International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Sustainable Energy Forecasting: A Machine Learning Framework for India's Power Sector

Omkar Singh[1], Sherilyn Kevin[2], Shreya Muley[3]
[1] Head of Department, [2] Assistant Professor , [3] P.G. Student
[1,3] Department of Data Science, [2] Department of Information Technology
[1,2,3] Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai - 40010, India
[1]omkarsingh@tcsc.edu.in, [2]sherilynkevin@tcsc.edu.in, [3]shreyamuley31@gmail.com

----------**********----------

Abstract:

This study presents a comprehensive analysis of India's electricity generation trends, focusing on oil, gas, and wind energy sources, using advanced machine learning models. The research leverages data from the India Climate and Energy Dashboard (ICED) to explore state-wise temporal trends in electricity generation, aiming to provide insights into generation patterns, demand forecasting, and resource utilization. Three datasets—wind, oil, and gas electricity generation—were analyzed using ARIMA, LSTM, and Gaussian Mixture Models (GMM) to capture temporal dependencies, seasonal trends, and regional disparities in energy production. The ARIMA and LSTM models demonstrated strong performance in forecasting electricity generation, with LSTM showing particularly high accuracy in capturing complex temporal patterns. GMM clustering revealed significant regional disparities, highlighting states like Gujarat and Maharashtra as leaders in operational capacity. The study underscores the growing importance of renewable energy, particularly wind, while also emphasizing the challenges of grid integration and variability in generation. The findings offer valuable insights for energy planners and policymakers, supporting informed decision-making in resource allocation, grid management, and sustainable energy development. The research contributes to the broader goal of achieving energy security and sustainability in India, providing a robust methodological framework for future studies in energy forecasting and optimization.

Keywords — **Electricity Generation, Renewable Energy, Machine Learning, ARIMA, LSTM, Gaussian Mixture Model (GMM), Energy Forecasting, Grid Integration, Resource Optimization, Sustainable Development**

----------**********----------

I. INTRODUCTION

The demand for electricity in India has grown substantially due to rapid industrialization, urbanization, and evolving policy frameworks. As of March 2024, the nation's total installed power capacity stands at approximately 442 GW, derived from a mix of coal, natural gas, hydropower, solar, and wind energy. However, managing energy generation across states remains a challenge due to regional differences in resource availability, infrastructure, and consumption patterns. While coal continues to be a major energy source, increasing emphasis on renewable energy has introduced new challenges in grid integration and power stability. This study aims to analyze state-wise temporal trends in electricity generation using machine learning models, providing insights into generation patterns, demand forecasting, and resource utilization. By identifying fluctuations and long-term trends, the research supports informed decision-making in energy planning, policy formulation, and sustainable development. The study underscores the importance of data-driven methodologies in optimizing India's electricity generation landscape while ensuring efficient and equitable resource distribution.

II. LITERATURE REVIEW

Accurate electricity demand forecasting is critical for India's energy security, given its rapid urbanization, climate commitments, and persistent energy poverty. Globally, frameworks like D-FED (Bedi & Toshniwal, 2019) have advanced long-term demand forecasting by addressing temporal dependencies, offering insights adaptable to India's dynamic consumption patterns [1]. Regionally, hybrid machine learning models (ANN, LSTM, SVR) reviewed by Niharika and Singh (2022) demonstrate superior accuracy for India's electricity consumption trends, emphasizing the need for localized solutions [2]. Regression techniques, such as those applied by Rekhade and Sakhare (2021) to forecast sector-specific demand (industry, domestic, agriculture), further validate the utility of data-driven approaches when aligned with national datasets from India's Central Electricity Authority [3]. Chen et al. (2023) highlighted the role of socio-economic and climate variables in short-term regional forecasting, a methodology highly relevant to India's diverse climatic zones and economic disparities [5]. An et al. (2019) projected a 5.17% annual growth in India's power generation using probabilistic methods, stressing the urgency of balancing supply and demand to mitigate energy poverty [6]. At the state level, Jain et al. (2020) developed the KmFuzz model for Uttar Pradesh, combining XGBoost, LSTM, and fuzzy logic to achieve a MAPE of 1.94%, showcasing the value of granular data (15-minute blocks) for regional planning [7]. Similarly, Rahman et al. (2018) linked India's total electricity consumption (TEC) to GDP, forecasting 1,778,358 MW for 2030 and underscoring socio-economic drivers in energy policy [8]. Renewable integration remains pivotal, with Benti et al. (2023) advocating machine learning over traditional ARIMA models to address India's grid challenges [4]. Garg et al. (2021) achieved near-perfect solar generation forecasts (R^2 = 0.99) using GDP and consumption data, while Barbar et al. (2021) highlighted air conditioning and electric vehicles as key demand drivers, proposing customizable regression models for infrastructure scalability [9][10]. Despite advancements, persistent issues like data variability and interpretability gaps necessitate continued innovation [11]. This review consolidates India-centric forecasting advancements, emphasizing their role in policy formulation, climate resilience, and sustainable development.

