RESEARCH ARTICLE

# Multilevel Stacking with Domain-Adversarial Neural Network and Enriched Metadata for Improved Renewable Energy Output Forecasting

Prince Clement Addo[1*], Daniel Dankwah[1], Joshua Caleb Dagadu[1], Samuel Kofi Akpatsa[1], Oliver Kufour Boansi[1], William Asiedu[1], George Asante[1], Enoch Adjei Frimpong[2], Dong Miao[3], William Agyei Atuahene[1]

[1]Department of Information Technology, University of Skill Training and Entrepreneurial Development, Ghana
[2]Department of Computer Science, Cape Coast Technical University, Ghana
[3]School of Computer Science, University of Electronic Science and Technology of China, China
*Corresponding Author: [pcaddo@aamusted.edu.gh]

**To cite this article:** Addo, P. C., Dankwah, D., Dagadu, J. C., Akpatsa, S. K., Boansi, O. K., Asiedu, W., Asante, G., Frimpong, E. A., Miao, D., & Atuahene, W. A. (2026). Multilevel stacking with domain-adversarial neural network and enriched metadata for improved renewable energy output forecasting. Journal of Technology Education and Applied Sciences, 1(1), 32–45. https://doi.org/10.64922/jteas.v1i1.86

**ABSTRACT**

This study proposes a multilevel stacking approach that extends Wolpert's (1992) stacked generalisation, aiming to minimise forecasting error, reduce systematic bias, and enforce temporal domain invariance. It incorporates walk-forward out-of-fold and step-based walk-forward validation, enriched metadata, and a domain adversarial neural network (DANN). The base learners XGB and LSTM function as residual and latent encoders for the Random Forest with DANN meta-learner. DANN is integrated in meta-learning to minimise task regression loss and maximize domain discrimination loss through a gradient reversal layer. This fusion improves domain transition stability, as the model's high explanatory power with ^2 values exceed 0.98 across validation and test regimes. The MASE values below 0.07 on both regimes indicate strong predictive performance relative to a naïve benchmark, and NRMSE values around 0.025 confirm that the proposed stacking framework effectively captures temporal dependencies. Moreover, the model achieves a near-symmetric residual distribution, with a skewness of 0.37 on the validation set and -0.01 on the test set, suggesting that adversarial feature learning helps reduce asymmetric prediction errors and stabilise meta-learning. This proposed framework demonstrates empirically that stacking should not be evaluated solely on MAE/RMSE but on residual distribution behaviour under domain transition. Furthermore, SHAP analysis provides empirical evidence that enriched metadata significantly enhances the learning capacity, allowing the meta-learner to capture complex dependencies.

**KEYWORDS**

Multilevel Stacking, Renewable Energy Forecasting, Domain Adversarial Neural Network, Enriched Metadata, Walk-Forward Validation

***Publication History***

Received: 14 November 2025
Revised: 28 November 2026
Accepted: 5 April 2026
Available online: 30 July 2026

## Introduction

The stacked ensembles have improved generalisation in static contexts. Their direct application to sequential and time-dependent data is hindered by data leakage, concept drift, and structural irregularities (Tsymbal, 2004). Recent studies have attempted to incorporate temporal information into heterogeneous ensembles such as tree-based, convolutional, and recurrent networks (Zhang et al., 2018). This underscores the significance of base-learner diversity for predictive performance (Middlehurst et al., 2021). However, balancing accuracy, robustness, or interpretability is challenging, particularly for non-stationary energy data. The enriched metadata has contributed to improving ensemble performance, but systematic implementation in renewable energy domains is limited (Cerqueira et al., 2020). Studies explore the use of Random Forest as a meta-learner within a stacking framework. This is due to the Random Forest's ability to capture nonlinear dependencies in a diverse set of learners' outputs. An empirical study has confirmed that tree-based meta-learners are effective in learning complex relationships among predictive vectors (Krauss et al., 2017).

Moreover, a study on credit risk assessment shows that an ensemble of tree-based models outperformed a standalone model in predictive reliability and robustness (Lessmann et al., 2015). These findings collectively suggest that Random Forest meta-learning provides a flexible mechanism for synthesizing heterogeneous predictive signals while mitigating overfitting through ensemble diversity. On the contrary, there have been methodological challenges that sideline the reliability of stacking with the Random Forest meta-learner in dynamic regimes. The data distribution often evolves due to seasonal variations and distributional shifts. The predictive relationships learned by base-learners and the meta-learner become unstable, leading to degradation in forecasting accuracy (Tsymbal, 2004).

Even though Random Forest can capture nonlinear dependencies among base model predictions, it does not inherently and explicitly address the distributional mismatch that arises when the statistical properties of the data shift across regimes. This study proposes a multilevel stacking with a Domain Adversarial Neural Network (DANN), unlike traditional stacking methods that aggregate base learners without explicitly addressing temporal nonstationarity. Integrating a DANN in meta-learning extracts domain-invariant features from sequential data. By introducing a gradient reversal layer, the framework actively discourages reliance on domain-specific signals. This approach provides two major advantages. First, it enhances stability in predictions by reducing the impact of distributional shifts across different time periods, ensuring the meta-learner generalizes well to unseen sequences.

Second, it refines the meta-features, including base learner residuals, LSTM latent embedding, and engineered features, by filtering out domain-dependent noise, thereby improving interpretability and robustness. To the best of our knowledge, this study represents the first integration of DANN as a regularisation and loss function in meta-learning. By combining residual-aware features, temporal latent representations, and domain adversarial learning, the framework addresses both the accuracy and stability challenges of renewable energy prediction.

