17] Awasthi, A., Shukla, A. K., Murali Manohar, S. R., Dondariya, C., Shukla, K. N., Porwal, D., & Richhariya, G., "Review on sun tracking technology in solar PV system. Energy Reports," *Energy Reports*, vol. 6, pp. 392-405, 2020, doi: 10.1016/j.egyr.2020.02.004.
- [18] AL-Rousan, N., Mat Isa, N. A., & Mat Desa, M. K., "Efficient single and dual axis solar tracking system controllers based on adaptive neural fuzzy inference system," *Journal of King Saud University - Engineering Sciences*, vol. 32, no. 7, pp. 459–469, 2020, doi: 10.1016/j.jksues.2020.04.004.
- [19] M. Kharrich *et al.*, "Optimal design of an isolated hybrid microgrid for enhanced deployment of renewable energy sources in Saudi Arabia," *Sustainability*, vol. 13, no. 9, p. 4708, 2021, doi: 10.3390/su13094708.
- [20] Jose A. Carballo, Javier Bonilla, Lidia Roca, Manuel Berenguel, "New low-cost solar tracking system based on open source hardware for educational purposes," *Solar Energy*, vol. 174, , 826-836, 2018.
- [21] Jose A. Carballo, Javier Bonilla, Manuel Berenguel, Jesús Fernández-Reche, Ginés García, "New approach for solar tracking systems based on computer vision, low cost hardware and deep learning," *Renewable Energy*, vol. 133, pp. 1158-1166, 2019.
- [22] Reddy, S. R., "A machine learning approach for modeling irregular regions with multiple owners in wind farm layout design," *Energy*, vol. 220, 2021, doi: 10.1016/j.energy.2020.119691.
- [23] M. Seeger, "Gaussian processes for machine learning," *International Journal of Neural Systems*, vol. 14, no. 2, pp. 69–106, 2004, doi: 10.1142/S0129065704001899.
- [24] L. Dhanesh, *et al.*, "Enhanced and Energy-Efficient Program Scheduling for Heterogeneous Multi-Core Processors System," *Lecture Notes in Electrical Engineering*, vol. 665, pp. 737–747, 2020, doi: 10.1007/978-981-15-5262-5_55.
- [25] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, "Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs," *Sensors*, vol. 19, no. 9, 2019, doi: 10.3390/s19092206.

BIOGRAPHIES OF AUTHORS



Somasundaram Deepa    received her B.E. from K.S.R College of Technology, affiliated to Periyar University, in 2003, M.E from Annamalai University in 2005. She completed her Ph.D. degree from Sathyabama university in 2013. Presently, she is working as a professor in the Department of EEE at Panimalar Engineering College, Chennai. She has published a more than 30 papers in international and national journals. Her area of interest is power system, optimisation technique. She has more than 15 years of experience in teaching field. She can be contacted at email: dee_soms123@yahoo.co.in.






Jayanthi Arumugam    is currently working as assistant professor in the Department of Computer Science and Engineering at Velammal Engineering College, Chennai, India. Her research interests include data mining and machine learning. She can be contacted at email: jayanthiarumugamk@gmail.com.






Raguraman Purushothaman    is currently working as an assistant professor in the Department of CSE (artificial intelligence), Madanapalle Institute of Technology & Science, Angallu, Madanapalle, Andhra Pradesh. His area of interest includes theory of computation, design and analysis of algorithms, image processing and data science. He can be contacted at email: yuvaragu.pt@gmail.com.



D. Nageswari    is an assistant professor at R.M.K. College of Engineering and Technology, holds a Ph.D. in Electrical and Electronics Engineering from Anna University (2022) and an M.E. in Power Electronics and Drives from R.M.K. Engineering College, where she received the University Gold Medal. She has published over 18 papers, is a member of four professional societies, and her research focuses on optimization techniques, artificial intelligence, and smart grids. She can be contacted at email: nageswari@rmkcet.ac.in.



L. Rajasekhara Babu    is an assistant professor in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation (KLEF), Andhra Pradesh. He holds a Ph.D. in Computer Science from Bharathidasan University and has expertise in medical data mining, deep learning, and bioinformatics. With numerous publications and awards, his research spans in machine learning, big data, and ontology development. He can be contacted at email: rajasekharlingisetty@kluniversity.in.