III. METHODOLOGY

3.1. Dataset Collection:

To analyze trends in India's electricity generation, data was sourced from the India Climate and Energy Dashboard (ICED), accessible at https://iced.niti.gov.in/energy/electricity/generation, developed by NITI Aayog, which provides publicly available records on India's electricity generation. Three datasets were selected:

- **Wind Generation (July 2024)**: Contains operational capacities of wind power plants by state.
- **Oil & Gas Generation (July 2024)**: Details operational capacities of oil and gas plants across India.
- **Daily Generation (2015–2024)**: Provides daily electricity generation values for oil, gas, and wind sources.

These datasets allow for regional and temporal analysis of India's energy trends.

3.2. Data Processing:

- **Time Alignment**: Monthly and daily datasets were synchronized.
- **Data Cleaning**: Missing values were imputed, and duplicates were removed.
- **Unit Standardization**: Capacities were standardized to megawatts (MW), and generation to million units (MU) for consistency.
- **Filtering**: Only data for oil, gas, and wind sources was retained in the daily generation dataset, focusing the analysis on these key resources.

3.3. Data Pre-processing:

To prepare for time series forecasting with ARIMA and Prophet:

- Data was normalized and resampled for consistency.
- Temporal granularity was unified across datasets.
- Missing values were handled to maintain data continuity.

3.4. Feature Engineering

To enhance predictive accuracy, the following features were extracted:

- **Seasonality Indicators**: Monthly, weekly, and daily patterns.
- **Lag Features**: Recent historical values to capture trends.

3.5. Model Selection

Model selection involved choosing ARIMA for time series forecasting, LSTM for capturing complex temporal dependencies, and GMM for clustering regional energy generation patterns.

3.6. Model Training

Models were trained on preprocessed data, with hyperparameters optimized for ARIMA (p, d, q), LSTM (epochs, batch size), and GMM (number of clusters).

3.7. Model Evaluation

Models were evaluated using MAE, MSE, and MAPE for ARIMA and LSTM, and silhouette score for GMM. Residual analysis ensured no significant autocorrelation in errors.

3.8. Cross-Validation

Time-series cross-validation was applied to ensure model robustness and avoid overfitting.

3.9. Forecast Integration

The forecasts generated by ARIMA and LSTM were integrated with clustering results from GMM to provide a comprehensive view of future electricity generation trends.

3.10. Documentation and Reproducibility

Comprehensive documentation and version control ensured methodological reproducibility.

3.11. Ethical Considerations and Fair Use

Ethical considerations ensured transparent and responsible use of forecasts for sustainable energy practices.

IV. RESULTS AND DISCUSSION

4.1. Oil and Gas

4.1.1. Exploratory Data Analysis

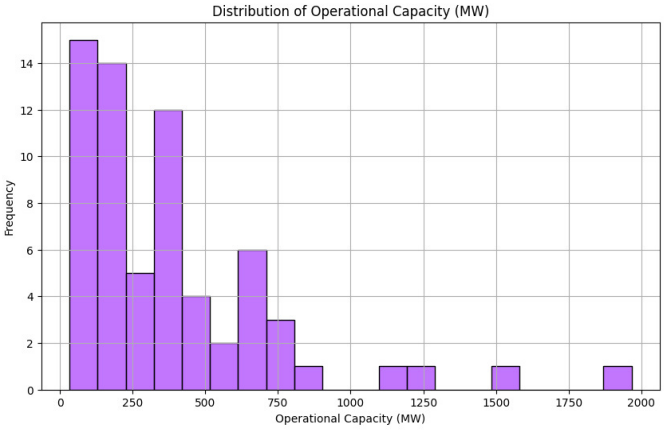


Fig.1. Distribution of Operational Capacity for Oil and Gas (MW)

Fig.1. shows the distribution of operational capacity (MW). Most units fall within the 0–500 MW range, with frequency peaking above 14. The distribution is left-skewed, indicating a predominance of smaller capacity units.