## Method

### Problem Formulation

Let the renewable and weather condition dataset be represented as a chronological time series.

$$D = \{(X_t, y_t)\}_{t=1}^{T}, \; X_t \in R^d, y_t \in R$$

Where $\boldsymbol{X_t} \in \boldsymbol{R^d}$ denotes the feature vectors of the base and engineered features, and $\boldsymbol{y_t} \in \boldsymbol{R}$ represents the target variable (Energy delta [wh]), which corresponds to energy production change. This dataset is a structured time series dataset sourced from Kaggle, a reputable online platform (https://www.kaggle.com/datasets/samanemami/renewable-energy-and-weather-conditions). The data spans five years from 2017 to 2022, with 196,776 records shown in Table 1.

**Table 1**

*Descriptive statistics of Numerical Values in renewable and Weather conditions Data*

| *Numerical variables* | *Count* | *Mean* | *Std.* | *Minimum* | *25%* | *50%* | *75%* | *Maximum* |
|---|---|---|---|---|---|---|---|---|
| Energy delta [Wh] | 196776 | 5.7300 | 1044.82 | 0.00 | 0.00 | 0.00 | 577.00 | 5020.00 |
| GHI | 196776 | 3.2596 | 52.172 | 0.00 | 0.00 | 1.600 | 46.80 | 229.20 |
| Temperature | 196776 | 9.7905 | 7.9954 | -16.60 | 3.600 | 9.300 | 15.700 | 35.800 |
| Pressure | 196776 | 1.0152 | 9.5857 | 977.0 | 1010.0 | 1016.0 | 1021.0 | 1047.0 |
| Humidity | 196776 | 7.98105 | 15.604 | 22.00 | 70.00 | 84.00 | 92.00 | 100.0 |
| Sunlight Time | 196776 | 211.721 | 273.90 | 0.00 | 0.00 | 30.00 | 390.00 | 1020.00 |
| Sunlight_Time/day_length | 196776 | 2.65186 | 0.3290 | 0.00 | 0.00 | 0.050 | 0.5300 | 1.00 |
| Hour | 196776 | 1.14989 | 6.9218 | 0.00 | 5.00 | 11.00 | 17.00 | 23.00 |

The study assumes that the dataset follows a normal distribution. If the data deviates from this assumption, it may warrant the use of transformation or non-parametric modelling approaches.

The study further conducts a data distribution test using the Shapiro-Wilk test, the Kolmogorov-Smirnov (KS) test, and the Anderson-Darling test.


https://doi.org/10.64922/jteas.v1i1.86

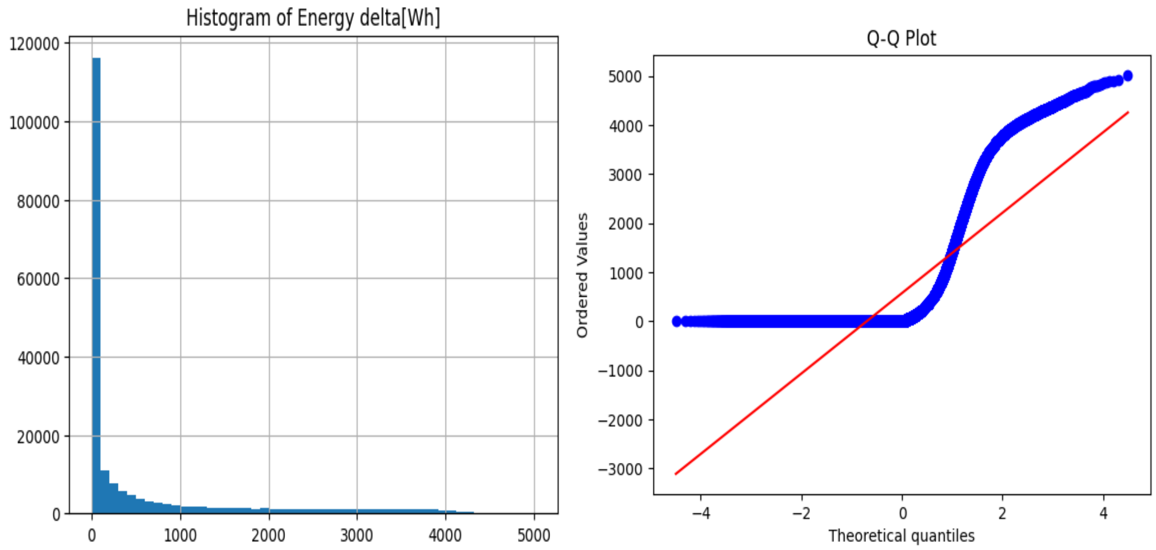
The Shapiro-Wilk test shows a test statistic ($W$) of 0.616 and a $p$-value less than 0.001. This value falls significantly below the conventional alpha level of 0.05. This rejects the null hypothesis, indicating a considerable deviation of the target variable from normality.

Similarly, the Kolmogorov-Smirnov (KS) test shows a test statistic of 0.292 with a corresponding $p$-value less than 0.001. This result confirms that the distribution of the target variable does not align with the characteristics of a Gaussian distribution.

The Anderson-Darling test yields a test statistic of 32,018.751, which substantially exceeded all critical values at the 15%, 10%, 5%, 2.5%, and 1% significance levels. This finding rejects the null hypothesis and reinforces the conclusions drawn from the Shapiro-Wilk and KS tests. The study ascertains that the results of all three statistical tests consistently demonstrate that the target variable is not normally distributed. Given the apparent non-Gaussian nature of the data, it is therefore applicable to consider a data transformation technique or a non-parametric model in the proposed multilevel stacking framework in Figure 1.

