Model Capacity Alignment under Data Scarcity

Tree-Based Ensembles versus Deep Recurrent Networks for Daily Solar Radiation Forecasting in West Africa

M. Rasheed | University of Hull | M.RASHEED2-2023@hull.ac.uk

1st International Online Conference on Inventions (IOCIV 2026) - Energy Security and Sustainable Development

Key message: With only ~570 training samples per country, tree ensembles generalise strongly; LSTM models stay near baseline.

0.9777
Best R2 (XGBoost, Nigeria)

~700
Daily observations per country

20
Important engineered predictors

~0.05
Deep recurrent R2 baseline

Research objective

Solar forecasting literature often assumes large high-frequency datasets. Many emerging energy systems do not have that luxury.

- Test whether model capacity matches the data-scarce daily forecasting regime.
- Compare tree ensembles and deep recurrent networks under identical preprocessing and chronological validation.

Data and study region

Countries: Nigeria, Ghana, Senegal
Period: September 2021 to November 2023
Scale: 711-715 daily observations per country
Split: 80/20 chronological split, yielding approximately 569 training samples and 142-143 test samples.

- Daily ground-station measurements from the World Bank energydata.info platform.
- The countries represent different irradiance persistence and volatility regimes.

Methodology at a glance

1. Raw weather inputs: 11 base meteorological variables.
2. Temporal embedding: lags, rolling means/standard deviations, seasonal encoding and interaction terms.
3. Models compared: Random Forest, XGBoost, LSTM and CNN-LSTM.
4. Validation: chronological 80/20 split to avoid temporal leakage.

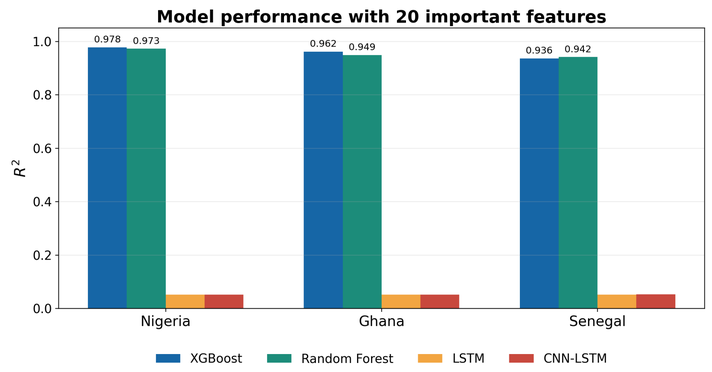
Learning-theory rationale

An LSTM with input dimension d and hidden size h has parameter count:

P = 4(dh + h2 + h)

With d approximately 20 and h = 32, this is about 6,784 parameters. With fewer than 600 training samples - and lower effective sample size under autocorrelation - recurrent networks are underconstrained.

Main result: capacity alignment wins

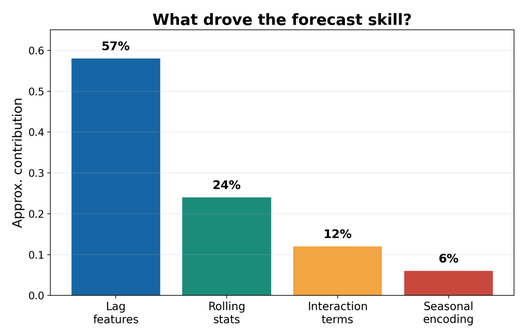


- Tree ensembles substantially outperform recurrent architectures across all countries.
- XGBoost gives the strongest overall performance, with R2 up to 0.9777.

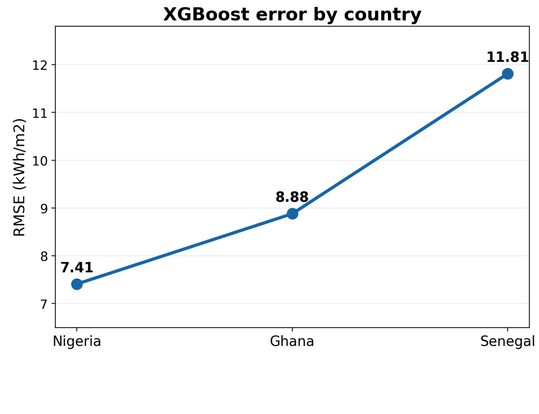
Best-performing configuration

Country	XGBoost R2	RF R2	XGB RMSE
Nigeria	0.9777	0.9728	7.41
Ghana	0.9618	0.9493	8.88
Senegal	0.9363	0.9417	11.81

Feature signal: lags dominate



Geographic predictability gradient



Why the deep models failed here

- The training set is small relative to the recurrent parameter count.
- Autocorrelation reduces the effective independent sample size.
- Dropout and early stopping help but do not create missing information.
- Tree ensembles impose stronger structural regularisation through bagging, shrinkage, depth limits, and subsampling.

Deployment implications

- Use XGBoost or Random Forest as robust baselines for data-scarce solar forecasting.
- Prioritise engineered temporal features before increasing model complexity.
- CPU-only training and quarterly retraining are practical for utilities and mini-grid operators.

Conclusion

In small-sample daily solar forecasting, model capacity must match dataset scale. For ~700 observations per country, structured temporal features plus tree ensembles are more reliable than high-capacity recurrent neural networks.

Take-home recommendations

1. Start simple
Use XGBoost/Random Forest as first-line models in small-N meteorological settings.

2. Engineer persistence
Prioritise lag and rolling features before adding deep sequence architectures.

3. Scale before deep learning
Recurrent models may need larger, higher-frequency, multi-year datasets to justify their capacity.

References: Breiman (2001); Chen & Guestrin (2016); Bishop (2006); Goodfellow et al. (2016); World Bank energydata.info (2023).

Conflict of interest: none declared. AI disclosure: poster layout/copy editing prepared with AI assistance; scientific content/results from author manuscript.