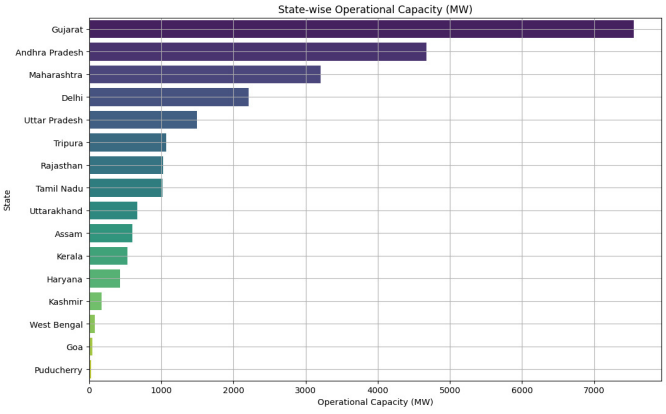


Fig.2. State - wise Operational Capacity for Oil and Gas (MW)

Fig.2. presents state-wise operational capacity (MW). Gujarat and Andhra Pradesh have the

highest capacities, highlighting regional disparities in energy infrastructure.

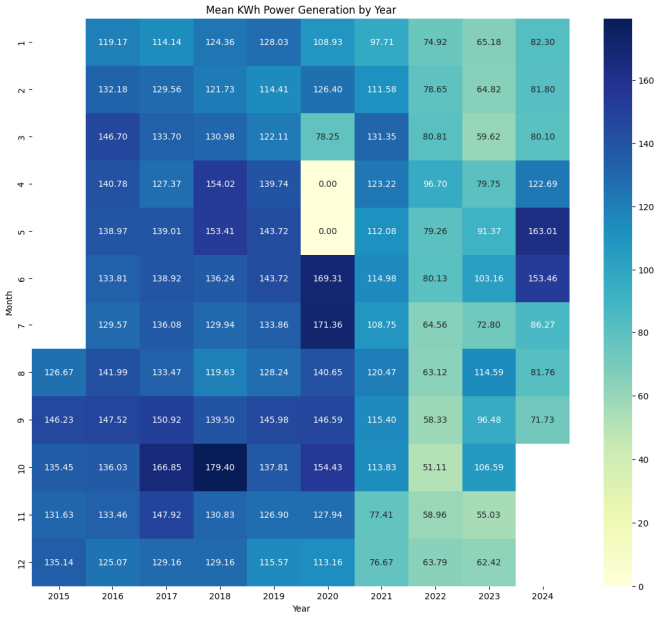
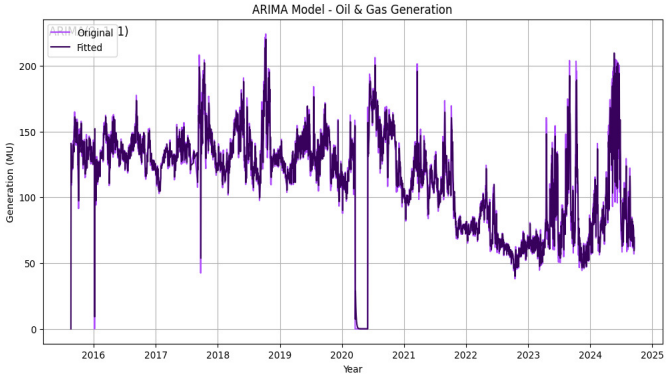


Fig.3. Mean KWh Power Generation by Year for Oil and Gas

Fig.3. shows monthly electricity generation (KWh) from 2015 to 2024. Peaks in 2021 (179.40 KWh) and dips in 2017 (108.93 KWh) indicate fluctuations, likely due to seasonal variations or policy changes. Recent trends suggest improved efficiency.