**Figure 1**

*Histogram and Q-Q plot depict the Data Distribution*




The forecasting objective of the study is to learn a mapping $\mathcal{F}: \boldsymbol{X_t} \rightarrow \boldsymbol{Y_t}$ that minimizes prediction error under non-stationary temporal conditions. However, the renewable energy and weather condition dataset exhibits an inherent

concept drift:

$$P_t(y|x) \neq P_{t+\triangle}(y|x) \quad (1)$$

distributional instability:

$$P_t(x) \neq P_{t+\triangle}(x) \quad (2)$$

and a domain-induced variance caused by evolving weather regimes. The stacking approach, which uses a Random Forest as a meta-learner, assumes independent and identically distributed observations. Consequently, their performance degrades when applied to sequential data distributional instability.

This study proposes a multilevel stacking that integrates gradient boosted residual modelling, temporal latent feature extraction, and adversarial domain adaptation. The study employs two complementary validation strategies: walk-forward out-of-fold (OOF) on the training set and step-based walk-forward validation on the validation and testing sets.



https://doi.org/10.64922/jteas.v1i1.86

The walk forward OOF on the training set serves as a leakage-free prediction for every time point. Each observation is predicted only by a model trained strictly on past data folds, ensuring temporal causality is respected (Tashman, 2000).

The step-based validation is performed on both the validation and test sets to mimic a forecast of a real-world deployment. The evaluation replicates a rolling adaptation, testing how models perform under concept drift and non-stationarity (Gama et al., 2014).

**Base Level Model Formulation**

The study employs an Extreme Gradient Boosting (XGB) as one of the two base-level learners to model nonlinear relationships between base features and target variables.

The boosting model predicts:

$$\hat{y}_t = \sum_{k=1}^{k} f_k (X_t) \tag{3}$$

Where $\boldsymbol{f_k}$ represents decision trees learned sequentially.

The residual errors are computed as:

$$r_t = y_t - \hat{y}_t \tag{4}$$

to capture unexplained variance and structural irregularities in the dataset. Residual dynamics are subsequently used as a meta-feature to guide the meta-learner. The lagged residuals are constructed as:

$$r^{(l)} = r_{t-1}, l = 1,2,3 \tag{5}$$

allowing the meta-learner to capture temporal persistence in prediction errors.

Moreover, as the tree-based models struggle with sequential dependencies, the study employs Long Short-Term Memory (LSTM) as a base learner to model and extract temporal latent features that are not observable in static features.

For a sequence of length $\boldsymbol{L}$

$$X_t = [X_{t-L+1}, \dots X_t] \tag{6}$$

The LSTM call updates are defined as:

$$f_t = \sigma(W_t[h_{t-1}, X_t] + b_f) \tag{7}$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{8}$$

$$\acute{c}_t = tanh(W_c[h_{t-1}, X_t] + b_c) \tag{9}$$

$$C_t = f_t \odot C_{t-1} + j_t \odot \acute{c}_t \tag{10}$$

$$h_t = O_t \odot \tanh(C_t) \tag{11}$$

The latent representation extracted from the bottleneck layer is

$$H_t \in R^k \tag{12}$$

which summarizes temporal dependencies in the input sequence.

The metadata combines complementary information sources such as residual dynamics, temporal latent features, weather condition and engineered features to form an enriched metadata. The enriched metadata is:

$$Z_t = [r_t^{m_1}, r_{t-1}^{m_1}, r_{t-2}^{m_1}, r_{t-3}^{m_1}, H_t, X_t] \tag{13}$$



https://doi.org/10.64922/jteas.v1i1.86

This enriched metadata allows the meta-learner to model predictive uncertainty, temporal structure, and weather context.

**Meta-Level Model Formulation**

The study employs Random Forest as the meta-learner to model nonlinear relationships. The prediction function of a Random forest is:

$$\hat{y}_t = \frac{1}{B}\sum_{b=1}^{B} T_b\,(Z_t) \qquad (14)$$

Where $T_b$ denotes an individual decision tree. Notwithstanding the strength of Random Forest, it suffers from sensitivity to distributional shifts,

$$P_{train(M)} \neq P_{test(M)} \qquad (15)$$

The inability to explicitly model temporal dependencies and vulnerability to domain-induced variance. These challenges prompted the integration of the domain adaptation mechanism. The Domain Adversarial Neural Network employed consists of a feature extractor$F(\bullet)$, a task predictor $G(\bullet)$, and a domain classifier $D(\bullet)$.

The adversarial training objective is to minimise prediction loss while maximizing domain confusion.

Task loss:

$$\mathcal{L}_y = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \qquad (16)$$

Domain classification loss:

$$\mathcal{L}_d = -\sum_{t=1}^{N} d_t \log(\hat{d}_t) \qquad (17)$$

And the combined objective:

$$\mathcal{L} = \mathcal{L}_y - \lambda\mathcal{L}_d \qquad (18)$$

Where $\boldsymbol{\lambda}$ controls the strength of adversarial learning. A gradient reversal layer multiplies the gradient by $-\boldsymbol{\lambda}$ during backpropagation, forcing the feature extractor to learn domain-invariant representations. The domain classifier output provides a probability estimate:

$$P(d_t|v_t)$$

Indicates the likelihood that a sample belongs to a shifted domain.

Samples are reweighted as:

$$w_t = 1 - P(d_t|v_t) \qquad (19)$$

These weights are incorporated into Random Forest training, emphasizing samples with stable distributions while reducing the influence of drifted observations. The final meta-learning stage trains the Random Forest using the purified features.

$$v_t = f(Z_t) \qquad (20)$$

and importance weights $\boldsymbol{w_t}$.

The final predictor:

$$\hat{y}_t = \frac{1}{B}\sum_{b=1}^{B} T_b\,(V_t) \qquad (21)$$


https://doi.org/10.64922/jteas.v1i1.86

This domain-adaptive learning process reduces prediction instability caused by evolving regimes.

The main objective function is to minimise forecasting error, reduce systematic bias, and enforce temporal domain invariance.

Thus:

$$\theta_{\hat{y}_t} = arg_{\theta_{\hat{y}_t}}^{min} \frac{1}{N} \sum_{t=1}^{N} \omega_t (y_t - \hat{y}_t(v_t, \theta_{\hat{y}_t}))^2 \quad (22)$$

where:

- $\theta_{\hat{y}_t}$represents the parameters of the Random Forest meta-learner.
- $N$ is the number of training samples.

---

Proposed Multilevel Stacking with DANN Algorithm

---

Input:

Renewable energy and Weather condition dataset $D = \{(X_t,\ y_t)\}_{t=1}^{T}$, where features $X_t \in \mathbb{R}^d$, target $y_t \in \mathbb{R}$.

Engineered features $\ell_t$={ lags, anomaly, differences of weather features }

Base learners $\mathcal{M}_j = \mathcal{M}_1 and\ \mathcal{M}_2$ and Meta-learner $\mathcal{M}_l$

Step size $S = 240$ and number of walk forward folds $K = 5$

Sequence length =5, Number of domains $D_n = 4$, and Latent dimension $p = 8$

Step 1: Walk forward Out of fold (OOF) on the Training set

1.1 Partition training set $D_{train}$ into $K$ chronological folds

1.2 For each $K = 1 \dots k$:

Define history $D_{1:(k-1)}$ as training, fold $k$ as validation

Train each base learner $\mathcal{M}_1 = XGB$ on $D_{1:(k-1)}$

Predict $\hat{y}_k^{(j)}$ on fold $k$

Compute residual $r_k^{(j)} = y_k - \hat{y}_k^{(j)}$

Train base learner $\mathcal{M}_2 = LSTM$ on $X_{t-L:t} \to \hat{y}_t$

Extract latent representation from hidden layers

$H_t = \Phi(X_{t-L:t}) \in \mathbb{R}^p$

Obtain:

$Z_{train}, Z_{val}, and\ Z_{test}$

1.3 Concatenate all folds to construct the OOF Metadata

---



https://doi.org/10.64922/jteas.v1i1.86

$Z_{train}$={$r^{m_1}, r^{m1}_{lag1}, r^{m1}_{lag2}, r^{m1}_{lag3}, H_1 \ldots H_p, X_t$}

Step 2: Step-based walk-forward forecasting on the Validation and Test set

2.1 For each Validation or Test set:

Train $\mathcal{M}_1, \mathcal{M}_2$ $on$ $(X_{hist}, y_{hist})$

Predict $\hat{y}^{(j)}_{t:t+s}$ on unseen data

Compute residual and extract latent features

Construct Enriched metadata:

$Z_{val}, Z_{test}$={$r^{m_1}, r^{m1}_{lag1}, r^{m1}_{lag2}, r^{m1}_{lag3}, H_1 \ldots H_p, X_t$}

Where:

$Z_t$={$Z_{train}, Z_{val}, Z_{test}$}

Step 3: Domain Adversarial Feature Purification

Domain construction:

$d_t \in \{1 \ldots D_4\}$

3.1 DANN Optimization

Feature extractor F(•) with gradient reversal

$\mathcal{L} = \mathcal{L}_{task}(y, \hat{y}) + \alpha\mathcal{L}_{domain}(d, \dot{d})$

3.2 Domain-informed sample weighting

Compute domain loss:

$$\mathcal{L}_d = -\sum_{t=1}^{N} d_t \log(\dot{d}_t)$$

Define weight:

$\omega_t = \exp(-\gamma L_d)$

$\mathcal{M}_l$ using the purified features

$v_t = f(Z_t)$

Step 4: Meta-learner Training and Evaluation

4.1 Train Meta-learner $\mathcal{M}_l$ on $v_{t_{train}}$, Validate on $v_{t_{val}}$, tune hyperparameters with TimeSeriesSplit

4.2 Retrain $\mathcal{M}_l$ on $v_{t_{train}} \cup v_{t_{val}}$

4.3 Evaluate final performance on $v_{t_{test}}$

Output:



https://doi.org/10.64922/jteas.v1i1.86

| |
|---|
| Final prediction $\hat{y}_t$ |
| Evaluation, Residual diagnostics, and SHAP interpretation |

**Results**

This section presents the ablation results for the proposed multilevel stacking with DANN and a traditional stacking framework (without DANN). These two frameworks differ in how they integrate DANN into Random Forest meta-learning.

Both frameworks incorporate XGB residual, residual lag features, LSTM latent features, base, and engineered features as Enriched metadata. This enriched metadata is based on the assertion that ensembles of diverse models with uncorrelated errors yield superior performance within stacking frameworks.

The objective of the study is to preserve the latent statistical structure of the time-series process; no scaling transformation is applied to the dataset. The evaluation, therefore, focuses on prediction accuracy, temporal stability, distributional robustness, residual diagnostics, and SHAP interpretability across the validation and test regimes.

**Performance Accuracy**

The predictive accuracy of the Random Forest meta-learner with and without DANN is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Scaled Error (MASE), Normalized RMSE (NRMSE), and R-square ($R^2$).