4.1.2. ARIMA Model



Statistic	coef	std err	z	P>|z|	[0.025	0.095]
ar.L1	0.7046	0.023	30.676	0.000	0.660	0.750
ar.L2	-0.1460	0.010	-14.897	0.000	-0.165	-0.127
ma.L1	-0.7738	0.023	-34.070	0.000	-0.818	-0.729
sigma2	128.9214	0.727	177.360	0.000	127.497	130.346

Fig.4. ARIMA Model for Oil and Gas Generation

Fig.4. presents ARIMA model results for oil and gas generation, showing original data, fitted values, and residuals. The model effectively captures trends, and the ADF test (p = 0.0002) confirms stationarity.

Table 1. SARIMAX Results for ARIMA Model Diagnostics and Goodness-of-Fit Metrics for Oil and Gas Generation

Statistic	Value	Statistic	Value
Dep. Variable	Generation (MU)	**Ljung-Box (L1) (Q)**	0.00
Model	ARIMA(2, 1, 1)	**Prob(Q)**	0.98
No. Observations	3278	**Heteroskedasticity (H)**	1.57
Log Likelihood	-12611.769	**Prob(H) (two-sided)**	0.00
AIC	25231.539	**Jarque-Bera (JB)**	238288.02
BIC	25255.918	**Prob(JB)**	0.00
HQIC	25240.268	**Skew**	-0.13

Table 2. presents SARIMA (2,1,1) model results, capturing seasonality in time-series data. Key statistics include Log-Likelihood, AIC, and BIC for model fit evaluation. The AR(1) (0.7046) shows strong positive correlation, while MA(1) (-0.7738) suggests negatively correlated errors. Diagnostic tests indicate no autocorrelation (Ljung-Box p = 0.98) but suggest non-normality (Jarque-Bera p = 0.00) and potential heteroskedasticity (H test p = 0.00).

Table 2. SARIMAX Results: Oil and Gas Generation

4.1.4. LSTM Model

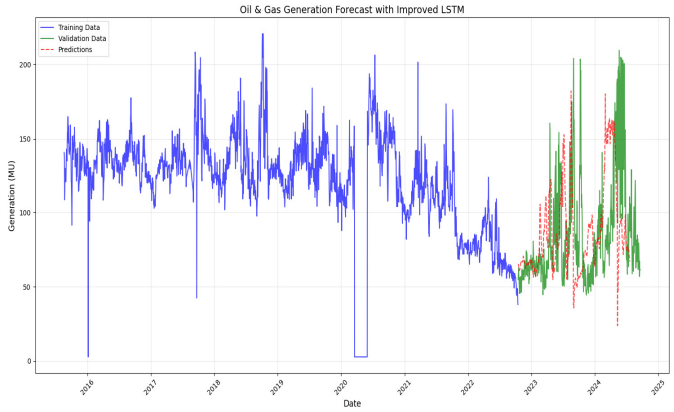


Fig.5. LSTM Model for Oil and Gas Generation

Fig.5. illustrates LSTM-based forecasting for oil and gas generation, showing training, validation, and predicted values. The model effectively captures trends, aligning closely with actual data.

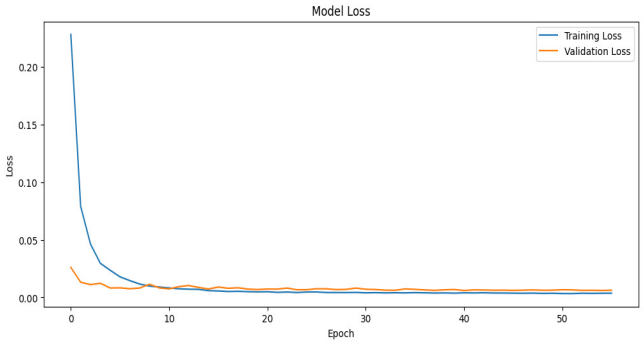


Fig.6. LSTM Model loss graph for Oil andGas Generation

Fig.6. depicts training vs. validation loss over epochs, showing effective model learning. Both losses decrease and stabilize around epoch 30, with no signs of overfitting.

4.1.5. GMM Model

Table 3. shows GMM clustering results for electricity generation data, identifying two clusters:

- Cluster 0 (2,331 points, mean: 133.94 MU) represents high-generation periods.
- Cluster 1 (947 points, mean: 65.12 MU) represents low-generation periods.