**Table 2**

*The Quantitative Performance Accuracy*

| *Stacking* | *Dataset* | *MAE* | *RMSE* | *MASE* | *NRMSE* | $R^2$ |
|---|---|---|---|---|---|---|
| Proposed (With DANN) | Validation | 64.25 | 127.22 | 0.068 | 0.025 | 0.985 |
| | Testing | 57.99 | 121.82 | 0.065 | 0.025 | 0.987 |
| Random Forest (Without DANN) | Validation | 55.57 | 124.61 | 0.059 | 0.024 | 0.985 |
| | Testing | 55.05 | 126.77 | 0.061 | 0.026 | 0.985 |

The quantification performance from both stacking frameworks indicates significant generalisation capacity. Each model achieved an $R^2$ of 0.98 on the validation and testing set, confirming robustness during model selection. More notably, the test set results demonstrate near-perfect alignment between predicted and observed renewable energy output. On the contrary, the Random Forest without DANN outperformed the proposed framework with DANN, as indicated by MAE and RMSE. There is a slightly lower error in validation, with MAE of 55.58Wh and RMSE of 124.61Wh, compared with MAE of 64.25Wh and RMSE of 127.23Wh for the DANN enhanced framework.

Similarly, the test MAE of the model without DANN is 55.05Wh, marginally lower than 57.99Wh observed for the proposed model with DANN configuration. This difference corresponds approximately to 2-9Wh, suggesting that each configuration provides highly accurate predictions. The MASE values below 0.07 for both configurations indicate strong predictive performance relative to a naïve benchmark, confirming that the stacking framework effectively captures temporal dependencies within the energy generation process. Likewise, NRMSE values around 0.025 show that prediction errors constitute only a small fraction of the target range.



https://doi.org/10.64922/jteas.v1i1.86

**Figure 2**

*Validation set of Actual and Predicted plot of the Multilevel Stacking with DANN*

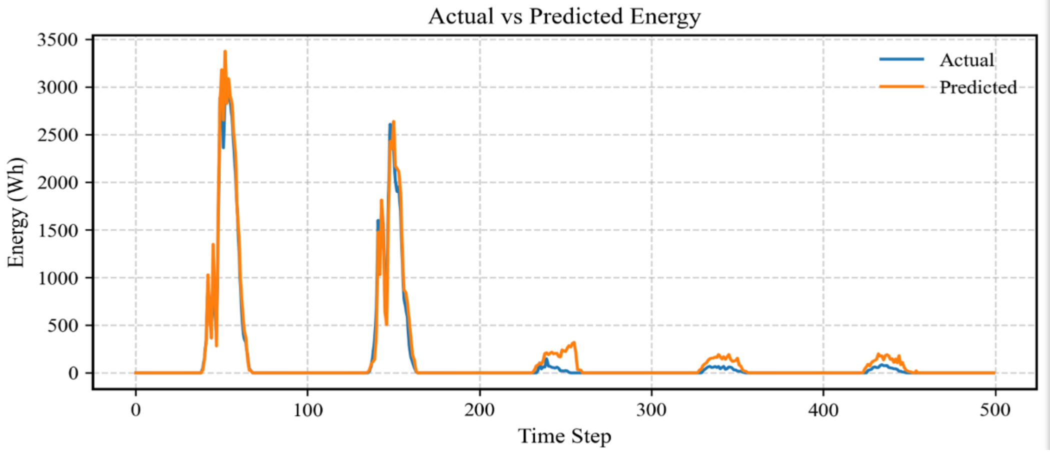
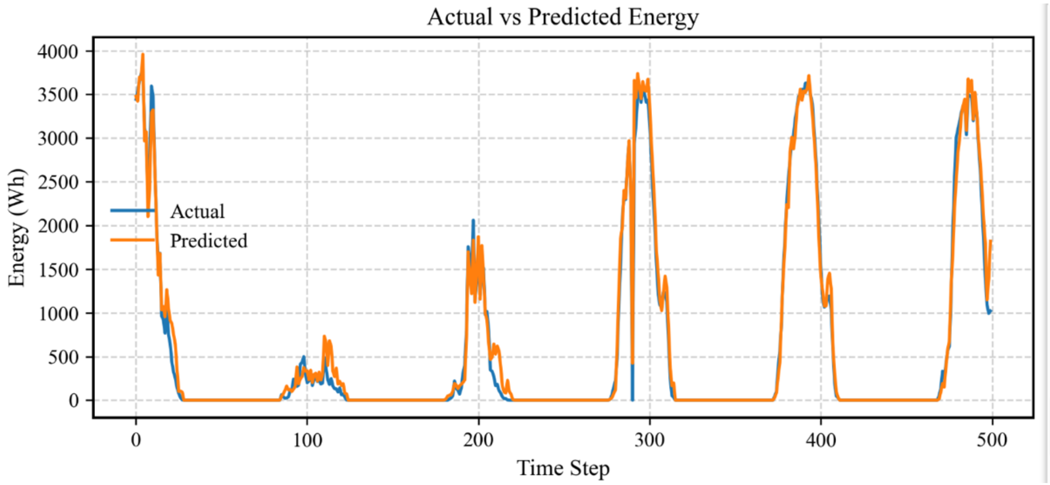



**Figure 3**

*Testing set of Actual and Predicted plot of the Multilevel Stacking with DANN*




## Temporal Stability

The study evaluates the ability of the model to maintain consistent predictive performance across time regimes, as shown in Table 3.

**Table 3**

*Temporal Stability across Time Regimes*

| *Stacking* | *Dataset* | *Error Variance* | *Drift Magnitude (ΔMAE)* | *Prediction Variance* |
|---|---|---|---|---|
| Proposed (With DANN) | Validation | 462.27 | 3.27 | $1.04\times10^6$ |
| | Test | 770.67 | 84.55 | $1.08\times10^6$ |
| Random Forest (Without DANN) | Validation | 459.25 | 8.78 | $1.04\times10^6$ |
| | Test | 885.88 | 43.99 | $1.09\times10^6$ |

The validation error variance for the Random Forest with and without DANN, shown in Table 3, is almost the same and an approximate 3.02 difference, indicating comparable variability in short-term predictions. However, the differences in energy move clearly in the test regime, where the DANN-enhanced framework exhibits a lower error variance of 770.68 compared with 885.88 of the Random Forest without DANN


https://doi.org/10.64922/jteas.v1i1.86

configuration. This reduction suggests that adversarial representation learning improves generalisation stability under unseen temporal conditions.

The drift magnitude is the change in MAE across time windows. The validation drift magnitude is substantially lower with the DANN-enhanced configuration of 3.27Wh compared with the Random Forest without DANN configuration of 8.78Wh. This indicates an improved short-term temporal consistency.

On the contrary, the test drift magnitude is larger in the DANN-enhanced configuration of 84.56Wh than in the configuration without DANN of 43.99Wh. This suggests that although adversarial training stabilises early temporal regimes, the long-term regimes still exhibit residual non-stationary dynamics.

**Robustness to Distribution Drift**

The study evaluates distributional robustness using the Population Stability Index (PSI), Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, domain classification accuracy, and Maximum Mean Discrepancy (MMD).

**Table 4**

*Robustness to Distribution Drift*

| *Stacking* | *Dataset* | *PSI* | *KL Divergence* | *JS Divergence* | *DCA* | *MMD* |
|---|---|---|---|---|---|---|
| Proposed (With DANN) | Validation | 0.013 | 0.010 | 0.002 | 0.824 | 0.00048 |
| | Test | 0.006 | 0.005 | 0.001 | | |
| Random Forest (Without DANN) | Validation | 0.038 | 0.024 | 0.004 | 0.834 | 0.00041 |
| | Test | 0.019 | 0.010 | 0.001 | | |

The results indicate that DANN substantially reduces distributional divergence between training and unseen regimes. Across all divergence metrics, the DANN-enhanced configuration shows consistently small divergence values. This indicates an improved alignment between training and test distributions.

However, the MMD is higher for the DANN configuration, which may reflect the tradeoff between enforcing domain invariance and preserving local feature variability. The domain classifier accuracy is approximately 0.82 and 0.83 in both Random Forest with DANN and without DANN, respectively. This indicates that temporal regimes remain partially distinguishable.

**Residual Diagnostics**

The study conducts residual diagnostics on the predicted results to assess whether prediction errors are unbiased and temporally uncorrelated. This is how each framework is effectively mitigating distribution shifts, as evidenced by improved performance from validation to testing.

**Table 5**

*Residual Diagnostic Summary*

| *Stacking* | *Dataset* | *Number of Samples* | *Mean* | *Standard Deviation* | *Min* | *Max* | *Skew* |
|---|---|---|---|---|---|---|---|
| Proposed (With DANN) | Validation | 29516 | 2.883 | 127.195 | -1495.16 | 1070.53 | 0.36530 |
| | Testing | 29517 | -1.854 | 121.806 | -1624.60 | 1490.07 | -0.01115 |
| Random Forest | Validation | 29516 | -9.186 | 124.275 | -2199.86 | 1829.26 | -0.30789 |
| (Without DANN) | Testing | 29517 | -14.22 | 125.974 | -2196.43 | 1704.18 | -0.67411 |

The residual mean shows a near-zero mean in the validation and test set of the DANN-enhanced configuration. The validation mean of 2.88Wh and test mean of -1,85Wh, suggesting minimal systematic bias. The residual standard deviation is 127.20Wh on the validation and 121.81Wh on the test set, which is consistent with the RMSE values.



https://doi.org/10.64922/jteas.v1i1.86

Similarly, the Random Forest without DANN configuration shows a negative bias residual mean of -9.19Wh on validation and -14.22Wh on the test set. Although the difference is relatively small, it indicates a greater systematic under-prediction compared with the DANN-enhanced configuration.

The residual skewness shows differences between the two models. The DANN model shows a near-symmetric residual distribution with 0.37 skewness on validation and -0.01 on the test set. Whereas the Random Forest without DANN configuration exhibits negative skewness, particularly in the test set of -0.67. This suggests that the adversarial feature learning helps reduce asymmetric prediction errors.

**SHAP Analysis and Interpretation**

This section demonstrates the quantification results from the SHAP global importance with a drift score and stability diagnostics. The feature drift score quantifies the degree of distributional change across temporal regimes, and a stability ratio indicates the consistency of feature contributions under temporal shift.

**Table 6**

*SHAP Global Feature Importance, Drift Score, and Stability Ratio*

| *Rank* | *Feature* | *Importance* | *Drift Score* | *Stability ratio* |
|---|---|---|---|---|
| 1 | xgb_resid | 163.57 | 48.12 | 0.948630 |
| 2 | latent_4 | 132.00 | 5.37 | 0.616725 |
| 3 | latent_6 | 94.25 | 7.75 | 0.753409 |
| 4 | latent_0 | 81.19 | 6.80 | 0.836090 |
| 5 | sunlightTime/daylenght | 66.99 | 1.82 | 0.870719 |
| 6 | res_lag1 | 59.80 | 8.56 | 0.415484 |
| 7 | latent_1 | 58.27 | 3.68 | 0.290659 |
| 8 | latent_2 | 54.32 | 2.43 | 0.392489 |
| 9 | sunlightTime | 43.45 | 7.09 | 0.898036 |
| 10 | latent_5 | 36.12 | 1.40 | 0.259783 |

The xgb_resid represents the XGB residual error, and emerged as the most influential predictor with an importance score of 163.57. This shows the reliance of the meta-learner on error-correction signals. However, the drift score of 48.12 suggests that the residual distribution undergoes substantial changes across temporal regimes. Despite the high drift score, the stability ratio of 0.95 suggests that the predictive contribution of xgb_resid remains highly consistent.