Table 3. GMM Results: Oil and Gas Generation

Cluster	Count	Mean Generation (MU)	Min Generation (MU)	Max Generation (MU)
0	2331	133.94	96.24	224.10
1	947	65.12	0.00	96.06

4.2. Wind

4.2.1. Exploratory Data Analysis

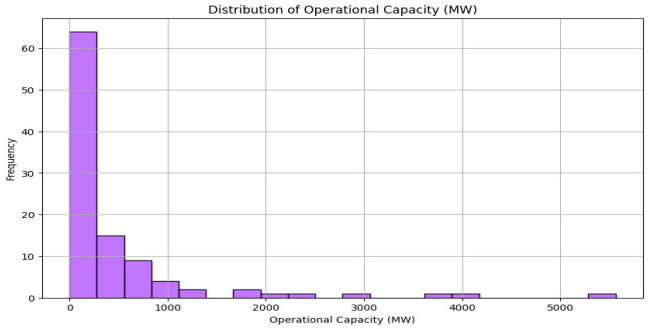


Fig.7. Distribution of Operational Capacity for Wind (MW)

Fig.7. shows the distribution of power plant capacities. Most plants have capacities below 1,000 MW, with a right-skewed distribution indicating fewer high-capacity plants.

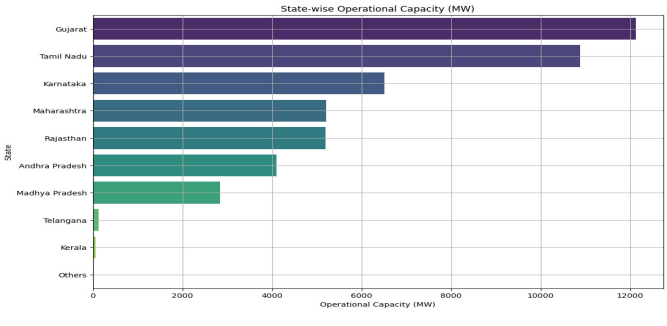


Fig.8. State - wise Operational Capacity for Wind (MW)

Fig.8. presents a bar chart of state-wise renewable energy capacity in India. Gujarat, Tamil Nadu, and Karnataka lead in operational capacity, followed by Maharashtra and Rajasthan. Andhra Pradesh shows growth, while Telangana, Kerala, and other states have lower capacity, indicating potential for expansion.

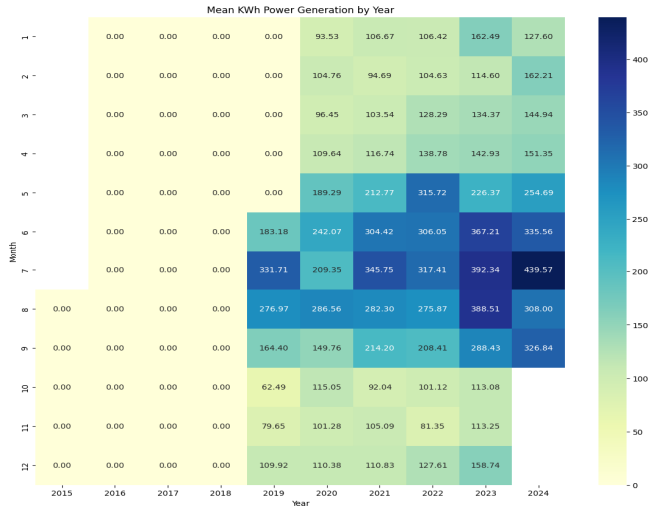


Fig.9. Mean KWh Power Generation by Year for Wind

Fig.9. presents a heatmap of average monthly kWh power generation by year. The color intensity indicates generation levels, showing an overall increase over time. A seasonal pattern is evident, with peaks in July (high summer demand) and lows in January. Variability across years may be influenced by weather, economic activity, or operational changes.

4.2.2. ARIMA Model

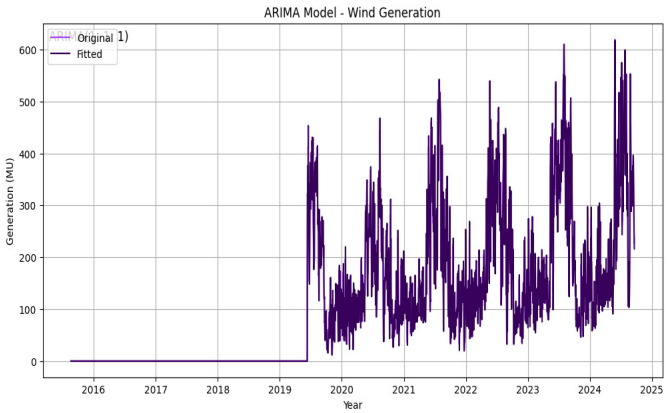


Fig.10. ARIMA Model for Oil and Gas Generation

Fig.10. presents a time series plot of wind generation data with an ARIMA model fit. The blue line represents the original data, while the purple line shows the fitted values. The model captures the overall trend and seasonality but shows some discrepancies, indicating possible missing factors. The ADF test (p-value: 0.014) confirms stationarity.

Table. 4 presents SARIMAX (1,1,1) model results for time series forecasting. The model accounts for autoregressive (AR), differencing (I), and moving average (MA) components. Key findings: AR(1) (-0.1718) and MA(1) (0.4613) significantly impact predictions. The Ljung-Box test (p=0.44) confirms no residual autocorrelation, while the Jarque-Bera test (p=0.00) suggests non-normal residuals. The model effectively captures trends but may have limitations due to residual distribution.

Statistic	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.095]**
ar.L1	-0.1718	0.037	-4.630	0.000	-0.245	-0.099
ma.L1	0.4613	0.034	13.629	0.000	0.395	0.528
sigma2	1023.7967	12.780	80.111	0.000	998.749	1048.845

Table 4. SARIMAX Results: Wind Generation

Statistic	Value	Statistic	Value
Dep. Variable	Generation MU)	Ljung-Box (L1) (Q)	0.61
Model	ARIMA(1, 1, 1)	Prob(Q)	0.44
No. Observations	3280	Heteroskedasticity (H)	inf
Log Likelihood	-16016.606	Prob(H) (two-sided)	0.00
AIC	32039.213	Jarque-Bera (JB)	4676.21
BIC	32057.498	Prob(JB)	0.00
HQIC	32045.760	Skew	-0.04

4.2.3. LSTM Model

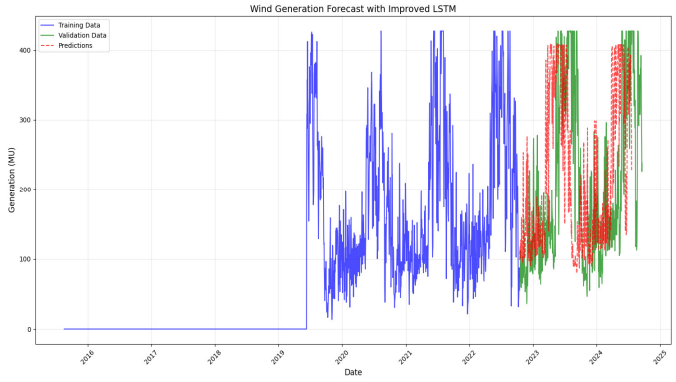


Fig.11. LSTM Model for Wind Generation

Fig.11. presents a time series plot of wind generation data with a forecast using an Improved LSTM model. The blue line represents training data, the green line shows validation data, and the red dotted line indicates the forecast. The model accurately captures seasonal trends, demonstrating its effectiveness for wind generation forecasting.

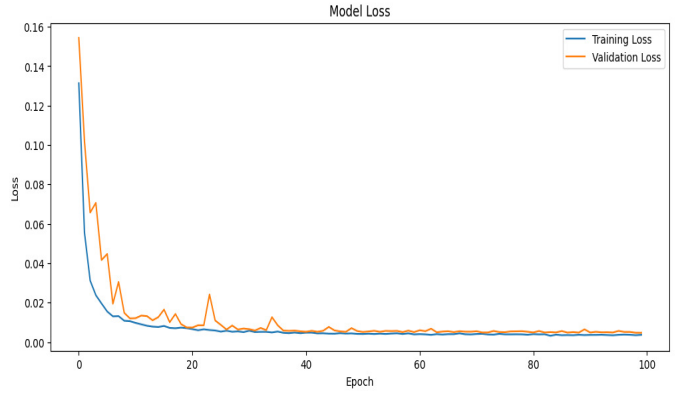


Fig.12. LSTM Model loss graph for Wind Generation

Fig.12. shows training and validation loss curves over 100 epochs. Training loss decreases sharply, while validation loss flattens after ~10 epochs, indicating overfitting. Techniques like early stopping and regularization can help mitigate this.