Moreover, the top-ranked predictors correspond to latent features extracted from the LSTM temporal encoder. The latent_4 has the highest importance of 132.00 compared with latent-6 and latent_0, which have the importance of 94.25 and 81.19, respectively. The drift score associated with these features ranges between 2 and 8. This suggests that the temporal representations extracted by the LSTM are less sensitive to distributional shifts, and their stability ratio, which ranges from 0.62 to 0.84, shows consistency across regimes.

The weather features, namely sunlightTime/dayLength and sunlightTime, appear within the top influential predictors. The drift score of the sunlightTime/dayLength of 1.82, indicating that the feature remains highly stable across temporal regimes. The stability ratio of 0.87 confirms that sunlightTime/dayLength contributes consistently to the prediction process.

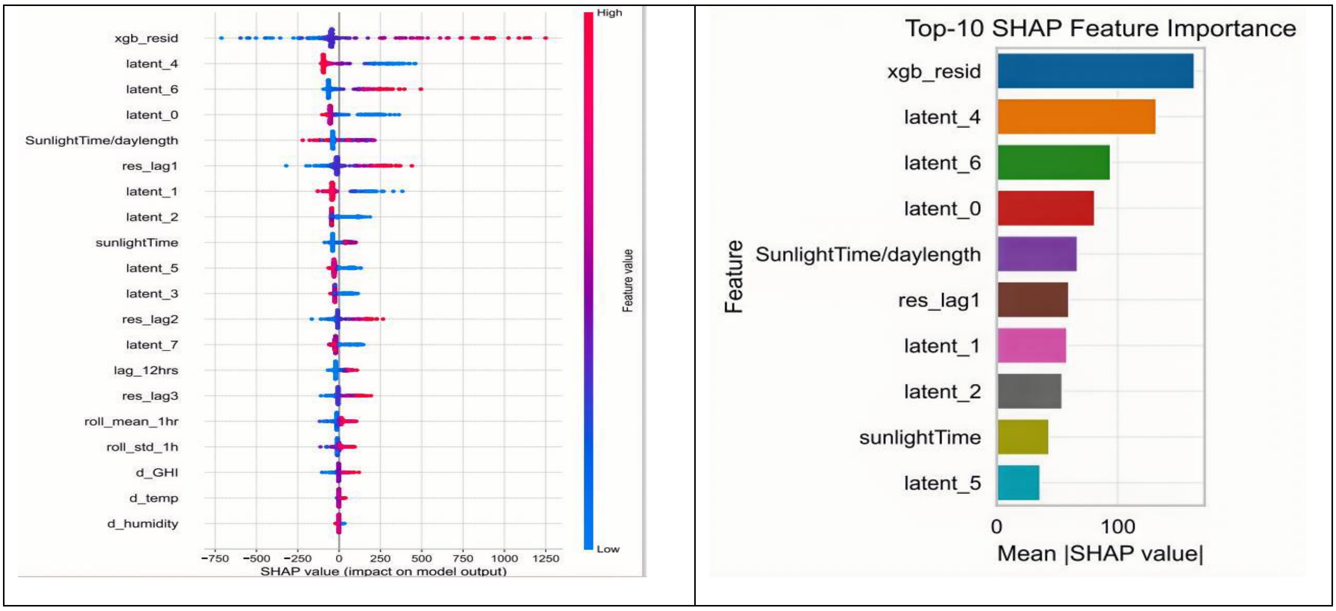
Similarly, sunlightTime shows a high stability ratio of 0.90 and reinforces the importance of integrating physical domain knowledge into meta-learning in renewable energy.