4.2.4. GMM Model

Table 5 presents clustering results of electricity generation using a Gaussian Mixture Model (GMM). Two clusters were identified: Cluster 0 (2,296 points, low generation: 0.00–145.74 MU, mean: 40.04 MU) and Cluster 1 (984 points, high generation: 146.04–619.44 MU, mean: 280.52 MU). This distinction aids targeted energy planning and resource allocation.

Table 5. GMM Results for Wind Generation

Cluster	Count	Mean Generation (MU)	Min Generation (MU)	Max Generation (MU)
0	2296	40.0444	0.00	145.74
1	984	280.5225	146.04	619.44

Number of clusters: 2
Log-likelihood of the model: -1.2136

V. FUTURE WORK

This study provides a foundation for improving electricity generation forecasting and resource optimization in India. Future research could integrate additional data sources (e.g., weather, economic indicators) to enhance model accuracy. Advanced techniques like Transformer-based models and real-time adaptive forecasting can further improve predictions. Addressing regional disparities in electricity generation, optimizing renewable energy integration, and studying long-term scenario analysis will be crucial for policy planning. Ethical considerations, transparency, and cross-country comparisons can further refine India's energy strategies for a sustainable future.

VI. CONCLUSION

This study analyzed India's electricity generation trends using ARIMA, LSTM, and GMM models, capturing temporal and regional patterns. LSTM showed high forecasting accuracy, making it a valuable tool for energy planning. GMM clustering revealed regional disparities, with Gujarat and Maharashtra leading in capacity, highlighting the need for balanced resource distribution. Wind energy analysis emphasized its growing role but also underscored integration challenges, necessitating research on storage and grid stability. The findings offer data-driven insights for optimizing resource allocation, enhancing infrastructure, and supporting renewable adoption.

This study's methodological framework can be extended to other energy sources and regions, contributing to energy security, sustainability, and informed policy decisions for India's evolving energy landscape.

VII. REFERENCES

[1] Bedi J, Toshniwal D. Deep learning framework to forecast electricity demand. Applied Energy. 2019;238:1312-26
[2] Niharika HK, Singh J. Electricity consumption forecasting system of India: A review.
[3] Rekhade R, Sakhare D. Forecasting sector-wise electricity consumption for India using various regression models. Current Science. 2021;121(3):365-71. https://doi.org/10.18520/CS/V121/I3/365-371.
[4] Benti NE, Chaka MD, Semie AG. Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. Sustainability. 2023;15(9):7087.
[5] Chen G, Hu Q, Wang J, Wang X, Zhu Y. Machine-learning-based electric power forecasting. Sustainability. 2023;15(14):11299.
[6] An Y, Zhou Y, Li R. Forecasting India's electricity demand using a range of probabilistic methods. Energies. 2019;12(13):2574. https://doi.org/10.3390/EN12132574.
[7] Jain R, Jain N, Gupta Y, Chugh T, Chugh T, Hemanth D. A modified fuzzy logic relation-based approach for electricity consumption forecasting in India. International Journal of Fuzzy Systems. 2020;22:461-75. https://doi.org/10.1007/S40815-019-00704-Z.
[8] Rahman H, Selvarasan I, Begum AJ. Short-term forecasting of total energy consumption for India-a black box based approach. Energies. 2018;11(12):3442. https://doi.org/10.3390/en11123442.
[9] Barbar M, Mallapragada DS, Alsup M, Stoner R. Scenarios of future Indian electricity demand accounting for space cooling and electric vehicle adoption. Scientific Data. 2021;8(1):178.
[10] Garg I, Raj A, Jajoria A. Solar power production forecasting in India using machine learning approach. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). 2021. p. 1265-72. IEEE.
[11] Various studies (summarized across the provided articles).