https://doi.org/10.64922/jteas.v1i1.86

**Figure 4**

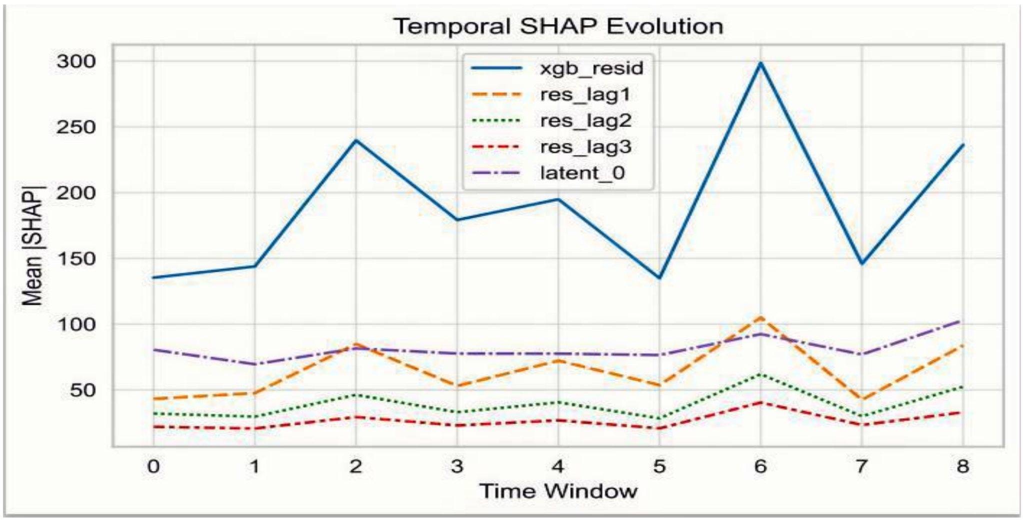
*Plot of Global SHAP Importance*




To verify the invariance of SHAP importance distribution across diverse domains, a Jensen-Shannon Divergence (JSD) is conducted on the training, validation, and testing sets. The divergence of 0.0669 between the training and validation, 0.0415 on training and testing, and 0.0309 on the validation and testing. All the values are less than 0.1, so the SHAP importance distribution across domains is nearly indistinguishable.

**Figure 5**

*The Temporal SHAP Evolution*




Similarly, Spearman's rank correlation is 0.919 between training and validation, 0.935 between training and testing, and 0.991 between validation and testing. These correlations confirm feature consistency across temporal regimes and no structural importance collapses.

**Conclusion**

The proposed multilevel stacking framework aims to minimise forecasting error, reduce systematic bias, and enforce temporal domain invariance. The integration of Walk-forward validation prevents data leakage by



https://doi.org/10.64922/jteas.v1i1.86

training on historical data and predicting on unseen data. The base learners XGB and LSTM function as residual and latent encoders to provide the structure of errors in the metadata. The fusion of residual, short-term residual lags and LSTM latent embedding improves domain transition stability. This is relative to both model configuration achieving a high explanatory power with $R^2$ values exceeding 0.98 across validation and test regimes.

Similarly, the MAE difference of the two configurations, of which the Random Forest without DANN configuration, exceeds approximately 2-9Wh on both regimes. This is extremely small relative to the maximum production level of 5020Wh, suggesting that integration of the enriched metadata helps both configurations provide highly accurate predictions.

Even though the quantitative performance results show that domain adversarial regularisation does not primarily improve point prediction accuracy, it instead influences distributional and temporal robustness.

The MASE values below 0.07 on both validation and test regimes indicate strong predictive performance relative to a naïve benchmark. This confirms that the stacking framework effectively captures temporal dependencies. Likewise, NRMSE values around 0.025 demonstrate that prediction errors constitute only a small fraction of the target range. The temporal stability results indicate that the configuration with DANN achieves a substantially lower drift magnitude of 3.27 Wh in the validation set, indicating improved short-term temporal consistency. Conversely, it reduces to 84.56Wh in the testing set. This shows an inherent difficulty of high-frequency energy production data where weather conditions fluctuation introduces an abrupt shift in the data distributions.

Across all divergence metrics, the proposed stacking framework shows consistently small divergence values. This indicates an improved alignment between the training and unseen distributions. The residual skewness shows how the integration of DANN in the Random Forest meta-learning has improved the model. The configuration with DANN achieves a near symmetric residual distribution with 0.37 skewness on the validation and -0.01 skewness on the testing set, compared to the Random Forest without DANN configuration, exhibiting negative skewness. This suggests that the adversarial feature learning helps reduce asymmetric prediction errors.

This proposed framework demonstrates empirically that stacking should not be evaluated solely on MAE/RMSE but on residual distribution behaviour under domain transition. Furthermore, SHAP analysis provides empirical evidence that enriched metadata significantly enhances the learning capacity, allowing the meta-learner to capture complex dependencies. Further research could explore other domain adaptation methods beyond DANN, such as Maximum Mean Discrepancy (MMD) or Adversarial Discriminative Domain Adaptation (ADDA) (Gretton et al., 2012; Tzeng et al., 2017), and experiment with other advanced ensemble methods (e.g., Gradient Boosting Machines, Neural Networks) to ascertain if alternative combinations yield further improvements in forecasting accuracy and stability to potentially enhance feature invariance and model robustness across diverse temporal regimes.

**Acknowledgement**

The authors would like to acknowledge all individuals whose constructive comments and suggestions helped to improve the quality of this paper. No external funding was received for the completion of this work.

**Author Contribution Statement**

Conceptualization: PCA (lead), DD (equal)

Data Curation: WA (lead), GA (equal), EAF (supporting)

Formal Analysis: PCA (lead), SKA (supporting), JCD (supporting)

Methodology: DD (equal), SKA (equal), JCD (lead), WAA (equal)

Software: WA (lead), GA (equal), EAF (supporting)



https://doi.org/10.64922/jteas.v1i1.86

Validation: PCA (lead), SKA (supporting), JCD (equal), WA (supporting), GA (supporting), DM (supporting), OKB (supporting)

Writing – Original Draft: PCA (lead), DD (equal), DM (supporting)

Writing – Review & Editing: PCA (equal), WAA (equal), EAF (equal), OKB (equal)

**Conflict of Interest Statement**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**AI Disclosure Statement**

The authors confirm that no artificial intelligence or generative AI tools were used in the design of this study.

**References**


Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. https://doi.org/10.1145/2523813

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702. https://doi.org/10.1016/j.ejor.2016.10.031

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., & Bagnall, A. (2021). HIVE-COTE 2.0: A new meta ensemble for time series classification. *Machine Learning*, *110*(11), 3211–3243. https://doi.org/10.1007/s10994-021-06057-9

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. *Computer Science Department, Trinity College Dublin*, *106*(2), 58.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6848–6856). https://doi.org/10.1109/CVPR.2018.00716





https://doi.org/10.64922/jteas.v1i1.86